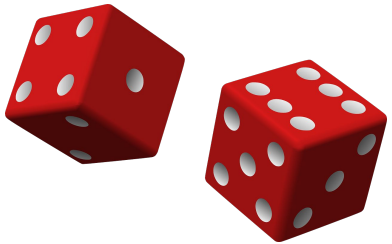
Análise e Pré-processamento de Dados

Introdução

Dados-Informação-Conhecimento

Dados-Informação-Conhecimento



- Dados
  - Elementos conhecidos de um problema.
  - Desprovidos de significado quando considerados isoladamente.

Dados-Informação-Conhecimento

- Informação
  - Um conjunto estruturado de dados.
  - Possuem utilidade e podem gerar ações.

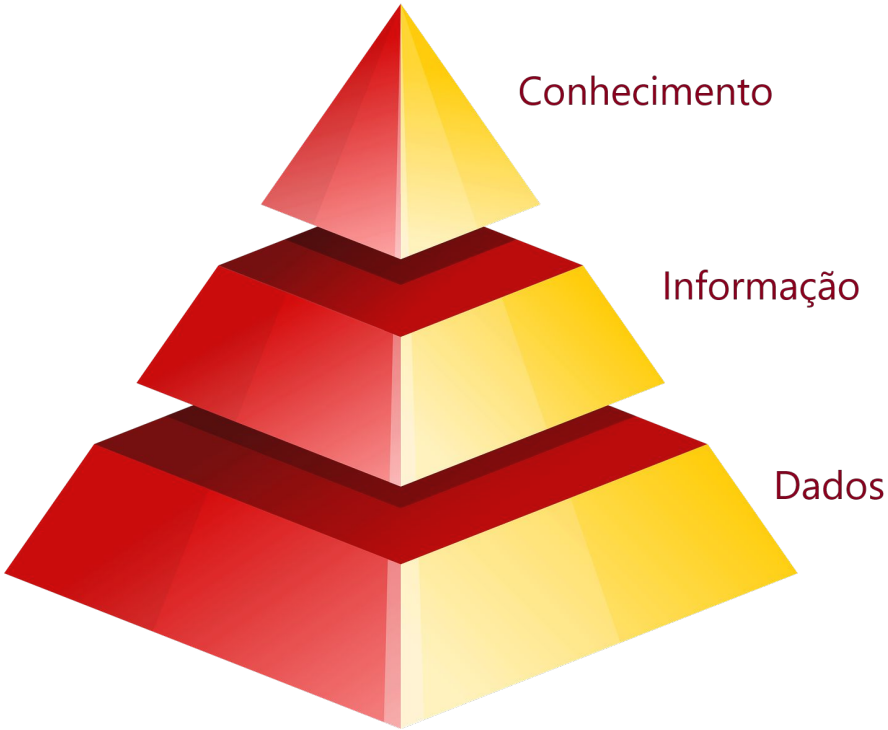
Dados-Informação-Conhecimento

- Conhecimento
  - Síntese de múltiplas fontes de informação.
  - Possui elevado significado e utilidade para o suporte à tomada de decisões.

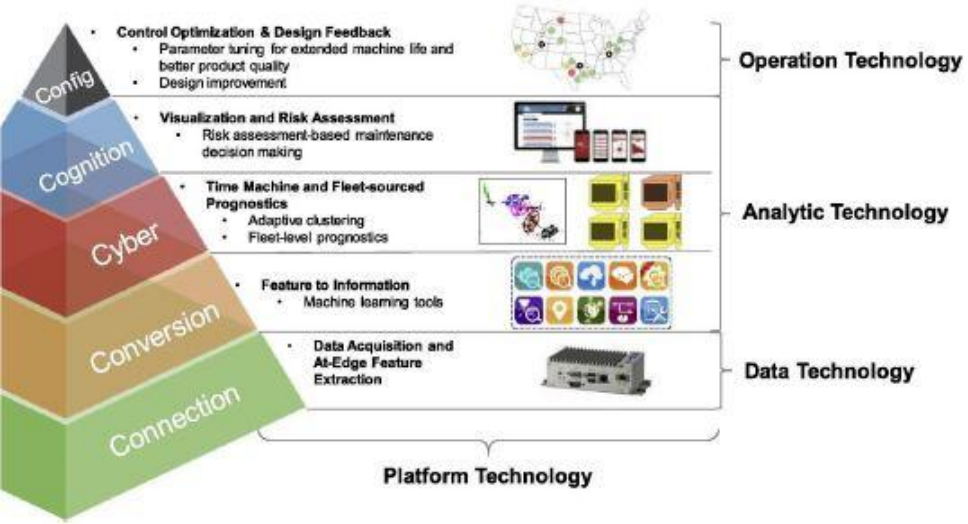
Dados-Informação-Conhecimento

- Computador manipula informações contidas em sua memória.
- **Instruções**: comandam o funcionamento da máquina e determinam a maneira como os dados devem ser tratados.
- **Dados**: informação que devem ser manipulada pelo computador.

Dados-Informação-Conhecimento



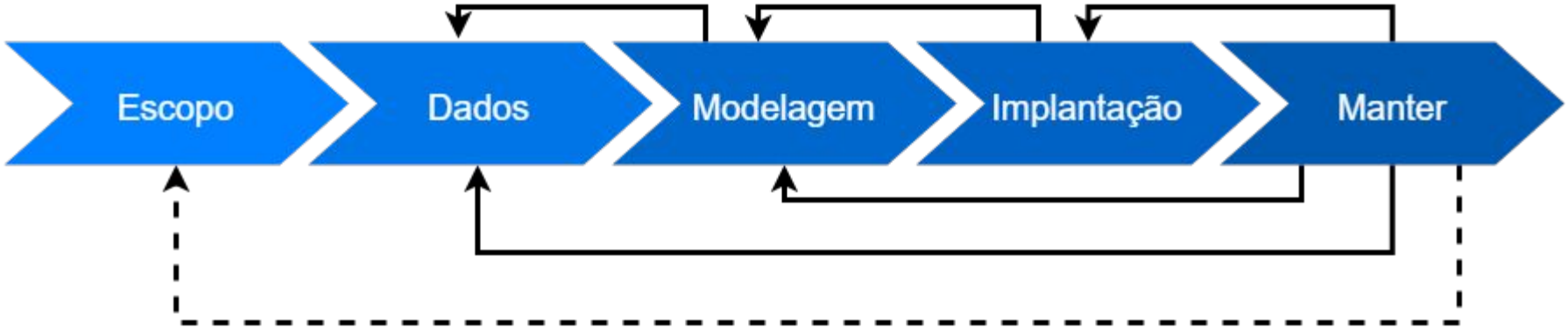
Dados-Informação-Conhecimento

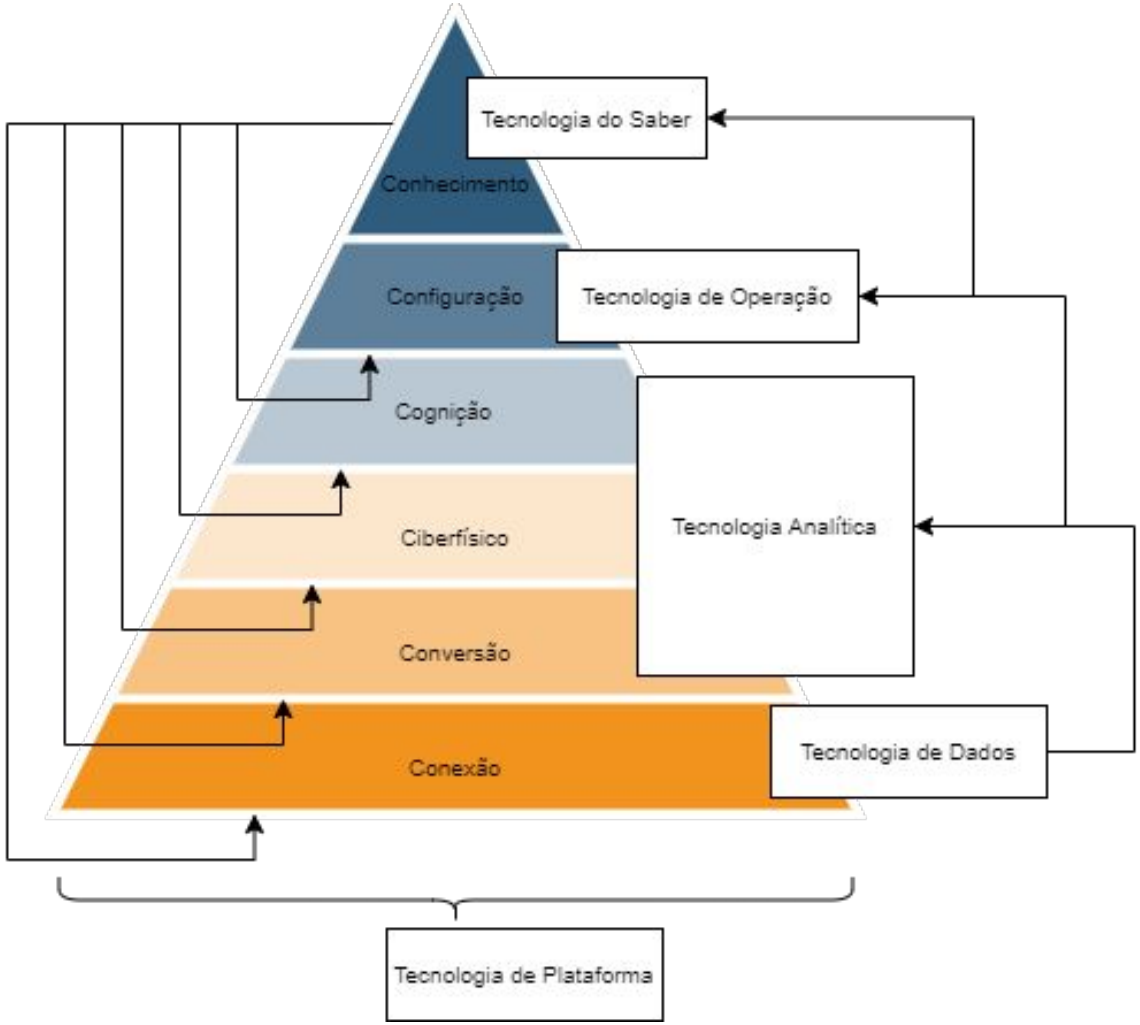


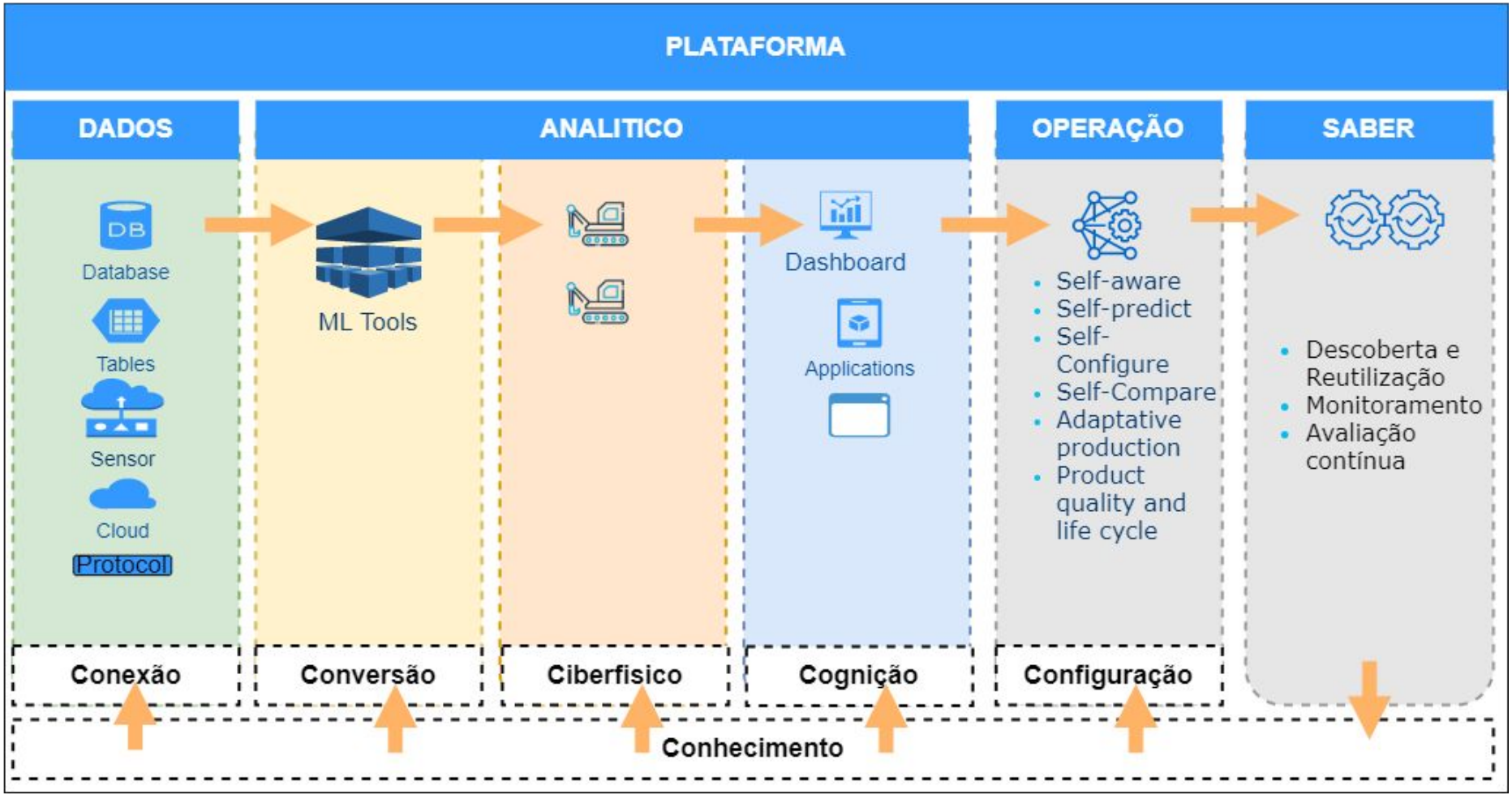
LEE, J.; BAGHERI, B.; KAO, H.-A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manufacturing letters, Elsevier, v. 3, p. 18–23, 2015.

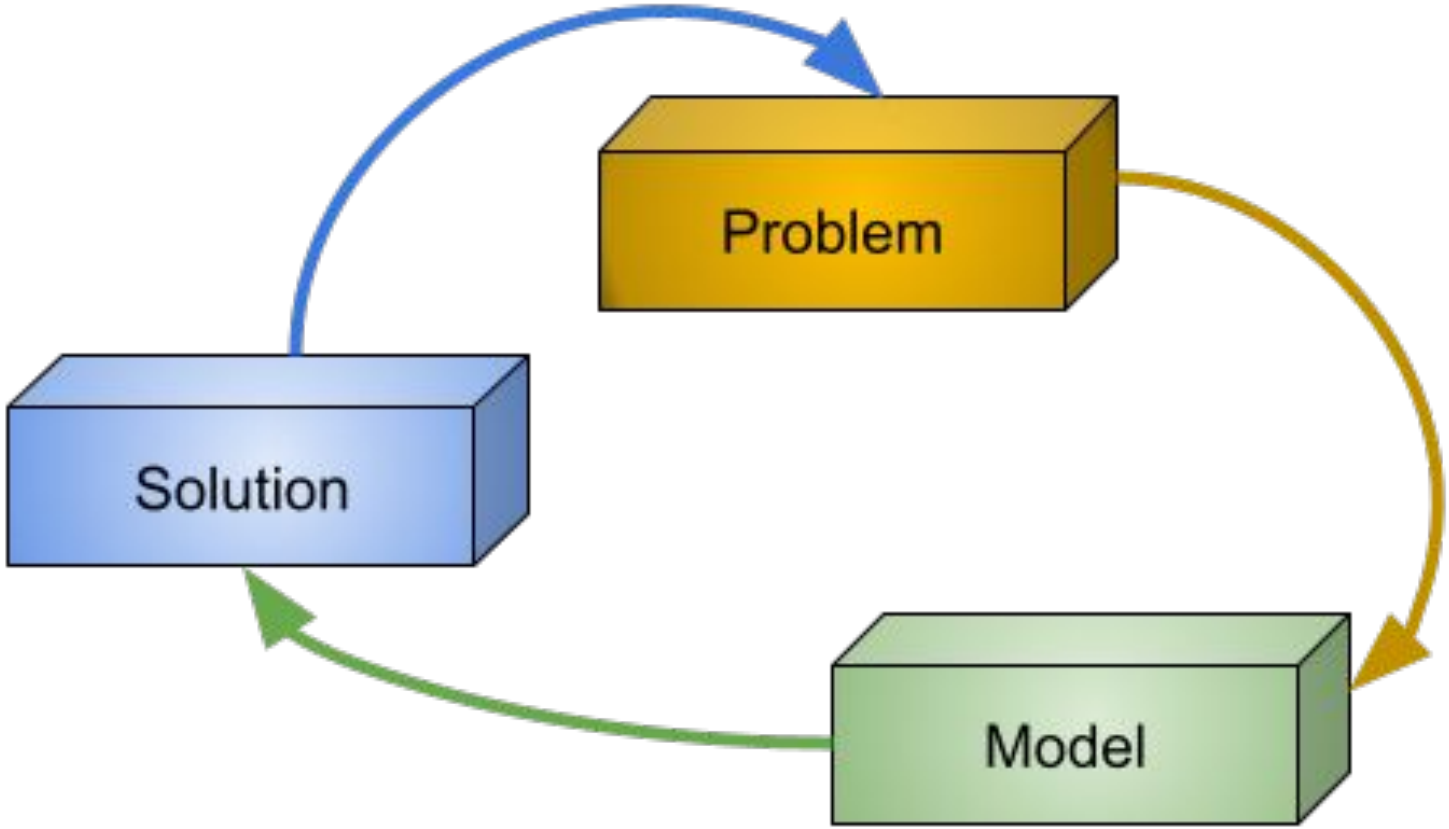
- Arquitetura 5C
  - **Connection:** identifica o equipamento e mecanismo adequados para adquirir dados.
  - **Conversion**: converte os dados para informação significativa.
  - **Cyber**: analisa e simula (Digital Twin) os dados coletados em massa.
  - **Cognition**: aproveita as informações geradas das camadas anteriores para identificar falhas diagnosticar problemas.
  - **Configuration**: atua como controle para tornar as máquinas autoconfiguráveis, tomando decisões corretivas e preventivas.

IA na Indústria - Pipeline




Tecnologia do Saber
Conhecimento
Configuração
Tecnologia de Operação
Cognição
Ciberfísico
Tecnologia Analítica
Conversão
Conexão
Tecnologia de Dados
Tecnologia de Plataforma


PLATAFORMA
DADOS
ANALITICO
OPERAÇÃO
SABER
DB
Database
Tables
Sensor
Cloud
Protocol
ML Tools
Dashboard
Applications
Self-aware
Self-predict
Self-Configure
Self-Compare
Adaptative production
Product quality and life cycle
Descoberta e Reutilização
Monitoramento
Avaliação contínua
Conexão
Conversão
Ciberfisico
Cognição
Configuração
Conhecimento


Problem
Model
Solution

Otimização Matemática

• Pesqui. Oper. 27 (1) • Abr 2007 • https://doi.org/10.1590/S0101-74382007000100009 COPIAR

Um modelo de otimização para o problema integrado de dimensionamento de lotes e programação da produção em fábricas de refrigerantes

Claudio F. M. Toledo Paulo M. França Reinaldo Morabito Alf Kimms SOBRE OS AUTORES

International Journal of Production Research
Vol. 47, No. 11, 1 June 2009, 3097–3119

Taylor & Francis
Taylor & Francis Group

Multi-population genetic algorithm to solve the synchronized and integrated two-level lot sizing and scheduling problem

C.F.M. Toledo[a], P.M. França[b], R. Morabito[c] and A. Kimms[d]*

[a]*Departamento de Ciência da Computação, Universidade Federal de Lavras, C.P. 3037, 37200-000, Lavras, MG, Brazil;* [b]*Departamento de Matemática, Estatística e Computação, Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia, Rua Roberto Simonsen, 305 19060-900, Presidente Prudente, SP, Brazil;* [c]*Departamento de Engenharia de Produção, Universidade Federal de São Carlos, C.P. 676, 13565-905, São Carlos, SP, Brazil;* [d]*Department of Technology and Operations Management, University of Duisburg-Essen, 47048 Duisburg, Germany*

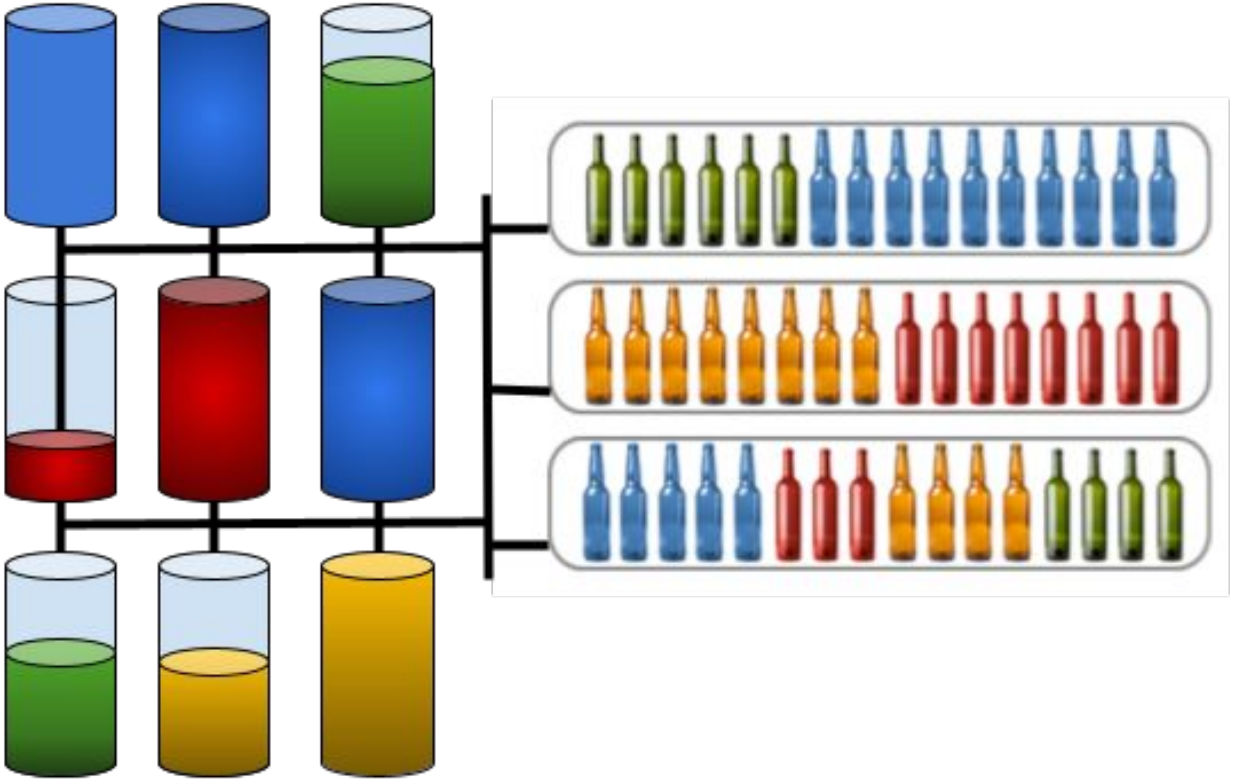
Research Article | Open Access
Volume 2015 | Article ID 182781 | https://doi.org/10.1155/2015/182781

Show citation

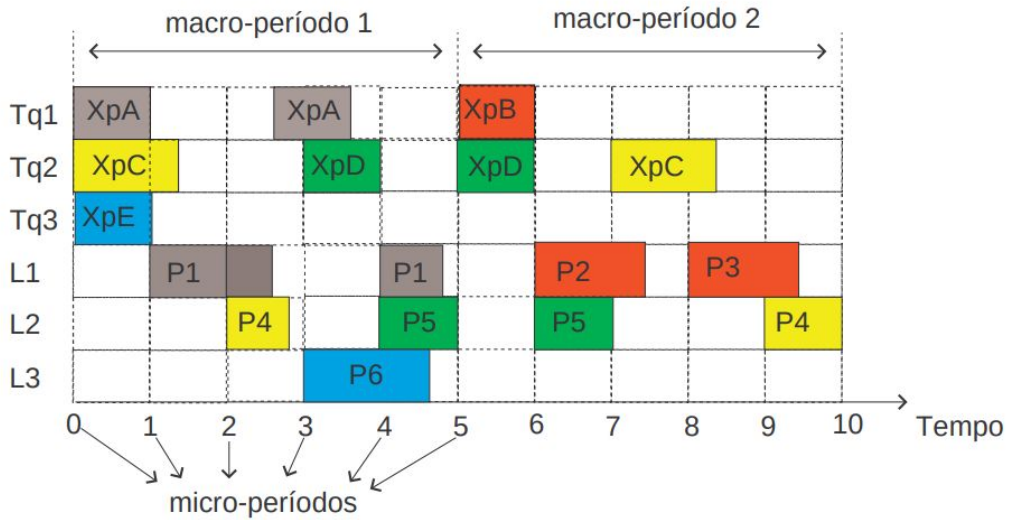
The Synchronized and Integrated Two-Level Lot Sizing and Scheduling Problem: Evaluating the Generalized Mathematical Model

Claudio F. M. Toledo,[1] Alf Kimms,[2] Paulo M. França,[3] and **Reinaldo Morabito**[4]
Show more

Problema Industrial



Modelo Matemáticc



$$p_{jl}q_{jls} \leq Cx_{jls}$$

$$z_{ijls} \geq x_{jls} + x_{il,s-1} - 1$$

$$\sum_{j=1}^{J} p_{jl}q_{jls} \leq Cu_{ls}$$

$$\epsilon u_{ls} \leq \sum_{j=1}^{J} q_{jls}$$

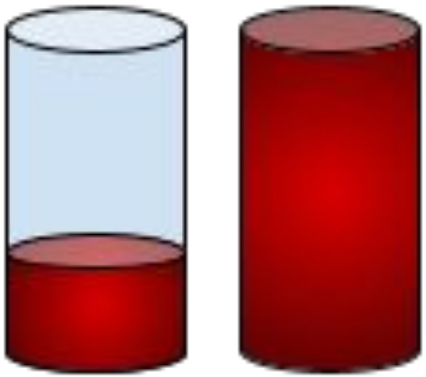
$$u_{ls} \geq u_{l,s+1}$$

q_{jls} Número de produtos j produzidos na linha l no lote s.

x_{jls} = 1, se o lote s na linha l pode ser usado para produzir o produto j; 0, caso contrário.

u_{ls} = 1, se uma quantidade é efetivamente produzida no lote s na linha l; 0, caso contrário.

Modelo Matemático



$$\bar{q}_{jks} \leq \overline{Q}_k \bar{x}_{jks}$$

$$\bar{q}_{jks} \leq \overline{Q}_k \bar{u}_{ks}$$

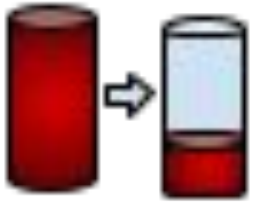
$$\sum_{j=1}^{\bar{J}} \bar{q}_{jks} \geq \underline{Q}_k \bar{u}_{ks}$$

$\bar{q}_{jks}$ Quantidade de xarope j armazenada no tanque k no lote s.

$\bar{x}_{jks}$ = 1, se o lote s pode ser usado para armazenar o xarope j no tanque k; 0, caso contrário.

$\bar{u}_{ks}$ = 1, se lote s é usado efetivamente para armazenar xarope no tanque k; 0, caso contrário.

Modelo Matemático



$$\bar{x}_{jks} - \bar{x}_{jk,s-1} \leq u_{ks}$$

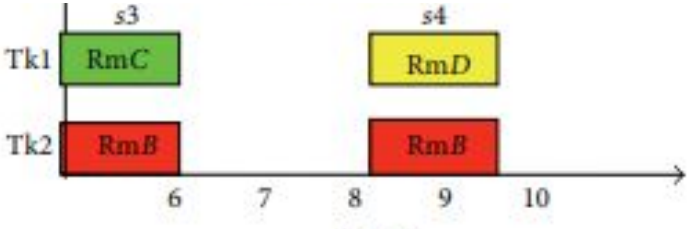
$$\bar{z}_{ijks} \geq \bar{x}_{jks} + \bar{x}_{ik,s-1} - 2 + \bar{u}_{ks}$$

$\bar{x}_{jks}$ = 1, se o lote s pode ser usado para armazenar o xarope j no tanque k; 0, caso contrário.

$\bar{u}_{ks}$ = 1, se lote s é usado efetivamente para armazenar xarope no tanque k; 0, caso contrário.

$\bar{z}_{ijks}$ Indica se o tanque k é ajustado ($\bar{z}_{ijks} = 1$) ou não ($\bar{z}_{ijks} = 0$) a partir do xarope i para o xarope j no início do lote s.

Modelo Matemático



$$\bar{z}_{ijks} \geq \bar{x}_{jks} + \bar{x}_{ik,s-1} - 2 + \bar{u}_{ks}$$

$$\overline{z}_{\mathrm{Rm}B,\mathrm{Rm}B,\mathrm{Tk}2,s4} \geq \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s4} + \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s3} - 2 + \overline{u}_{\mathrm{Tk}2,s4}$$
$$\overline{z}_{\mathrm{Rm}B,\mathrm{Rm}B,\mathrm{Tk}2,s4} \geq 1 + 1 - 2 + 1$$

$$\overline{z}_{\mathrm{Rm}C,\mathrm{Rm}D,\mathrm{Tk}1,s4} \geq \overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s4} + \overline{x}_{\mathrm{Rm}C,\mathrm{Tk}1,s3} - 2 + \overline{u}_{\mathrm{Tk}1,s4}$$
$$\overline{z}_{\mathrm{Rm}C,\mathrm{Rm}D,\mathrm{Tk}1,s4} \geq 1 + 1 - 2 + 1$$

$$\overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s4} = \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s3} = \overline{u}_{\mathrm{Tk}2,s3} = 1$$
$$\overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s4} = \overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s3} = \overline{u}_{\mathrm{Tk}1,s4} = 1$$

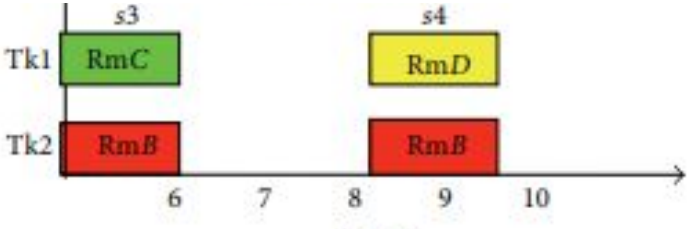
$$\bar{x}_{jks} - \bar{x}_{jk,s-1} \leq u_{ks}$$

$$\overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s4} - \overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s3} \leq \overline{u}_{\mathrm{Tk}1,s4} \Longrightarrow 1 - 0 \leq 1$$

$$\overline{x}_{\mathrm{Rm}C,\mathrm{Tk}1,s4} - \overline{x}_{\mathrm{Rm}C,\mathrm{Tk}1,s3} \leq \overline{u}_{\mathrm{Tk}1,s4} \Longrightarrow 0 - 1 \leq 1$$

$$\overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s4} - \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s3} \leq \overline{u}_{\mathrm{Tk}2,s4} \Longrightarrow 1 - 1 \leq 1$$

Modelo Matemático



$$\bar{z}_{ijks} \geq \bar{x}_{jks} + \bar{x}_{ik,s-1} - 2 + \bar{u}_{ks}$$

$$\overline{z}_{\mathrm{Rm}B,\mathrm{Rm}B,\mathrm{Tk}2,s4} \geq \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s4} + \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s3} - 2 + \overline{u}_{\mathrm{Tk}2,s4}$$
$$\overline{z}_{\mathrm{Rm}B,\mathrm{Rm}B,\mathrm{Tk}2,s4} \geq 1 + 1 - 2 + 1$$

$$\overline{z}_{\mathrm{Rm}C,\mathrm{Rm}D,\mathrm{Tk}1,s4} \geq \overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s4} + \overline{x}_{\mathrm{Rm}C,\mathrm{Tk}1,s3} - 2 + \overline{u}_{\mathrm{Tk}1,s4}$$
$$\overline{z}_{\mathrm{Rm}C,\mathrm{Rm}D,\mathrm{Tk}1,s4} \geq 1 + 1 - 2 + 1$$

$$\overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s4} = \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s3} = \overline{u}_{\mathrm{Tk}2,s3} = 1$$
$$\overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s4} = \overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s3} = \overline{u}_{\mathrm{Tk}1,s4} = 1$$

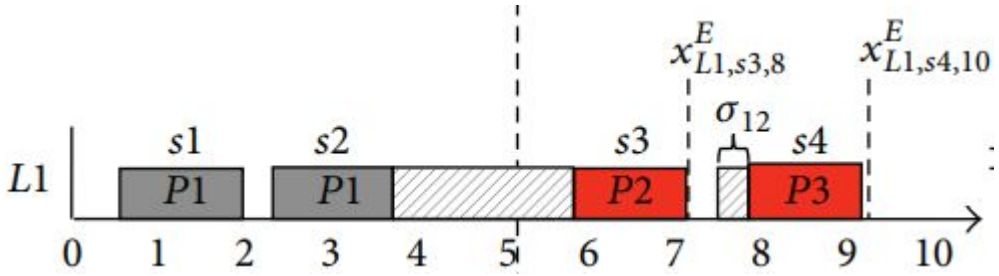
$$\bar{x}_{jks} - \bar{x}_{jk,s-1} \leq u_{ks}$$

$$\overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s4} - \overline{x}_{\mathrm{Rm}D,\mathrm{Tk}1,s3} \leq \overline{u}_{\mathrm{Tk}1,s4} \Longrightarrow 1 - 0 \leq 1$$

$$\overline{x}_{\mathrm{Rm}C,\mathrm{Tk}1,s4} - \overline{x}_{\mathrm{Rm}C,\mathrm{Tk}1,s3} \leq \overline{u}_{\mathrm{Tk}1,s4} \Longrightarrow 0 - 1 \leq 1$$

$$\overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s4} - \overline{x}_{\mathrm{Rm}B,\mathrm{Tk}2,s3} \leq \overline{u}_{\mathrm{Tk}2,s4} \Longrightarrow 1 - 1 \leq 1$$

Modelo Matemático



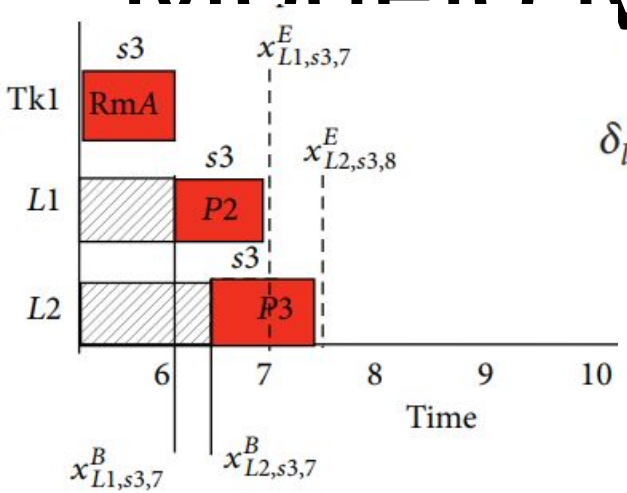
$$\sum_{\tau=(t-1)T^m+1}^{t\cdot T^m} C^m \tau x^E_{ls\tau} - \sum_{\tau=(t-1)T^m+1}^{t\cdot T^m} C^m \tau x^E_{l,s-1,\tau} \geq \sigma_{ls}$$

$$10x^E_{L1,s4,10} - 8x^E_{L1,s3,8} \geq \sigma_{L1,s4} + p_{P3,L1} q_{P3,L1,s4}$$

$x^E_{ls\tau}$ = 1, se o lote s na linha l termina no micro-período τ.

δ_{ls} Tempo no primeiro micro-período do lote que é reservado para tempo de troca e tempo ocioso.

Modelo Matemáti



$$q_{jls} = \sum_{k=1}^{\bar{L}} \sum_{\tau=(t-1)T^m+1}^{t\cdot T^m} q_{kjls\tau}$$

$$\delta_{ls} = \left(\sum_{\tau=(t-1)T^m+1}^{t\cdot T^m} \tau x^E_{ls\tau} - \sum_{\tau=(t-1)T^m+1}^{t\cdot T^m} \tau x^B_{l,s-1,\tau} + 1 \right) C^m - \sum_{j=1}^{J} p_{jl} q_{jls},$$

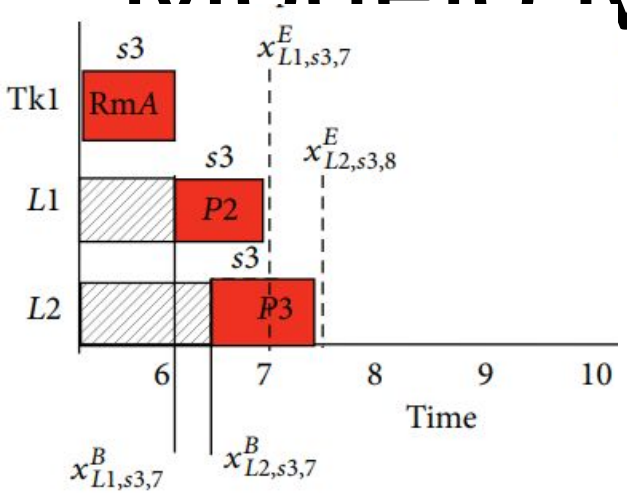
$x^E_{ls\tau}$ = 1, se o lote s na linha l termina no micro-período τ.

δ_{ls} Tempo no primeiro micro-período do lote que é reservado para tempo de troca e tempo ocioso.

$x^B_{ls\tau}$ = 1, se o lote s na linha l começa no macro-período τ.

$q_{kjls\tau}$ Quantidade de produto j, produzido na linha l, no micro-período τ e que pertence ao lote s, que utiliza xarope a partir do tanque k.

Modelo Matemático



$$\sum_{k=1}^{\bar{L}} \sum_{j=1}^{J} p_{jl} q_{kjls\tau} \leq \sum_{\tau'=(t-1)T^m+1}^{t \cdot T^m} C^m x^E_{ls\tau'},$$

$$\sum_{k=1}^{\bar{L}} \sum_{j=1}^{J} p_{jl} q_{kjls\tau} \leq \sum_{\tau'=(t-1)T^m+1}^{t \cdot T^m} C^m x^B_{ls\tau'},$$

$$p_{P2,L1} q_{\mathrm{Tk1},P2,L1,s3,7} \leq x^E_{L1,s3,7}$$
$$p_{P3,L2} q_{\mathrm{Tk1},P3,L2,s3,7} \leq x^E_{L2,s3,8}$$
$$p_{P3,L2} q_{\mathrm{Tk1},P3,L1,s3,8} \leq x^E_{L2,s3,8}$$

$$p_{P2,L1} q_{\mathrm{Tk1},P2,L1,s3,7} \leq x^B_{L1,s3,7}$$
$$p_{P3,L2} q_{\mathrm{Tk1},P3,L2,s3,7} \leq x^B_{L2,s3,7}$$
$$p_{P3,L2} q_{\mathrm{Tk1},P2,L1,s3,8} \leq x^B_{L2,s3,7}$$

$x^E_{ls\tau}$ = 1, se o lote s na linha l termina no micro-período τ.

δ_{ls} Tempo no primeiro micro-período do lote que é reservado para tempo de troca e tempo ocioso.

$x^B_{ls\tau}$ = 1, se o lote s na linha l começa no macro-período τ.

$q_{kjls\tau}$ Quantidade de produto j, produzido na linha l, no micro-período τ e que pertence ao lote s, que utiliza xarope a partir do tanque k.

Modelo Matemático

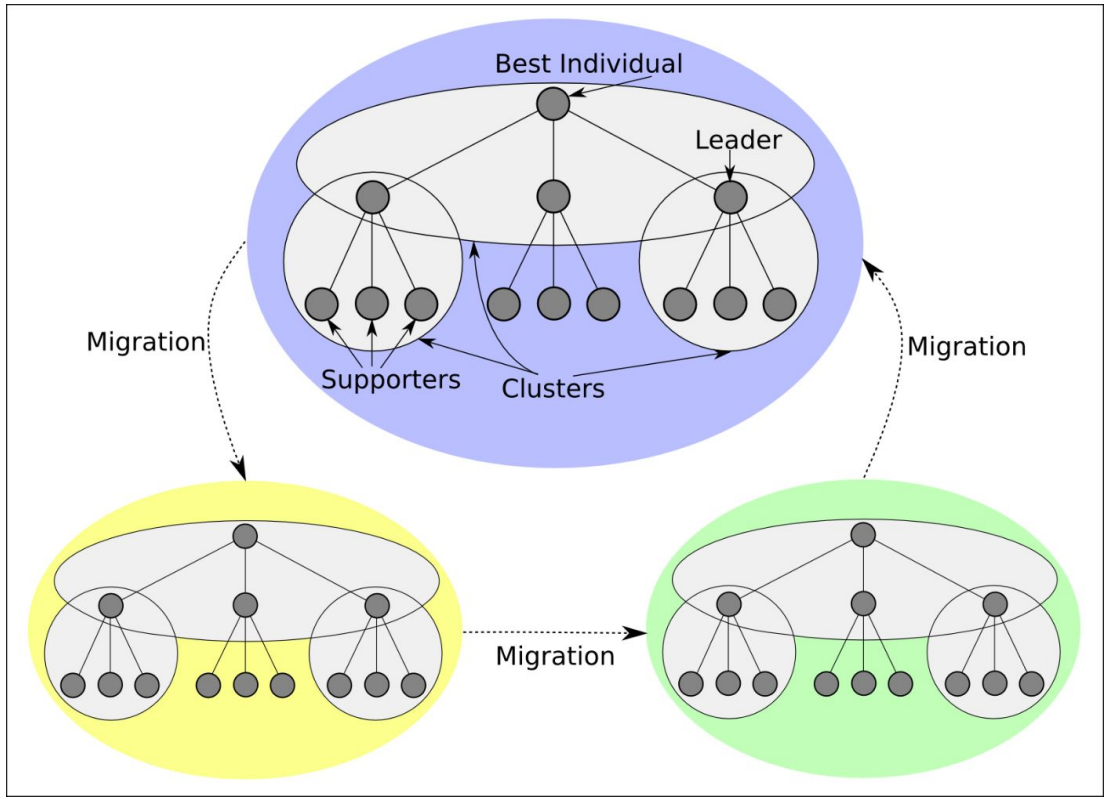
Instâncias Soluções Ótimas

L/Tk/J/Xp	T1	T2	T3	T4
2/2/2/1	10	10	2	1
2/2/3/2	10	6	3	1
2/2/4/2	10	6	3	0
3/2/2/1	10	1	0	0
3/2/3/2	10	0	0	0
3/2/4/2	10	1	0	0
4/2/2/1	2	0	1	0
4/2/3/2	8	1	0	0
4/2/4/2	7	0	0	0

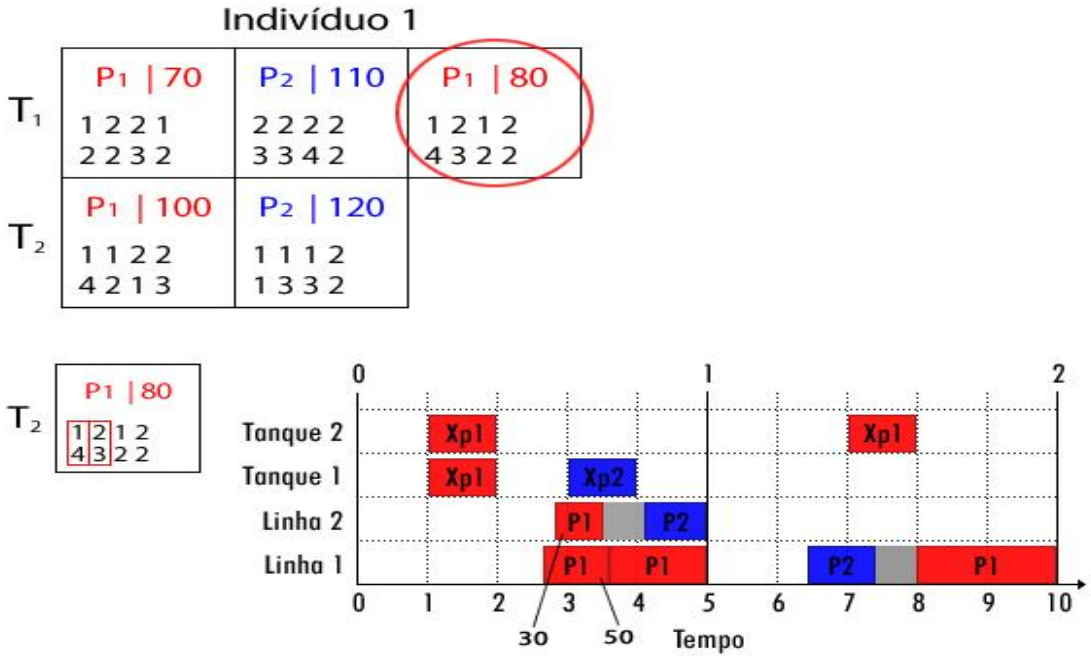
L: Linhas
Tk: Tanques
J: Produtos
Xp: Xarope

- Dificuldade para retornar soluções ótimas quando o número de parâmetros de entrada cresce.
- Dificuldade para retornar soluções factíveis para combinações do tipo:
  - 10 produtos, 5 tanques e 4 macro-períodos
  - 15 produtos, 8 tanques e 4 macro-períodos
  - 10 produtos, 5 tanques e 8 macro-períodos
  - 15 produtos, 8 tanques e 8 macro-períodos
  - 10 produtos, 5 tanques e 12 macro-períodos
  - 15 produtos, 8 tanques e 12 macro-períodos

Algoritmo Evolutivo



Codificação e Decodificação

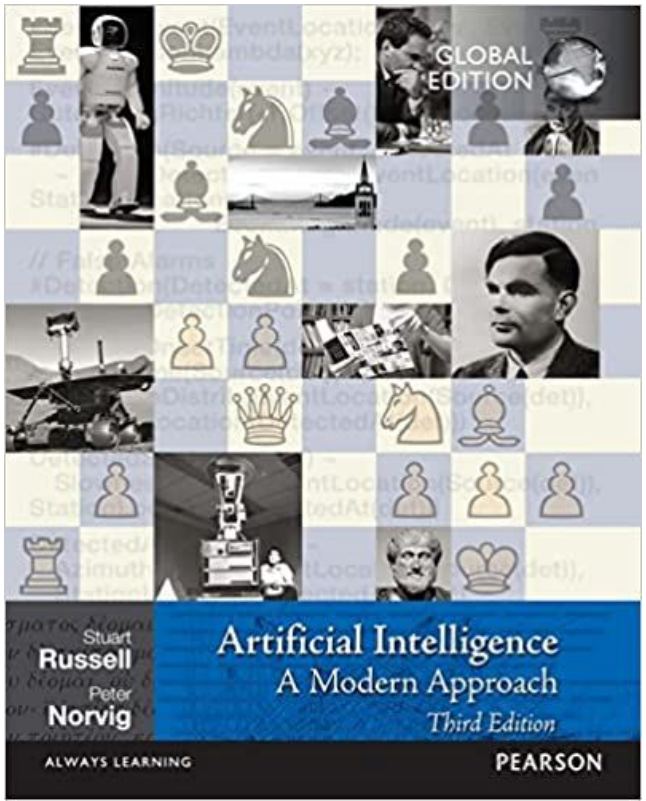


Resultados

Comb.	$L/\bar{L}/J/\bar{J}/T$	Comb.	$L/\bar{L}/J/\bar{J}/T$
A1	5/9/33/11/1	B1	6/10/52/19/1
A2	6/9/49/14/2	B2	6/10/56/19/2
A3	6/9/58/15/3	B3	6/10/65/21/3

Comb.	Solutions				Deviation – Dev(%)		
	Z^I	GA–Min	GA-Avg	GA-Max	GA-Min	GA-Avg	GA-Max
A1	1692098	1662098	1667098	1671098	-1.8	-1.5	-1.2
A2	3511910	3381357	3388059	3398510	-3.7	-3.5	-3.2
A3	5002677	4837433	4847207	4859924	-3.3	-3.1	-2.9
B1	3378205	3303500	3317500	3345500	-2.2	-1.8	-1.0
B2	4278521	4174522	4199463	4222860	-2.4	-1.8	-1.3
B3	7943402	7735818	7796636	7839039	-2.6	-1.8	-1.3

Autonomy



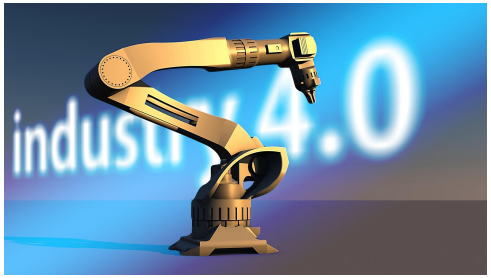
> “A rational agent should be autonomous—it should learn what it can to compensate for partial or incorrect prior knowledge.”

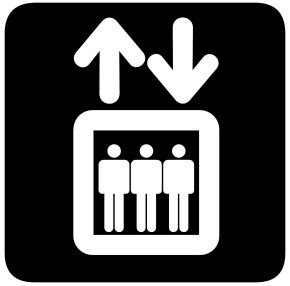
Situational Awareness

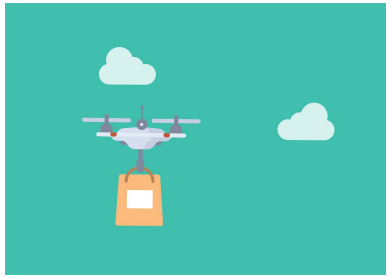
Perception of elements in the environment within a volume of time and space, understanding their meaning and projecting their condition in the near future. (ENDSLEY, 1988)

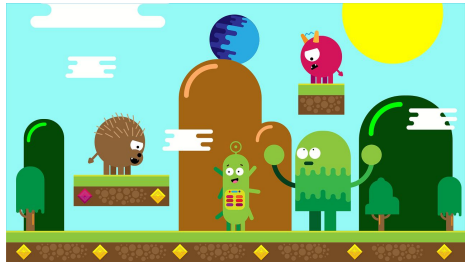
Endsley, Mica R. "Design and evaluation for situation awareness enhancement." Proceedings of the Human Factors Society annual meeting. Vol. 32. No. 2. Sage CA: Los Angeles, CA: Sage Publications, 1988.

Planning Domain Definition Language and Robot Operating System









PDDL

Planner 1

Planner 2

Planner 3

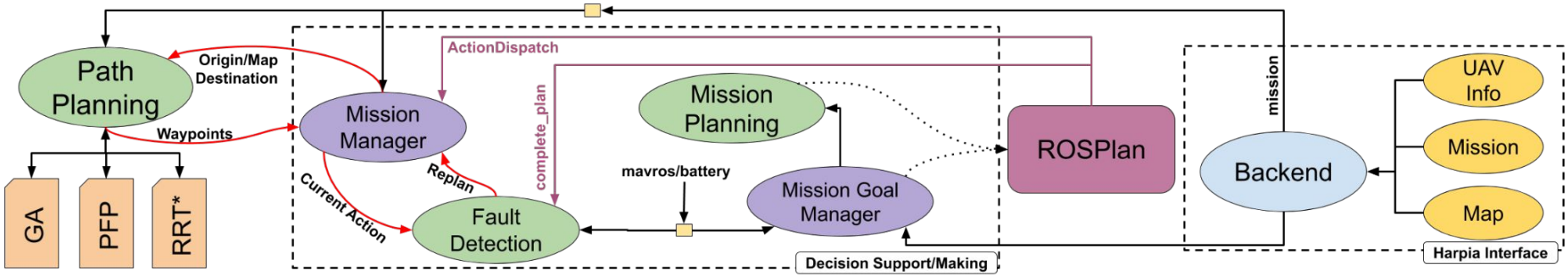
Planning Domain Definition Language (PDDL)

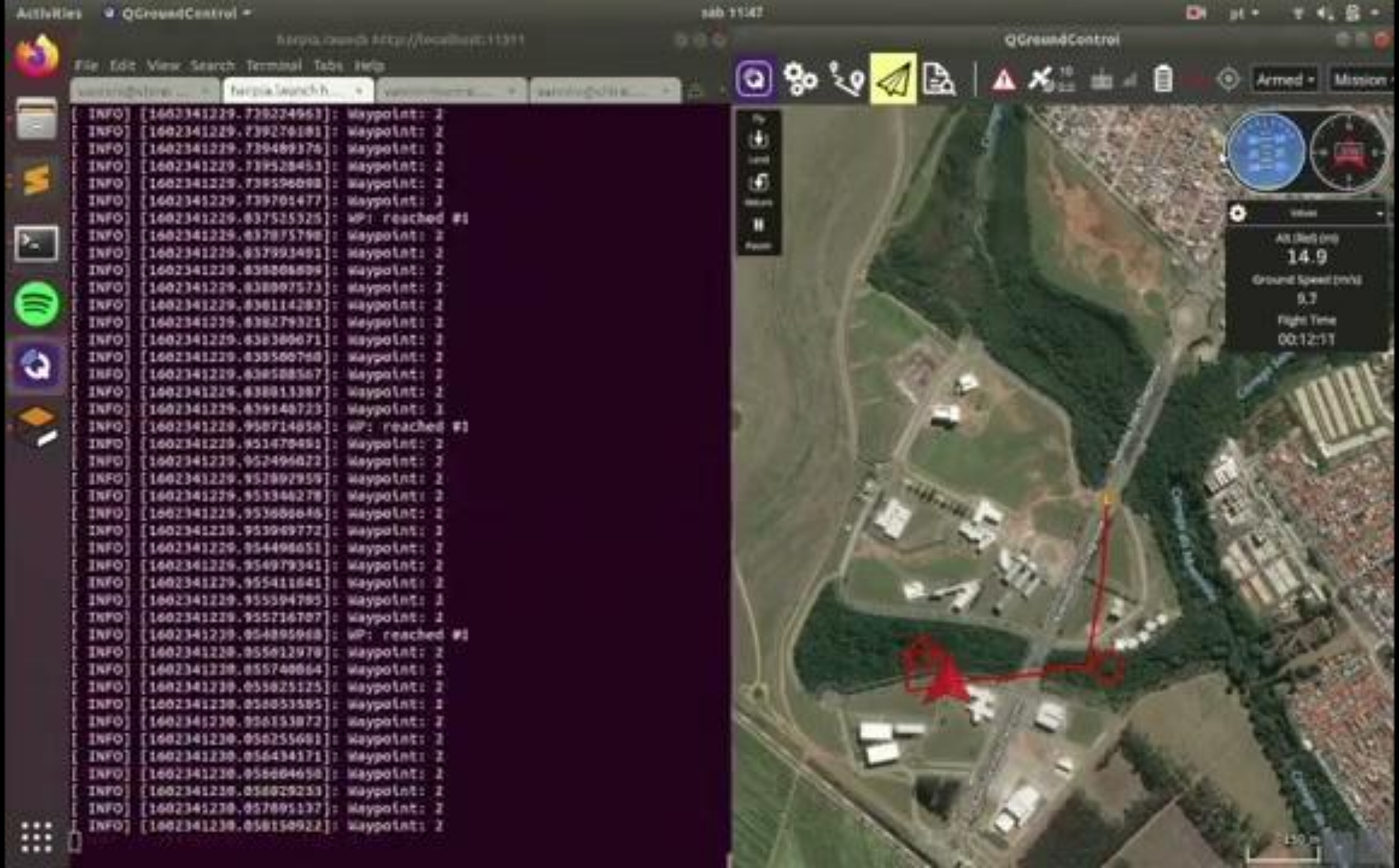
```
(define (domain harpia)
        (:requirements ...
        (:types region - object
                base  - region)
        (:functions
                (battery-amount)
                (distance ?from-region - region ?to-region -region) ; ; m
                (discharge-rate-battery) ; ; % / s
        ...)
        (:predicates
                (at ?region - region)
                (taken-image ?region - region)
                (picture-goal ?region - region)
        ...)
        (:durative-action go_to_picture ...
        (:durative-action go_to_pulverize ...
        (:durative-action go_to_base ...
        (:durative-action recharge_input ...
        (:durative-action pulverize_region ...
        (:durative-action take_imate ...
        (:durative-action recharge_battery ...
)
```

ICMC USP SÃO CARLOS

Robot Operating System

System Architecture




Activities
QGroundControl
sab 11:47
File Edit View Search Terminal Tabs Help
[INFO] [1602341229.738224963]: Waypoint: 2
[INFO] [1602341229.739270101]: Waypoint: 2
[INFO] [1602341229.739489376]: Waypoint: 2
[INFO] [1602341229.739528453]: Waypoint: 2
[INFO] [1602341229.739596008]: Waypoint: 2
[INFO] [1602341229.739701477]: Waypoint: 3
[INFO] [1602341229.837525525]: WP: reached #1
[INFO] [1602341229.837875798]: Waypoint: 2
[INFO] [1602341229.837993491]: Waypoint: 2
[INFO] [1602341229.838086899]: Waypoint: 2
[INFO] [1602341229.838097573]: Waypoint: 3
[INFO] [1602341229.838114283]: Waypoint: 2
[INFO] [1602341229.838279321]: Waypoint: 3
[INFO] [1602341229.838380671]: Waypoint: 3
[INFO] [1602341229.838580768]: Waypoint: 3
[INFO] [1602341229.838588567]: Waypoint: 3
[INFO] [1602341229.838813387]: Waypoint: 2
[INFO] [1602341229.839140723]: Waypoint: 3
[INFO] [1602341229.950714050]: WP: reached #1
[INFO] [1602341229.951470491]: Waypoint: 2
[INFO] [1602341229.952496822]: Waypoint: 2
[INFO] [1602341229.952892959]: Waypoint: 2
[INFO] [1602341229.953346278]: Waypoint: 2
[INFO] [1602341229.953686646]: Waypoint: 2
[INFO] [1602341229.953969772]: Waypoint: 3
[INFO] [1602341229.954498651]: Waypoint: 2
[INFO] [1602341229.954979341]: Waypoint: 2
[INFO] [1602341229.955411641]: Waypoint: 2
[INFO] [1602341229.955594785]: Waypoint: 3
[INFO] [1602341229.955716707]: Waypoint: 2
[INFO] [1602341230.054895968]: WP: reached #1
[INFO] [1602341230.055012970]: Waypoint: 2
[INFO] [1602341230.055740864]: Waypoint: 2
[INFO] [1602341230.056025125]: Waypoint: 2
[INFO] [1602341230.056053595]: Waypoint: 3
[INFO] [1602341230.056153872]: Waypoint: 2
[INFO] [1602341230.056255681]: Waypoint: 3
[INFO] [1602341230.056434171]: Waypoint: 2
[INFO] [1602341230.056604650]: Waypoint: 2
[INFO] [1602341230.056829253]: Waypoint: 2
[INFO] [1602341230.057095137]: Waypoint: 2
[INFO] [1602341230.058150922]: Waypoint: 2
QGroundControl
Armed
Mission
14.9
Ground Speed (m/s)
9.7
Flight Time
00:12:11
150 m

Análise de Dados

Conjuntos de Dados, tipos, escala e exploração de dados

Conjuntos de dados

- Matematicamente pode se descrito como:

$$X = [x_{i,j}] \qquad 1 \leq i \leq n \text{ e } 1 \leq j \leq d$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix}$$

n : número de objetos

d : dimensionalidade dos objetos

x_{ij} : valor da j-ésima característica do i-ésimo objeto

- *d* também é chamado de dimensionalidade do **espaço de objetos/espaço de entradas/espaço de atributos**.

Tabela 2.1 Conjunto de dados `hospital` com seus atributos

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Fonte: Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. D., & Carvalho, A. C. P. D. L. F. D. (2021). Inteligência artificial: uma abordagem de aprendizado de máquina.

Atributo Alvo

- **Atributo alvo** define o que se pretende prever.
- Podem identificar as **classes** ou categorias de pertencimento dos **objetos**, assumindo valores discretos {1,2,…n} e executando uma classificação dos objetos.
  - **Classe majoritária**: aquela classe k ∈{1,2,...n} com maior número de objeto,
  - **Classe minoritária**: aquela classe k ∈{1,2,...n} com menor número de objeto
  - Uma **classificação binária** ocorre quando temos apenas k ∈{1,2}
- Um problema de regressão ocorre quando o atributo alvo é numérico e contínuo.

Atributo Alvo

- **Problema de regressão**: capturar a correlação entre variáveis observadas em um conjunto de dados, quantificando se tais correlações são estatisticamente significativas ou não.
- Há métodos de regressão linear simples, regressão linear multivariada e regressão não linear para conjuntos de dados que demandam análises complexas.
- A previsão de valores em **séries temporais** é um caso de regressão onde há uma relação de periodicidade entre os valores tratados.

Atributo Alvo

Regressão linear simples: $Y = a + bX + u$

Regressão linear multivariada:

$Y = b + a_1 X_1 + a_2 X_2 + \ldots + a_d X_d + u$

$onde,$

Y : Variável dependente a ser predita, classificada ou explicada.

X ou X_i : Variável independente usada na determinação de Y.

b : Coeficiente linear

a ou a_i : Coeficiente angular da variável independente

u : regressão residual ou erro

Tipo de Atributo: Qualitativo ou Quantitativos

- **Atributos Qualitativos**
  - Também chamados de atributos simbólicos ou categóricos.
  - $x \in$ {pequeno, médio, grande} ou $x \in$ { {matemática, física, química}
  - Podem ser **ordenados em alguns casos** mas **não** se consegue aplicar **operações aritméticas** aos seus valores.
- **Atributos Quantitativos**
  - Os atributos quantitativos possuem valores numéricos.
  - Discretos: $x \in \{23, 45, 12\}$ ou contínuos: $x \in [0,1]$.
  - Podem ser **ordenados** e as **operações aritméticas** podem ser aplicadas.
  - Um caso especial são os atributos booleanos ou binários com $x \in \{0, 1\}$, representando falso ou verdadeiro, não ou sim, ausência ou presença, sucesso ou falha, etc.

Tabela 2.2 Tipo dos atributos do conjunto `hospital`

Atributo	Classificação
Id.	Qualitativo
Nome	Qualitativo
Idade	Quantitativo discreto
Sexo	Qualitativo
Peso	Quantitativo contínuo
Manchas	Qualitativo

Fonte: Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. D., & Carvalho, A. C. P. D. L. F. D. (2021). Inteligência artificial: uma abordagem de aprendizado de máquina.

Tipo de Atributo: Qualitativo ou Quantitativos

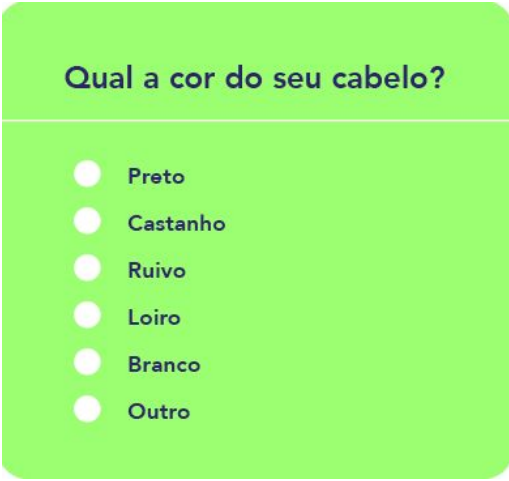
- **Atributos quantitativos** podem ter valores binários, inteiros ou reais.
- **Atributos qualitativos** são representados por uma **quantidade finita** de símbolos, nomes ou rótulos.
- **Atributos qualitativos** podem ser representados por números que indiquem, por exemplo, um Id.
- **Você não vai aplicar uma operação aritméticas sobre um Id, certo?**

Escala

- Estabelece as operações possíveis de serem aplicadas sobre o valor do atributo
- Há quatro escalas de estimativa de informação na estatística: nominal, ordinal, intervalo e racional (razão)
- **Nominais** e **ordinais** são do **tipo qualitativo**.
- **Intervalares** e **racionais** são do **tipo quantitativo**.

Escala

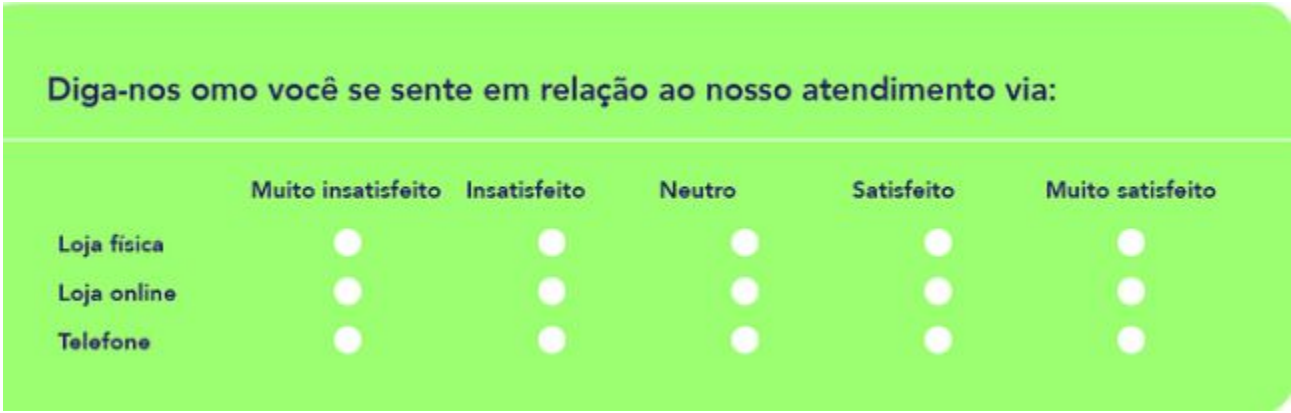
- A escala nominal estabelece rótulos ou nomes diferentes com a menor quantidade de informação possível.
  - Permite rotular atributos em classificações distintas, **sem qualquer significado numérico ou valor quantitativo**.
  - **Não** permite estabelecer uma r**elação de ordem** entre seus valores.
  - Algumas operações possíveis: = e ≠
  - Exemplo: {febre, tosse, coriza} ou {Mac, Linux, Windows}



Fonte: Conheça os 4 níveis de escala de medição

Escala

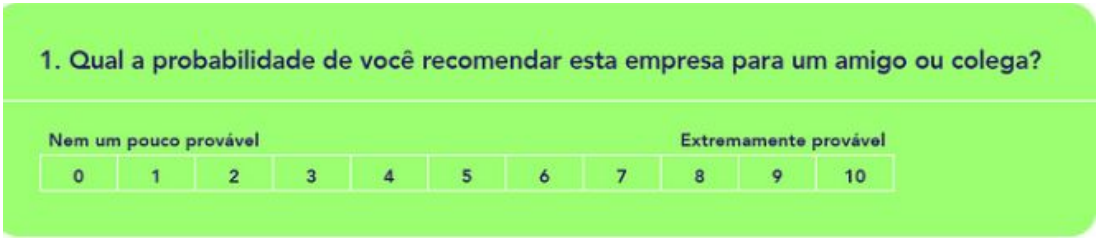
- A escala ordinal estabelece valores que permitem ordenar categorias ou classes representadas.
  - Permite **identificar** ou **categorizar** através de **comparações**.
  - Algumas operações possíveis: **=, ≠, <, >, ≤, ≥** podem ser utilizados.
  - Exemplo: {pequeno, médio,grande} ou {frio,morno,quente}



Fonte: Conheça os 4 níveis de escala de medição

Escala

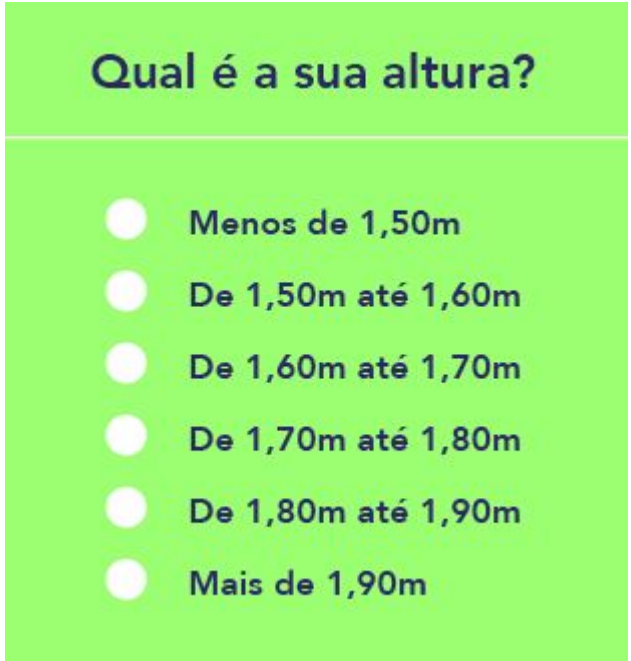
- A escala intervalar estabelece valores numéricos que pertencem a um intervalo.
  - **Evolução da escala ordinal** ao incorporar todas as suas propriedades.
  - Permite avaliar quanto determinados atributos estão distantes entre si em relação a determinada característica.
  - Algumas operações possíveis: **aritméticas, ordenamento, diferença de magnitude** entre dois valores (distância entre eles).



Fonte: Conheça os 4 níveis de escala de medição

Escala

- A escala racional estabelece valores numéricos com um significado absoluto.
  - Inclui as características da escala intervalar com a adição de que **não há valor numérico negativo**.
  - Temos um valor de “zero” como absoluto.
  - Permite **comparar proporções** ou medidas absolutas
  - Permite calcular medidas de tendência central como média, mediana, moda, etc.



Fonte: Conheça os 4 níveis de escala de medição

Tabela 2.3 Escala dos atributos do conjunto `hospital`

Atributo	Classificação
Id.	Nominal
Nome	Nominal
Idade	Racional
Sexo	Nominal
Peso	Racional
Manchas	Nominal
Temp.	Intervalar
#Int.	Racional

Fonte: Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. D., & Carvalho, A. C. P. D. L. F. D. (2021). Inteligência artificial: uma abordagem de aprendizado de máquina.

Explorando um Conjunto de Dados

- A estatística descritiva permite extrair de forma quantitativa as principais características de um conjunto de dados.
- Utiliza medidas que são calculadas rapidamente como a média e desvio padrão.
- Devemos considerar conjuntos com dados que possuem um único atributo (**dados univariados**) ou vários atributos (**dados multivariados**).

Dados Univariados

- Suponho um objeto x_i que apresenta um único atributo. Teremos um conjunto $X^j=\{x_1, x_2, ..., x_n\}$ com n objetos.
- Exemplo: uma tabela onde há uma única coluna com o peso (atributo do conjunto de dados) em que cada x_i representa o peso de uma pessoa (objeto).

Peso
79
67
92
43
52
72
87
67

Dados Univariados

- Vamos avaliar algumas medidas para amostras de dados univariados:
  - Medidas de localidade
  - Medidas de espalhamento
  - Medidas de dispersão

Peso
79
67
92
43
52
72
87
67

Medidas de Localidade

- Estatísticas capazes de extrair informações do conjunto de dados que indiquem pontos de referências da distribuição de dados.
- Métricas:
  - Moda
  - Média
  - Mediana
  - Quartis e Percentis

Medidas de Localidade

- **Moda:** maior frequência para o valor de um atributo e costuma ser aplicada a dados simbólicos

Manchas
Concentradas
Inexistentes
Espalhadas
Inexistentes
Uniformes
Inexistentes
Espalhadas
Uniformes

Est.
SP
MG
RS
MG
PE
RJ
AM
GO

Medidas de Localidade

- Seja X={x_1, x_2, ..., x_n} um conjunto de dados com n objetos, não necessariamente ordenados
- Média, mediana e percentil são mais utilizadas para atributos numéricos
- **Média**: temos o valor médio desse conjunto dado pela expressão abaixo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  - **Vantagem**: boa medida para encontrar o meio de um conjunto de valores mas apenas se estiverem distribuídos simetricamente.
  - **Desvantagem**: outliers.
  - Problema é reduzido usando a mediana.

Medidas de Localidade

- **Mediana:** Deve-se ordenar de forma crescente o conjunto de valores, aplicando em seguida a expressão abaixo para o cálculo da mediana.

$$mediana(X) = \begin{cases} \frac{1}{2}(x_i + x_{i+1}) & \text{Se n =2i} \\ x_{i+1} & \text{Se n=2i+1} \end{cases}$$

Exemplos:

- {17, 4, 8, 21, 4} $\Longrightarrow$ (4, 4, 8, 17, 21) $\Longrightarrow$ 8.
- {17, 4, 8, 21, 4, 15, 13, 9} $\Longrightarrow$ (4, 4, 8, 9, 13, 15, 17, 21) $\Longrightarrow$ (9 + 13)/2 = 11.

Medidas de Localidade

- Considere o conjunto de dados

  X={0, 70, 70, 80, 85, 90, 90, 90, 95, 100}, temos outlier = 0.

  - Moda = **90**
  - Média = **77** (Média sem o outlier = 85,5)
  - Mediana = (85+90)/2 = **87,5**

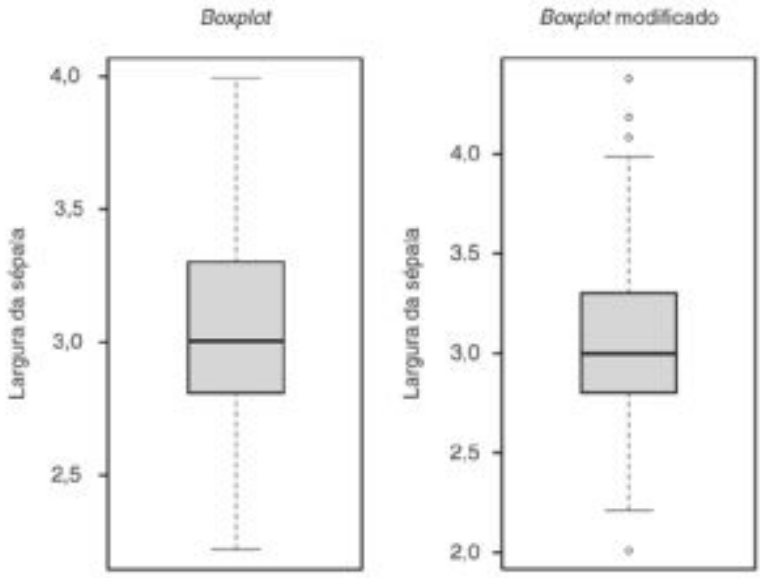
Medidas de Localidade

- **Quartis** e **percentis** também são utilizados para valores numéricos ordenados.
- Os chamados **quartis** dividem em quartos ao invés de metades:
  - **1º quartil (Q1) e 25º percentil (P25%)**: indicam que 25% dos valores estão abaixo dele.
  - **2º quartil (Q2) e 50º percentil (P50%):** é a mediana.
  - **3º quartil (Q3) e 75º percentil (P75%)**: indicam que 75% dos valores estão abaixo dele.

Medidas de Localidade

Algoritmo para o cálculo do percentil

```
1  Ordenar os n valores em ordem crescente
2  Calcular o produto np
3  se np não for um número inteiro  então
4       Arredondar para o próximo inteiro
5       Retornar o valor dessa posição na sequência
6  fim
7  senão
8       Considerar np = k
9       Retornar a média entre os valores nas posições k e k + 1
10 fim
```



Fonte: Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. D., & Carvalho, A. C. P. D. L. F. D. (2021). Inteligência artificial: uma abordagem de aprendizado de máquina.

Medidas de Espalhamento

- Medem a dispersão ou espalhamento de um conjunto de valores, indicando se estão amplamente espalhados ou relativamente concentrados em torno de um valor como a média, por exemplo.
- Métricas:
  - Intervalo;
  - Variância;
  - Desvio padrão.

Medidas de Espalhamento

- Considere novamente $X=\{x_1, x_2, ..., x_n\}$ como um conjunto de dados com n objetos, não necessariamente ordenados
- **Intervalo**: métrica simples que indica a dispersão máxima entre os valores do conjunto de dados:

$$Intervalo(X) = \max_{i=1...n}(x_i) - \min_{i=1...n}(x_i)$$

- Desvantagem: quando há um **reduzido número de valores extremos** e **alta concentração** de valores em torno de um ponto.

Medidas de Espalhamento

- **Variância**: considera o desvio quadrático em relação à média como segue

$$\sigma^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Logo, a variância também sofre efeito de outliers.
- O denominador *n – 1* é chamado correção de Bessel, aplicado para melhor estimar a variância em uma amostra de dados.
- **Desvio Padrão**: considera a raiz quadrada da variância.

$$\sigma(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Medidas de Espalhamento

- Outras medidas de espalhamento:
  - Desvio médio absoluto (AAD - Absolute Average Deviation):

$$AAD(X) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

  - Desvio mediano absoluto (MAD - Median Absolute Deviation):

$$MAD(X) = mediana(\{|x_1 - \bar{x}|, \ldots, |x_n - \bar{x}|\})$$

  - Intervalo interquartil (IQR - Interquartil Range):

$$IQR(X) = Q_3 - Q_1 = P_{75\%} - P_{25\%}$$

Medidas de Distribuição

- **Momento**: métrica definida em torno da média de um conjunto de valores.
- De forma geral, o momento amostral central é definido como:

$$\mathbb{E}_k(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

- **Média:** trata-se do primeiro **Momento em Relação à Origem**

$$\mathbb{E}_0(X) = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

- **Primeiro Momento em Relação à Média:** tem valor nulo

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \bar{x} = \bar{x} - \frac{1}{n} n\bar{x} = \bar{x} - \bar{x} = 0$$

Medidas de Distribuição

- **Variância:** trata-se do **Segundo Momento**

$$\mathbb{E}_2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sigma^2(X)$$

- **Terceiro Momento** permite calcular o **coeficiente de assimetria ou obliquidade**.

$$\mathbb{E}_3(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

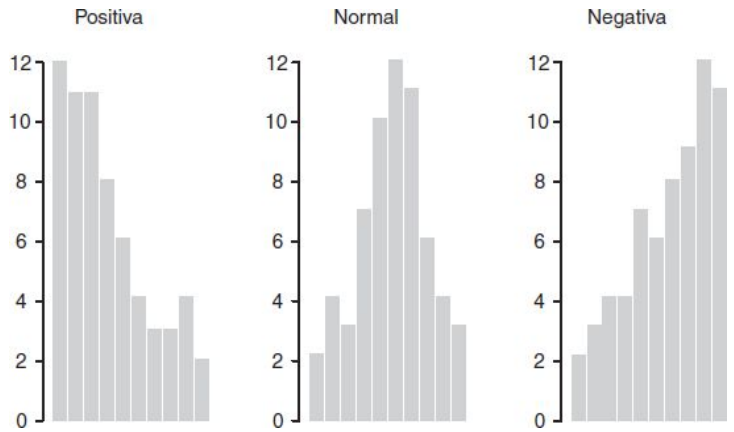
- **Obliquidade**: mede a simetria da distribuição dos dados em relação à média.

$$v = \frac{\mathbb{E}_3(X)}{\sigma^3}$$

- Divide-se pelo cubo do desvio padrão para deixar a métrica **independente de escala**.

Medidas de Distribuição

- O **obliquidade** indica a distribuição dos valores em um conjunto de dados como segue
  - Simétrica parar *v* = 0
  - Assimétrica à esquerda para v>0
  - Assimétrica à direita para v<0



Fonte: Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. D., & Carvalho, A. C. P. D. L. F. D. (2021). Inteligência artificial: uma abordagem de aprendizado de máquina.

Medidas de Distribuição

- **Quarto Momento** permite calcular o coeficiente de curtose.

$$\mathbb{E}_4(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

- **Curtose**: mede a dispersão através do achatamento da curva da função de distribuição dos valores.

$$c = \frac{\mathbb{E}_4(X)}{\sigma^4}$$

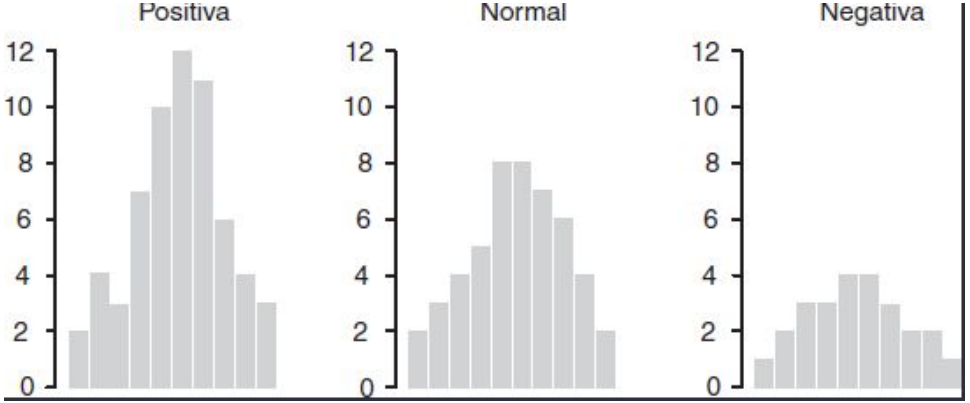
- Divide-se pelo desvio padrão elevado à quarta potência para deixar a métrica independente de escala.

- Temos c = 3 para uma distribuição normal com média 0 e variância 1, levando ao uso da fórmula com uma correção que permite a distribuição normal padrão ter curtose igual a 0.

$$c = \frac{\mathbb{E}_4(X)}{\sigma^4} - 3$$

Medidas de Distribuição

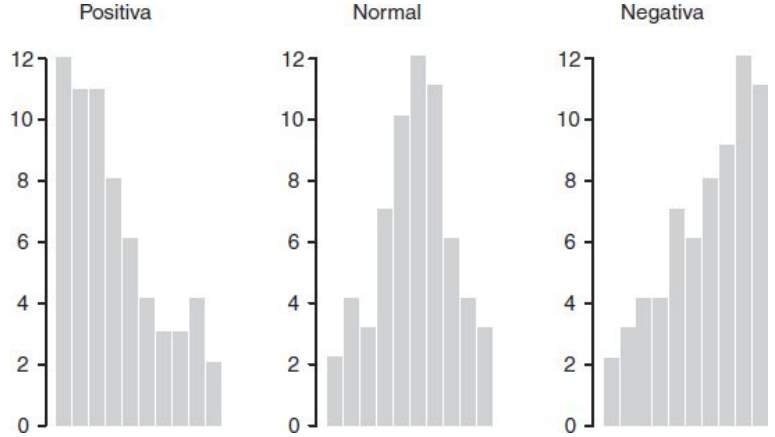
- c=0: histograma da distribuição dos dados com mesmo achatamento que uma distribuição normal.
- c>0: histograma de distribuição mais alto e concentrado do que a distribuição normal.
- c<0: histograma de distribuição mais achatado que a distribuição normal.



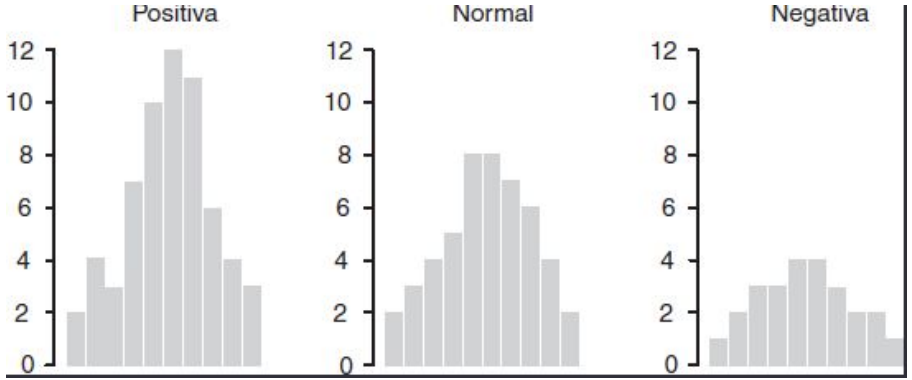
Fonte: Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. D., & Carvalho, A. C. P. D. L. F. D. (2021). Inteligência artificial: uma abordagem de aprendizado de máquina.

Medidas de Distribuição

- Média e desvio padrão são momentos que fornecem medidas de localidade e espalhamento, respectivamente.
- Obliquidade e curtose, terceiro e quarto momentos, fornecem medidas de distribuição dos valores.



Obliquidade



Curtose

Dados Multivariados

- Apresentam mais de um atributo de entrada
- Vamos considerar novamente a definição de **conjunto de dados**

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix}$$

n : número de objetos

d : dimensionalidade dos objetos

x_{ij} : valor da j-ésima característica do i-ésimo objeto

Dados Multivariados

- Tais atributos podem ser avaliados separadamente para se obter as medidas de localidade ou medidas de espalhamentos já definidas.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix} \Rightarrow x^j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \Rightarrow X = [x^1, \dots, x^d]$$

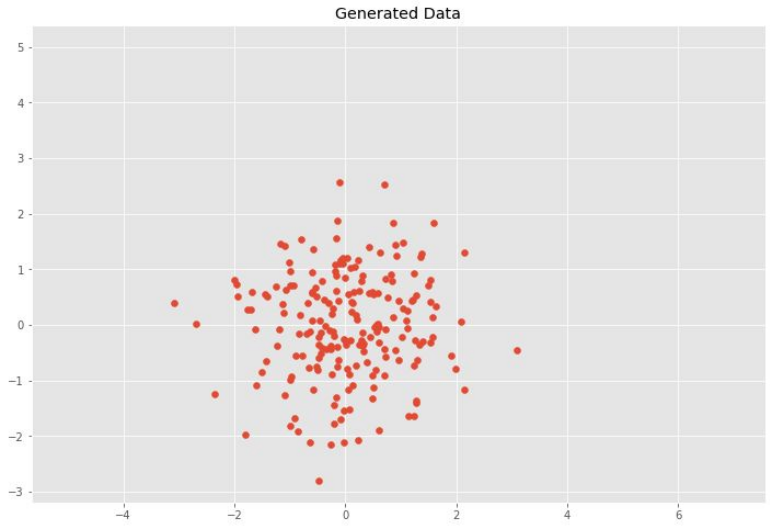
$$X = [x^1, \dots, x^d] \Rightarrow [\bar{x}^1, \dots, \bar{x}^d] \text{ ou } [\sigma_1^2, \dots, \sigma_d^2]$$

Dados Multivariados

- A análise de dados multivariados busca também estabelecer o relacionamento entre atributos.
- O espalhamento de um conjunto de dados multivariados é melhor aferido através da **covariância**
- **Covariância** mede a variância conjunta entre pares de variáveis aleatórias.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \Rightarrow x^p = \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix} \quad x^q = \begin{bmatrix} x_{1q} \\ x_{2q} \\ \vdots \\ x_{nq} \end{bmatrix}$$

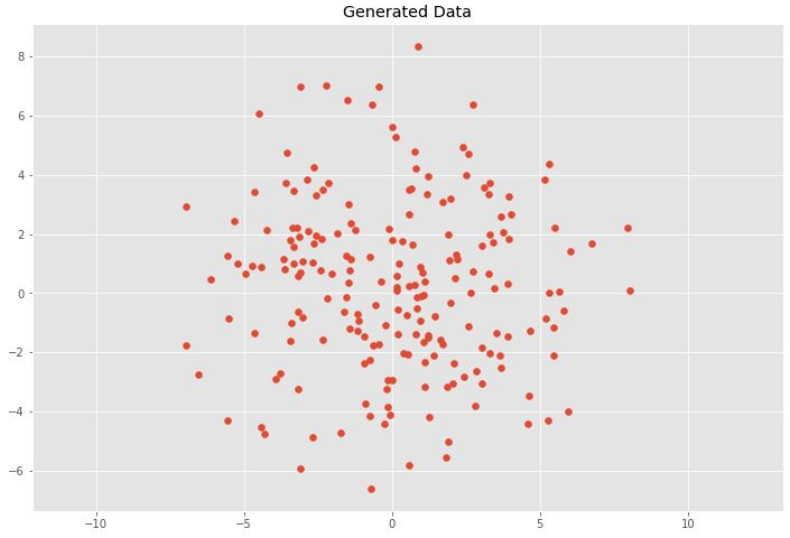
$$\sigma^2_{x^p,y^q} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ip} - \bar{x}_p)(x_{iq} - \bar{x}_q)$$



A **covariância** mede o grau com que os atributos variam juntos.

Baixa Variância

Alta Variância



Dados Multivariados

- Assim, para um conjunto de dados multivariados, podemos avaliar a **matriz de covariância.**
- **Matriz de covariância** é uma matriz 2x2 cujas entradas são variâncias e covariâncias associadas às diversas variáveis.

$$\sigma^2_{x^p,y^q} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)(x_{iq} - \bar{x}_q)$$

$$Cov(X) = \begin{bmatrix} \sigma^2_{11} & \sigma^2_{12} & \cdots & \sigma^2_{1n} \\ \sigma^2_{21} & \sigma^2_{22} & \cdots & \sigma^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{n1} & \sigma^2_{n2} & \cdots & \sigma^2_{nn} \end{bmatrix}$$

$Cov(X)$ é simétrica: $\sigma^2_{ij} = \sigma^2_{j,i}$

$\sigma^2_{ii} = \sigma^2$

Dados Multivariados

- A escala da variação de valores impacta o valor da covariância.
  - Por exemplo, dois atributos com variações de valores na escala de milhares podem ter um covariância maior que dois atributos mais fortemente relacionados, mas que apresentam variações de valores numa escala entre 0 e 1.

- **Correlação** é uma métrica que retira a influência da variação dos valores, sendo mais aplicada na exploração de dados multivariados.

$$corr(x^p, x^q) = \frac{\sigma^2_{x^p,x^q}}{\sigma_{x^p} \times \sigma_{x^q}}$$

Dados Multivariados

- **Correlação** é uma métrica que retira a influência da variação dos valores, sendo mais aplicada na exploração de dados multivariados.

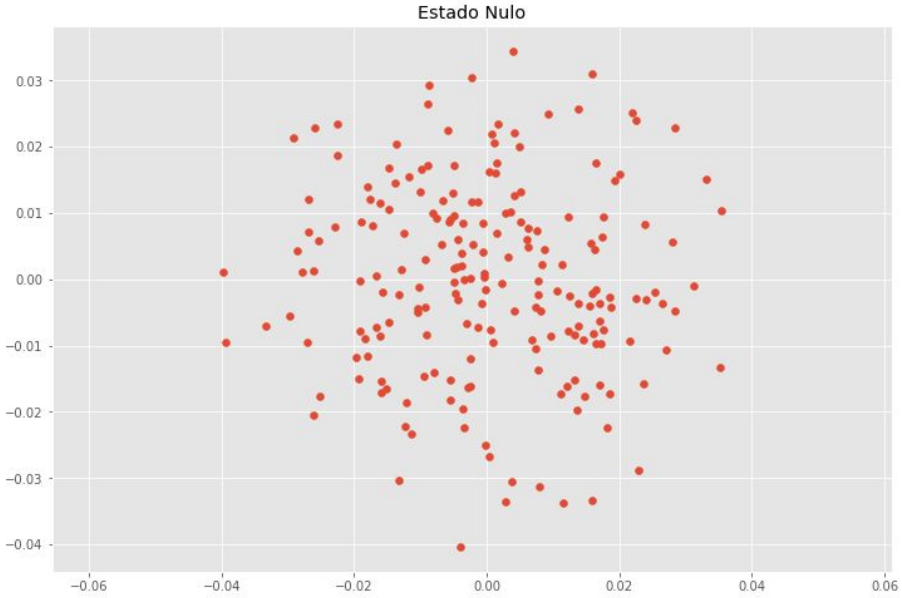
$$corr(x^p, x^q) = \frac{\sigma^2_{x^p, x^q}}{\sigma_{x^p} \times \sigma_{x^q}}$$

$$Corr(X) = \begin{bmatrix} 1 & corr(x^1, x^2) & \dots & corr(x^1, x^n) \\ corr(x^2, x^1) & 1 & \dots & corr(x^2, x^n) \\ \vdots & \vdots & \ddots & \vdots \\ corr(x^n, x^1) & corr(x^n, x^2) & \dots & 1 \end{bmatrix} \quad \begin{matrix} corr(x^p, x^q) \in [-1, +1] \\ corr(x^p, x^p) = 1 \end{matrix}$$

- A correlação tem valor 1 para elementos na diagonal e valores entre –1 (correlação negativa máxima) e +1 (correlação positiva máxima) nos demais casos.

Dados Multivariados

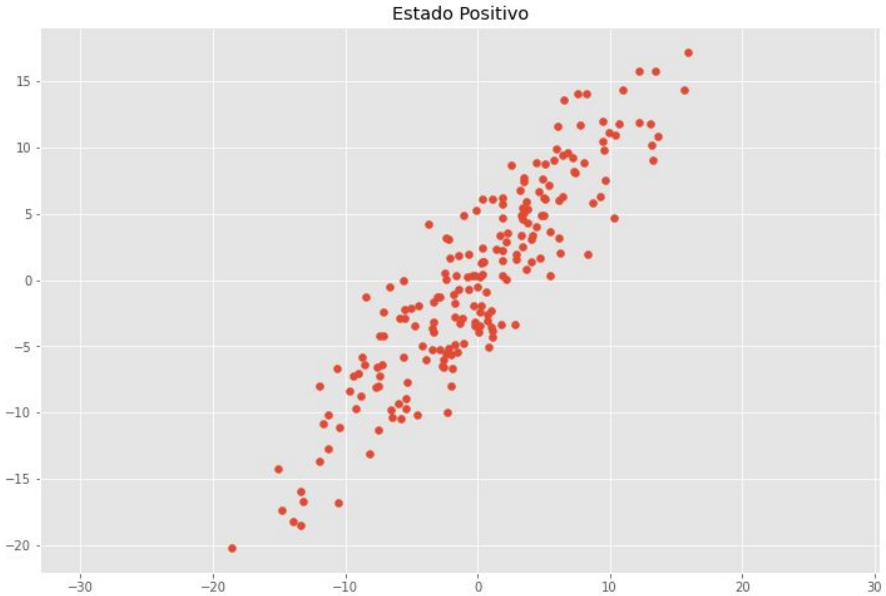
$corr(x^p, x^q) \approx 0$: indica que os atributos não têm um relacionamento linear.



- Nesse caso, não há uma relação entre as variáveis

Dados Multivariados

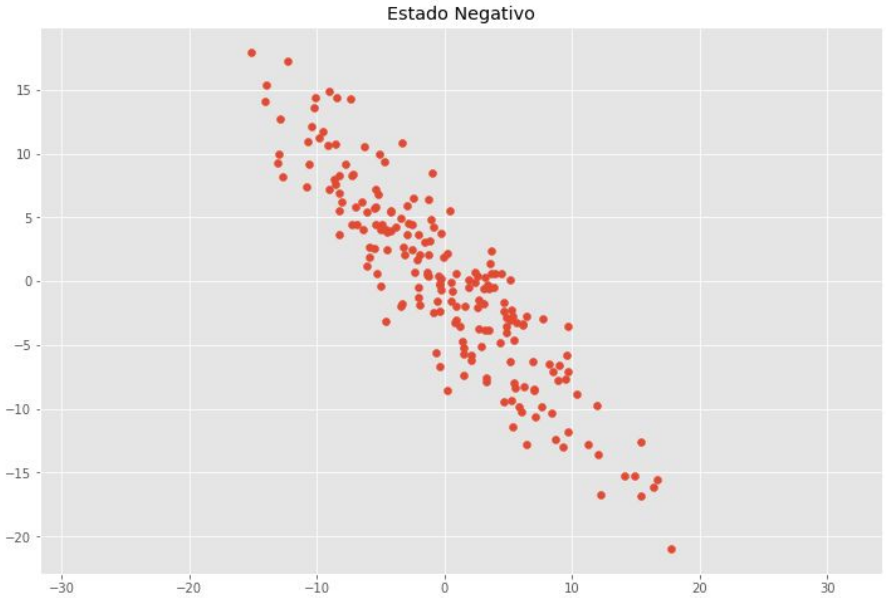
$corr(x^p, y^q) > 0$: indica que os atributos são diretamente relacionados.



- O aumento no valor de um atributo, aumenta o valor no outro.
- Variáveis seguem a mesma direção

Dados Multivariados

$corr(x^p, y^q) < 0$: indica que os atributos são inversamente relacionados.



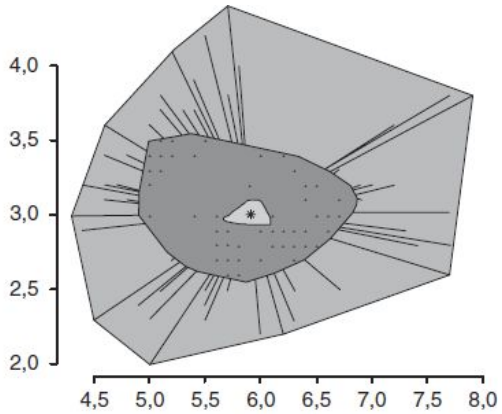
- O aumento no valor de um atributo, reduz o valor no outro.
- Variáveis seguem a direção oposta.

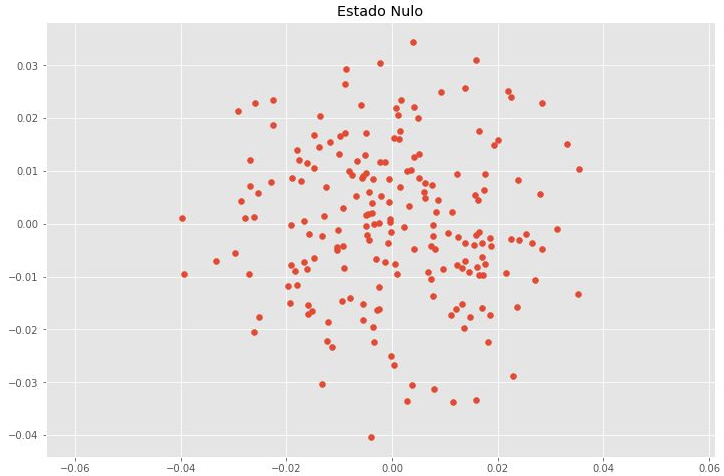
Dados Multivariados

- Essa interpretação se aplica também a covariância.
- Tanto na covariância quanto na correlação:
  - Quando dois atributos apresentam uma correlação positiva, o aumento do valor de um deles é geralmente acompanhado por um aumento no valor do outro.
  - Quando dois atributos têm uma correlação negativa, a redução no valor de um deles é geralmente acompanhada do aumento do valor do outro.

Dados Multivariados

- Recursos de visualização também facilitam a análise dos dados multivariados.
- Explorem recursos de visualização como

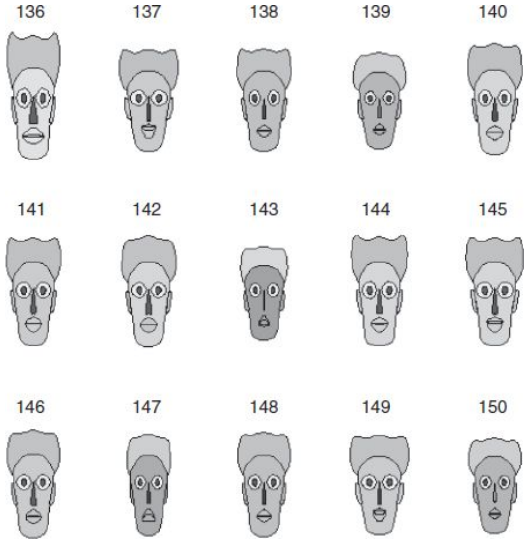




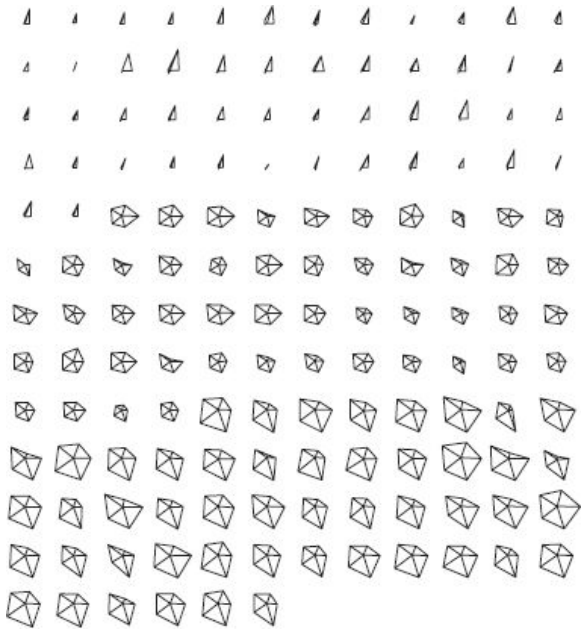
- bagplot é uma generalização bivariada do boxplot.
- scatter plot que permite visualizar a correlação linear entre dois atributos.

Dados Multivariados

- Recursos de visualização também facilitam a análise dos dados multivariados.
- Explorem recursos de visualização como



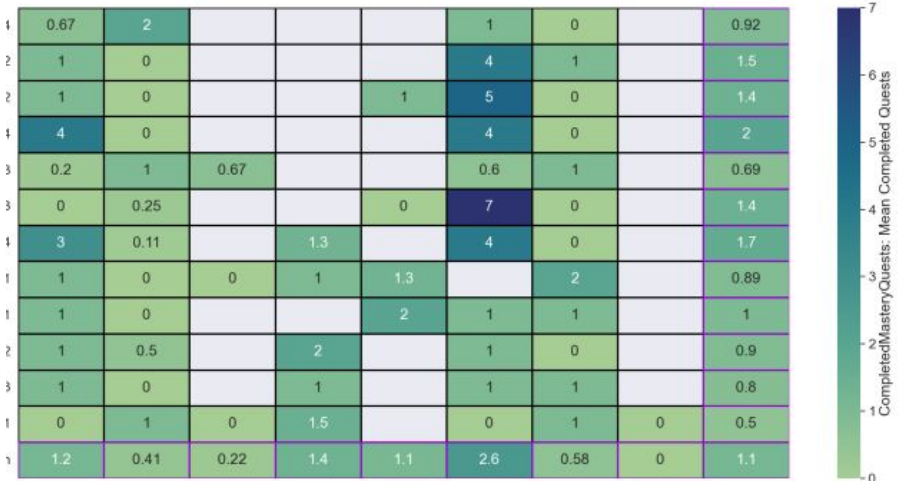
  - Diagrama de Chernoff representa atributos por características em uma face



  - Star plot utiliza figuras geométricas onde cada linha corresponde a um atributo e o tamanho da linha é proporcional ao valor do atributo.

Dados Multivariados

- Recursos de visualização também facilitam a análise dos dados multivariados.
- Explorem recursos de visualização como



  - Heatmaps ou mapas de calor são representações gráficas de dados que utilizam sistemas codificados por cores.