

# Introdução - Mineração de Dados

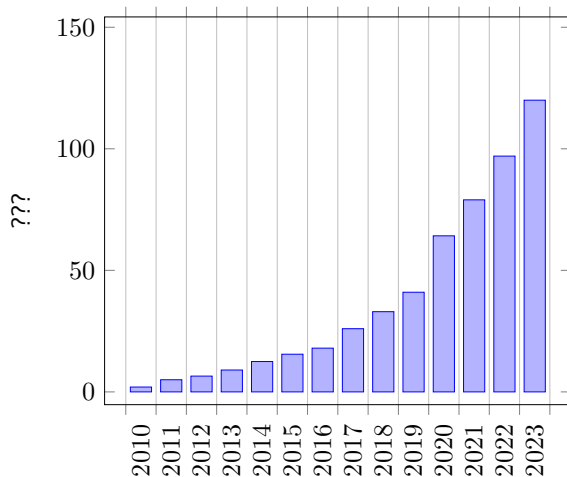
Alexandre Checoli Choueiri

02/08/2023

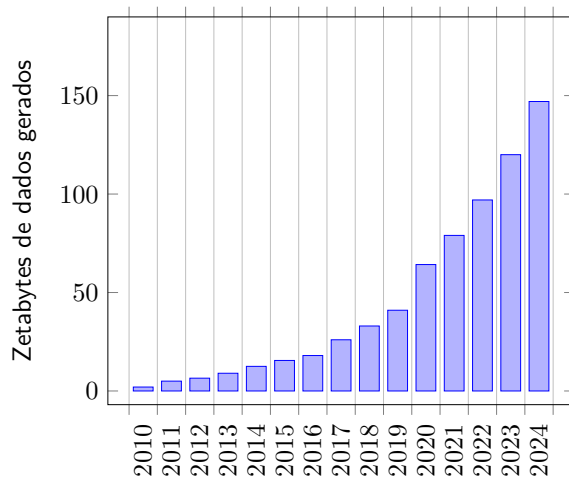
- ① Introdução
- ② Cadeia de valor da informação
- ③ Mineração de dados
- ④ Tarefas preditivas
- ⑤ Classificação
- ⑥ Regressão
- ⑦ Tarefas Descritivas
- ⑧ Agrupamento/Clusterização
- ⑨ Associação
- ⑩ Tarefa extra - mineração de processos
- ⑪ Sobre a disciplina

# Introdução

O que este gráfico representa?



## O que este gráfico representa?



O gráfico representa a **quantidade de dados gerados** (gerado, copiado ou consumido), em zetabytes por ano (com **estimativas**)

Estima-se que:

O gráfico representa a **quantidade de dados gerados** (gerado, copiado ou consumido), em zetabytes por ano (com **estimativas**)

Estima-se que:

1. Em 13 anos o consumo e geração de dados aumentou 74x.
2. No último ano foram gerados mais dados do que o produzido por toda a humanidade até 2014.

Extraír informações úteis de grandes conjuntos de dados é um desafio. Com o aumento das capacidades de processamento e armazenamento dos computadores, **enormes quantidades de dados são gerados diariamente**, demandando técnicas robustas, capazes de gerar informações e agregar valor.

1. *Web-data* e *e-commerce*
2. Compras em supermercados
3. Bancos/ transações com cartão de crédito
4. Sensores remotos em satélites e máquinas no chão de fábrica (indústria 4.0)



# Introdução

**EXERCÍCIO:** Considere o seguinte banco de dados, com informações a respeito da quantidades produzidas de um determinado produto.

ID ordem	$T_i$	$T_f$	Funcionário	Produto	Maquina	Qtde.	Refugo.
O1	28-02-2019: 11:05	28-02-2019: 11:50	Dwight	A	1	20	0
O1	28-02-2019: 12:10	28-02-2019: 12:30	Dwight	A	2	20	0
O1	28-02-2019: 12:35	28-02-2019: 13:55	Angela	A	3	20	1
O2	28-02-2019: 14:00	28-02-2019: 16:30	Oscar	B	1	30	1
O2	28-02-2019: 17:00	28-02-2019: 19:00	DeAngelo	B	2	30	4
O2	28-02-2019: 19:30	28-02-2019: 22:30	Oscar	B	3	30	0
O2	28-02-2019: 22:30	29-02-2019: 02:30	Michael	B	4	30	0
O3	29-02-2019: 03:00	29-02-2019: 03:22	Michael	A	1	10	4
O4	29-02-2019: 18:25	29-02-2019: 18:45	Jan	B	5	5	3

# Introdução

**EXERCÍCIO:** Considere o seguinte banco de dados, com informações a respeito da quantidades produzidas de um determinado produto.

ID ordem	$T_i$	$T_f$	Funcionário	Produto	Maquina	Qtde.	Refugo.
O1	28-02-2019: 11:05	28-02-2019: 11:50	Dwight	A	1	20	0
O1	28-02-2019: 12:10	28-02-2019: 12:30	Dwight	A	2	20	0
O1	28-02-2019: 12:35	28-02-2019: 13:55	Angela	A	3	20	1
O2	28-02-2019: 14:00	28-02-2019: 16:30	Oscar	B	1	30	1
O2	28-02-2019: 17:00	28-02-2019: 19:00	DeAngelo	B	2	30	4
O2	28-02-2019: 19:30	28-02-2019: 22:30	Oscar	B	3	30	0
O2	28-02-2019: 22:30	29-02-2019: 02:30	Michael	B	4	30	0
O3	29-02-2019: 03:00	29-02-2019: 03:22	Michael	A	1	10	4
O4	29-02-2019: 18:25	29-02-2019: 18:45	Jan	B	5	5	3

**O que poderíamos fazer com estes dados? Que informações poderiam ser investigadas?**

# Introdução

**EXEMPLO:** Considere o seguinte banco de dados, com informações a respeito da quantidade vendida de um determinado produto:

<b>Data</b>	<b>Produto</b>	<b>Qtde.</b>
01/01/2001	A	10
05/04/2021	B	30
05/04/2021	A	450
05/04/2021	D	2
...	...	...

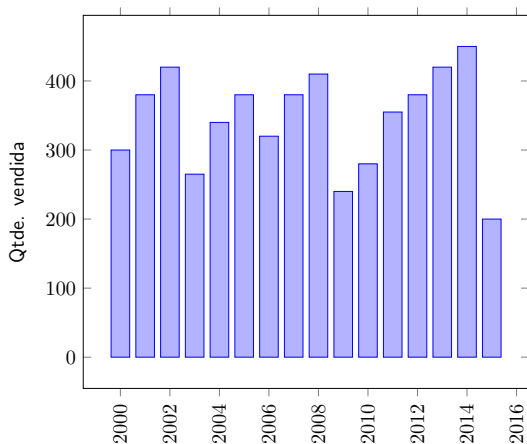
# Introdução

**EXEMPLO:** Considere o seguinte banco de dados, com informações a respeito da quantidade vendida de um determinado produto:

Data	Produto	Qtde.
01/01/2001	A	10
05/04/2021	B	30
05/04/2021	A	450
05/04/2021	D	2
...	...	...

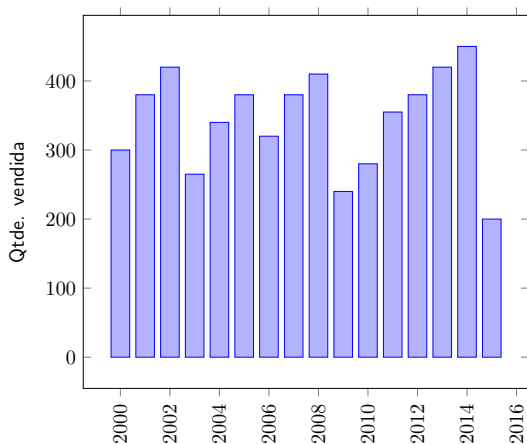
**Como poderíamos usar esses dados?**

# Introdução



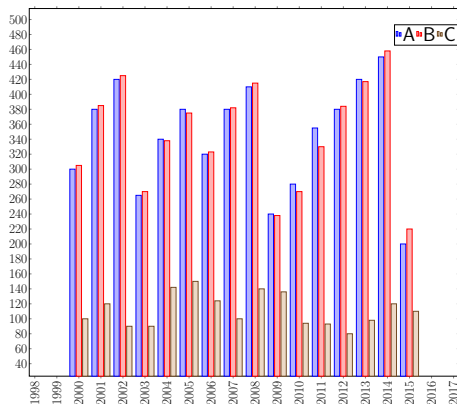
Uma forma é por meio de **gráficos**. Simplesmente plotando a soma de todas as vendas, agrupadas por ano, já nos mostra uma informação importante. **Consegue enxergar?**

# Introdução



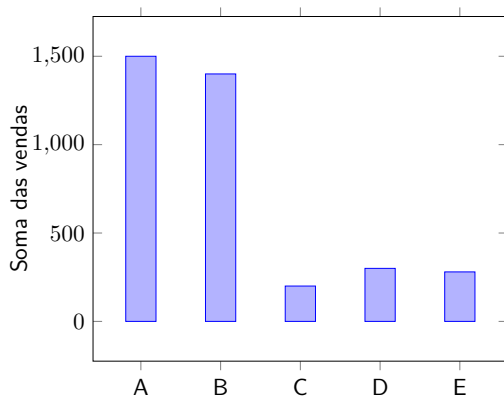
Uma forma é por meio de **gráficos**. Simplesmente plotando a soma de todas as vendas, agrupadas por ano, já nos mostra uma informação importante. **Consegue enxergar?** Note que existe um ciclo: de 3 em 3 anos as vendas crescem, para em seguida o ciclo se repetir (de um ponto mais baixo ou mais alto). **Com essa informação podemos investigar o que ocorre neste ciclo de 3 anos.**

# Introdução



Podemos **filtrar as vendas por produto**, e verificar se o padrão cíclico de vendas totais se mantém com todos eles. Note que pelo gráfico ao lado percebemos que o ciclo se mantém para os produtos A e B, **mas não para o C**. Novamente, isso gera uma nova pergunta que poderia ser examinada pelo gerente.

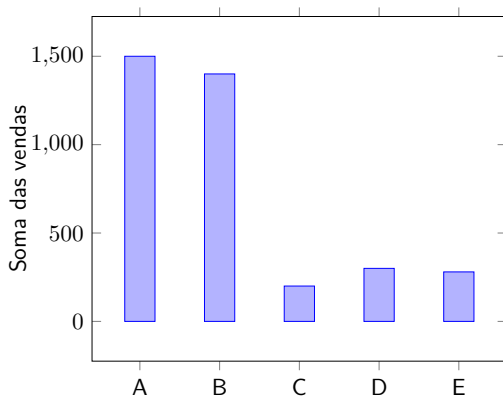
# Introdução



Uma análise bruta de **vendas totais por produto** também pode trazer alguma informação útil sobre quais são os produtos mais vendidos (carro chefe de vendas).

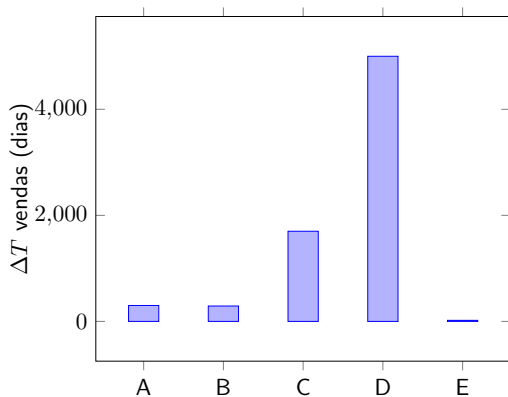


# Introdução



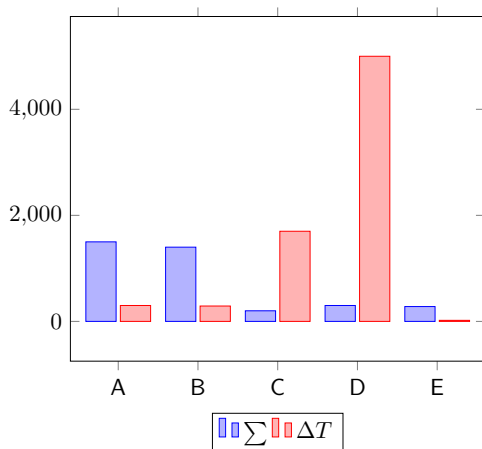
Uma análise bruta de **vendas totais por produto** também pode trazer alguma informação útil sobre quais são os produtos mais vendidos (carro chefe de vendas). **Mas será que esse gráfico realmente mostra o carro chefe?**

# Introdução



Lembre-se que existe a dimensão **tempo** no banco de dados, ou seja, por quanto tempo o produto foi vendido? O gráfico ao lado mostra o  $\Delta T$  existente entre a última data que o produto foi vendido e a primeira (em dias).

# Introdução



Colocando as duas informações em um mesmo gráfico podemos ter uma noção da densidade de vendas de cada produto. Para ficar mais evidente ainda seria ideal criar um gráfico com a razão  $\frac{\sum}{\Delta}$  vendas. Um produto que vendeu pouco, mas em um  $\Delta$  muito baixo, pode ser melhor do que um com muitas vendas em um  $\Delta$  grande.

## Cadeia de valor da informação

## Cadeia de valor da informação

O que fizemos no exemplo anterior pode ser enquadrado no processo de transformar **dados** em **informação**, que por sua vez está contido na chamada **cadeia de valor da informação**:

# Cadeia de valor da informação

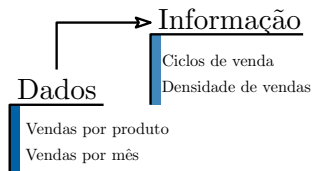
## Dados

Vendas por produto

Vendas por mês

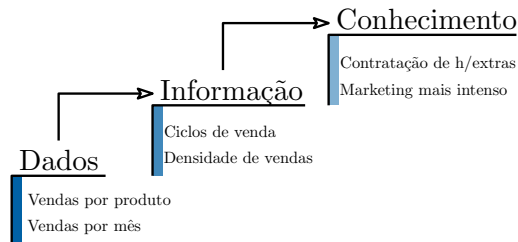
O que estava contido na tabela inicial de vendas dos produtos é considerado **dado**. Por si só, **os dados não são úteis para nada**.

# Cadeia de valor da informação



Quando geramos os gráficos, **transformamos os dados em informação**: por exemplo, descobrimos os ciclos de 3 em 3 meses, as densidades de venda. Mas ainda assim, **a informação por si só não é útil!**

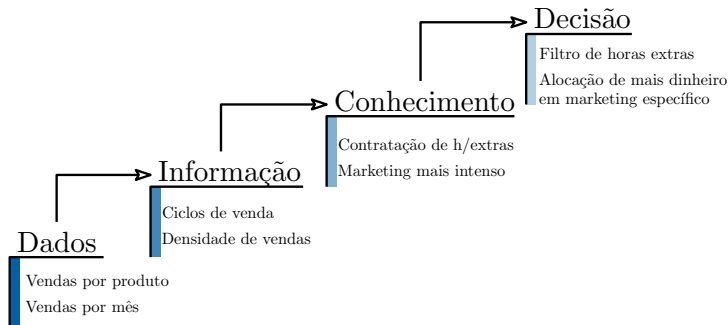
# Cadeia de valor da informação



Com base nessa informação, uma investigação mais minuciosa deve ser realizada, a fim de descobrir o **motivo** dos ciclos de vendas ocorrerem, bem como o motivo das densidades diferentes para cada produto. Isso agrega mais uma camada de valor à cadeia, **transformando a informação em conhecimento**.



# Cadeia de valor da informação



E a última etapa desta cadeia é usar esse conhecimento para auxiliar no processo de **tomada de decisão**. O que será feito de forma diferente, agora que sabemos o motivo de ocorrência da informação? Somente quando usamos o conhecimento para tomar decisões é que os dados que usamos realmente ganham valor na cadeia.

## Cadeia de valor da informação

Mas professor...esse ciclo parece mais uma daquelas ferramentas de produção que não **aguentamos mais decorar**:

## Cadeia de valor da informação

Mas professor...esse ciclo parece mais uma daquelas ferramentas de produção que não **aguentamos mais decorar**:

S5, 5W2H,  $6\sigma$ , PDCA, pilares da qualidade

## Cadeia de valor da informação

Mas professor...esse ciclo parece mais uma daquelas ferramentas de produção que não **aguentamos mais decorar**:

S5, 5W2H,  $6\sigma$ , PDCA, pilares da qualidade

E sinceramente todas poderiam ser chamadas simplesmente de...**bom senso**!

## Cadeia de valor da informação

Mas vocês sabem que eu não colocaria algo aqui **se realmente não tivesse algum valor.**

Embora essa cadeia de valor também seja algo intuitivo, e de certa forma com bom senso nem precisaríamos utilizá-la, existe uma pegadinha...Ao trabalharmos com dados, pela facilidade que temos de extrair informações (gráficos) dos mesmos (seja por linguagens de programação, excel, *power BI*, etc...) facilmente nos perdemos em tanta informação, e não a transformamos em conhecimento! **Sem conhecimento, não tomamos decisão. Na ânsia de gerar relatório bonitos e dinâmicos, geramos relatórios bonitos e dinâmicos que não agregam nada!**

## Cadeia de valor da informação

Mas vocês sabem que eu não colocaria algo aqui **se realmente não tivesse algum valor.**

Embora essa cadeia de valor também seja algo intuitivo, e de certa forma com bom senso nem precisaríamos utilizá-la, existe uma pegadinha...Ao trabalharmos com dados, pela facilidade que temos de extrair informações (gráficos) dos mesmos (seja por linguagens de programação, excel, *power BI*, etc...) facilmente nos perdemos em tanta informação, e não a transformamos em conhecimento! **Sem conhecimento, não tomamos decisão. Na ânsia de gerar relatório bonitos e dinâmicos, geramos relatórios bonitos e dinâmicos que não agregam nada!**

Então de certa forma o ciclo serve para nos lembrar de que: gerar gráficos não garante a geração de conhecimento, sem conhecimento não melhoramos o processo de tomada de decisão, e se não alteramos decisões, tudo foi em vão.

# Mineração de dados

# Mineração de dados

E será que o que fizemos com o banco de dados pode ser considerado **mineração de dados**? Vamos ver a definição:



# Mineração de dados

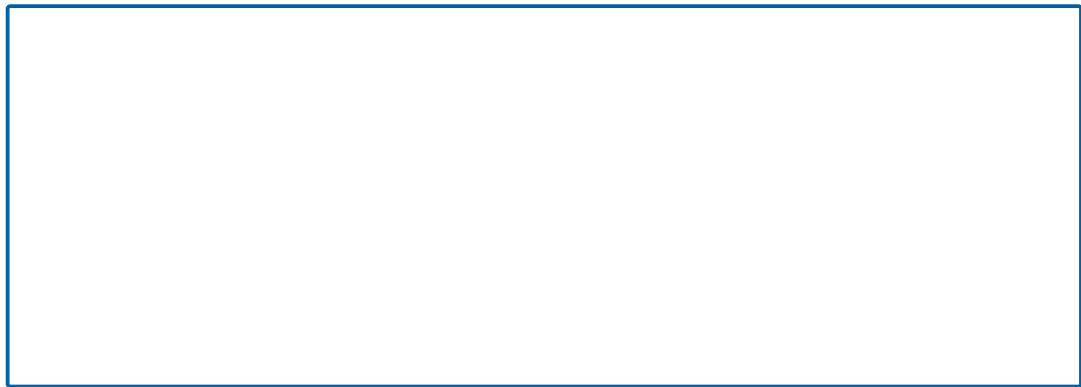
E será que o que fizemos com o banco de dados pode ser considerado **mineração de dados**? Vamos ver a definição:

## Definição

**Mineração de dados:** Extração **não trivial** de informação implícita, **previamente desconhecida** e **potencialmente útil** a partir dos dados

# Mineração de dados

Alguns termos da definição já podem ser destacados, a informação que extraímos do banco de dados de produtos era...



# Mineração de dados

Alguns termos da definição já podem ser destacados, a informação que extraímos do banco de dados de produtos era...

## 1. Desconhecida?

# Mineração de dados

Alguns termos da definição já podem ser destacados, a informação que extraímos do banco de dados de produtos era...

1. **Desconhecida?** *sim*, até gerarmos os gráficos não sabíamos da existência de ciclos e densidades de vendas de produtos.

# Mineração de dados

Alguns termos da definição já podem ser destacados, a informação que extraímos do banco de dados de produtos era...

1. **Desconhecida?** *sim*, até gerarmos os gráficos não sabíamos da existência de ciclos e densidades de vendas de produtos.
2. **Potencialmente útil?**

# Mineração de dados

Alguns termos da definição já podem ser destacados, a informação que extraímos do banco de dados de produtos era...

1. **Desconhecida?** **sim**, até gerarmos os gráficos não sabíamos da existência de ciclos e densidades de vendas de produtos.
2. **Potencialmente útil?** pela cadeia de valor, se o conhecimento que foi extraído for usado para melhorar a tomada de decisão, então **sim**! (existe o potencial de ser útil)!

# Mineração de dados

Alguns termos da definição já podem ser destacados, a informação que extraímos do banco de dados de produtos era...

1. **Desconhecida?** **sim**, até gerarmos os gráficos não sabíamos da existência de ciclos e densidades de vendas de produtos.
2. **Potencialmente útil?** pela cadeia de valor, se o conhecimento que foi extraído for usado para melhorar a tomada de decisão, então **sim**! (existe o potencial de ser útil)!
3. **Não trivial?**

# Mineração de dados

Alguns termos da definição já podem ser destacados, a informação que extraímos do banco de dados de produtos era...

1. **Desconhecida?** **sim**, até gerarmos os gráficos não sabíamos da existência de ciclos e densidades de vendas de produtos.
2. **Potencialmente útil?** pela cadeia de valor, se o conhecimento que foi extraído for usado para melhorar a tomada de decisão, então **sim**! (existe o potencial de ser útil)!
3. **Não trivial?** **talvez**...é aqui que demarcamos o limite mais importante da mineração de dados. Para encontrarmos as informações simplesmente fizemos alguns filtros e somas nos dados, de forma que podem ser consideradas **operações triviais**.



# Mineração de dados

Dessa forma percebemos que nem tudo que fazemos com dados pode ser considerado mineração de dados. **Relatórios bonitos e dinâmicos quase nunca são mineração de dados...**

# Motivação de criação e origens

Alguns problemas motivaram o surgimento da MD como uma disciplina:

# Motivação de criação e origens

Alguns problemas motivaram o surgimento da MD como uma disciplina:

1. **Escalabilidade:** o aumento no volume de dados exige algoritmos escaláveis, novas estruturas de dados e métodos de busca.

# Motivação de criação e origens

Alguns problemas motivaram o surgimento da MD como uma disciplina:

1. **Escalabilidade:** o aumento no volume de dados exige algoritmos escaláveis, novas estruturas de dados e métodos de busca.
2. **Alta dimensionalidade:** não é incomum se encontrar dados com centenas ou milhares de atributos (colunas), diferentemente de algumas dezenas como algumas décadas atrás. Técnicas de análise de dados tradicionais, feitas para lidar com poucos atributos, na maioria das vezes não desempenham bem para esses dados.

# Motivação de criação e origens

Alguns problemas motivaram o surgimento da MD como uma disciplina:

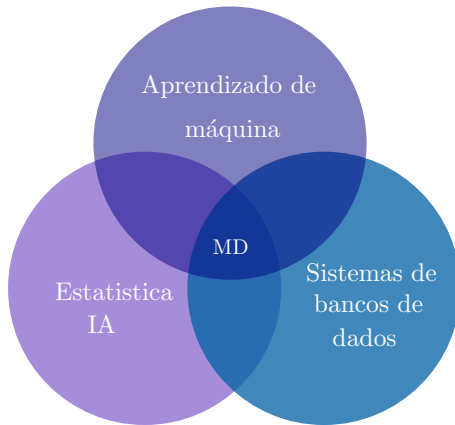
1. **Escalabilidade:** o aumento no volume de dados exige algoritmos escaláveis, novas estruturas de dados e métodos de busca.
2. **Alta dimensionalidade:** não é incomum se encontrar dados com centenas ou milhares de atributos (colunas), diferentemente de algumas dezenas como algumas décadas atrás. Técnicas de análise de dados tradicionais, feitas para lidar com poucos atributos, na maioria das vezes não desempenham bem para esses dados.
3. **Dados heterogêneos e complexos:** métodos de análise tradicionais geralmente lidam com dados de um mesmo tipo, ou contínuos ou categóricos. Recentemente surgiram objetos mais complexos que são armazenados em BDs, como páginas da *web* com textos semi-estruturados.

## Motivação de criação e origens

A mineração de dados surge então como um amálgama de disciplinas e conceitos, para lidar com esses problemas:

# Motivação de criação e origens

A mineração de dados surge então como um amálgama de disciplinas e conceitos, para lidar com esses problemas:



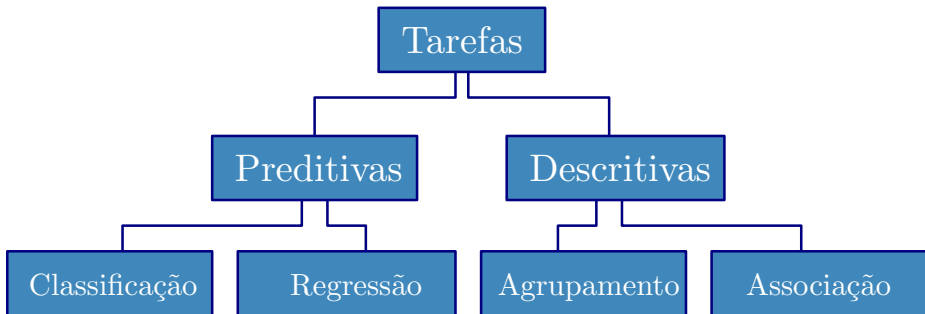
# Tarefas preditivas



# Tarefas da MD - Preditivas

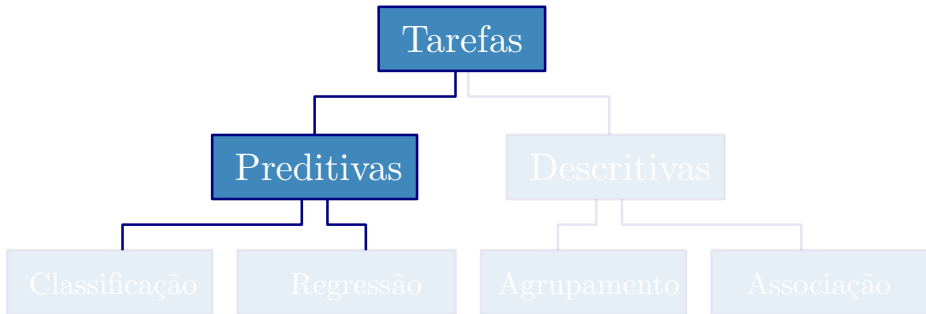
As aplicações de mineração de dados podem ser classificadas em 4 tarefas, sendo que estas são separadas por duas categorias.

# Tarefas da MD - Preditivas



As aplicações de mineração de dados podem ser classificadas em 4 tarefas, sendo que estas são separadas por duas categorias.

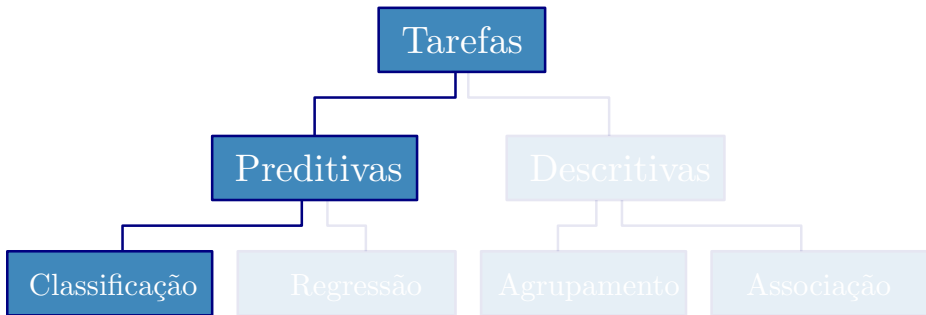
# Tarefas da MD



As **tarefas preditivas** buscam prever um atributo baseado em outros. Geralmente o valor a ser estimado é chamado de valor *target* ou **variável dependente**, enquanto os outros atributos usados para a predição são chamados de **variáveis independentes**.

# Classificação

# Tarefas da MD - Classificação



A **classificação** é a tarefa preditiva com maior número de aplicações atualmente. Vejamos do que se trata.

## Tarefas da MD - Classificação

**Classificação:** dado um conjunto de registros (dados), em que cada registro contém um conjunto de atributos, e um desses atributos se refere a uma classe (ou seja, um dado discreto ou categórico), a classificação visa encontrar uma função (modelo) que relacione o conjunto de atributos a uma classe. **Com essa função definida a partir dos dados, a mesma pode ser usada para prever valores com base em novos conjuntos de atributos.**

## Tarefas da MD - Classificação

Idade	Bebe	Fuma	Faz exercício	Diabetes
33	S	N	S	S
48	N	N	N	S
80	N	N	S	N
25	S	S	S	N
15	N	S	N	N
...	...	...	...	...

Considere o banco de dados acima que apresenta dados pessoais de pacientes um consultório médico, bem como se os mesmos possuem (S) ou não (N) diabetes.

## Tarefas da MD - Classificação

Idade	Bebe	Fuma	Faz exercício	Diabetes
33	S	N	S	S
48	N	N	N	S
80	N	N	S	N
25	S	S	S	N
15	N	S	N	N
...	...	...	...	...

Se considerarmos a variável **Diabetes** como *target*, a tarefa de classificação busca encontrar uma função que relacione as outras variáveis com a diabetes.



## Tarefas da MD - Classificação

Idade	Bebe	Fuma	Faz exercício	Diabetes
33	S	N	S	S
48	N	N	N	S
80	N	N	S	N
25	S	S	S	N
15	N	S	N	N
...	...	...	...	...

Dessa forma, teríamos um **modelo de classificação** que recebe como *input* um vetor do tipo:

$$x = [\text{Idade}, \text{Bebe}, \text{Fuma}, \text{Faz exercício}]$$

E retorna uma das duas classes possíveis para diabetes (S ou N).


## Tarefas da MD - Classificação

Idade	Bebe	Fuma	Faz exercício	Diabetes
33	S	N	S	S
48	N	N	N	S
80	N	N	S	N
25	S	S	S	N
15	N	S	N	???

Com esse modelo estimado, podemos prever se um novo paciente tem diabetes ou não, com base em seus outros atributos

## Tarefas da MD - Classificação

Quais outras aplicações poderiam ser criadas que usam a tarefa de classificação?

A large, empty rectangular box with a blue border, intended for the user to write their answer to the question above.

# Tarefas da MD - Classificação

Quais outras aplicações poderiam ser criadas que usam a tarefa de classificação?

1. **Diagnósticos médicos**

# Tarefas da MD - Classificação

Quais outras aplicações poderiam ser criadas que usam a tarefa de classificação?

1. **Diagnósticos médicos**
2. **Classificação de imagens**

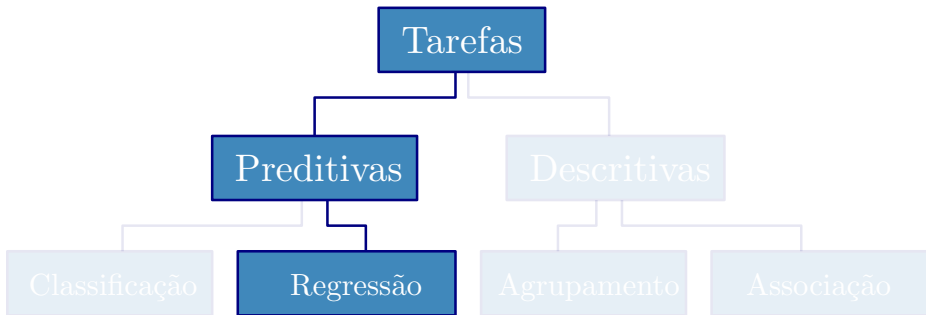
# Tarefas da MD - Classificação

Quais outras aplicações poderiam ser criadas que usam a tarefa de classificação?

1. **Diagnósticos médicos**
2. **Classificação de imagens**
3. **Classificação de clientes de banco para fornecimento de crédito**

# Regressão

# Tarefas da MD - Regressão



Como a classificação, a **regressão** também é uma tarefa preditiva, a única diferença é que o tipo da variável *target* não é uma classe, mas sim **numérica**.



## Tarefas da MD - Regressão

Nº quartos	Nº banheiros	Nº garagens	Sacada ?	Churrasqueira?	R\$/m <sup>2</sup>
4	3	1	S	N	10.000
4	4	2	S	S	12.000
2	1	1	N	N	6.000
...	...	...	...	...	

Considere o banco de dados acima que apresenta dados de uma pesquisa referente a características de imóveis e o custo por metro quadrado dos mesmos.

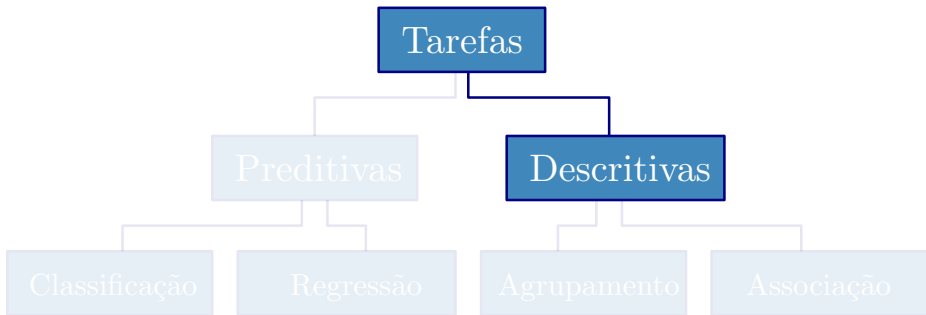
## Tarefas da MD - Regressão

N <sup>o</sup> quartos	N <sup>o</sup> banheiros	N <sup>o</sup> garagens	Sacada ?	Churrasqueira?	R\$/m <sup>2</sup>
4	3	1	S	N	10.000
4	4	2	S	S	12.000
2	1	1	N	N	6.000
...	...	...	...	...	

Um modelo de **regressão** poderia fazer a predição dos valores de m<sup>2</sup> com base nas características do imóvel.

# Tarefas Descritivas

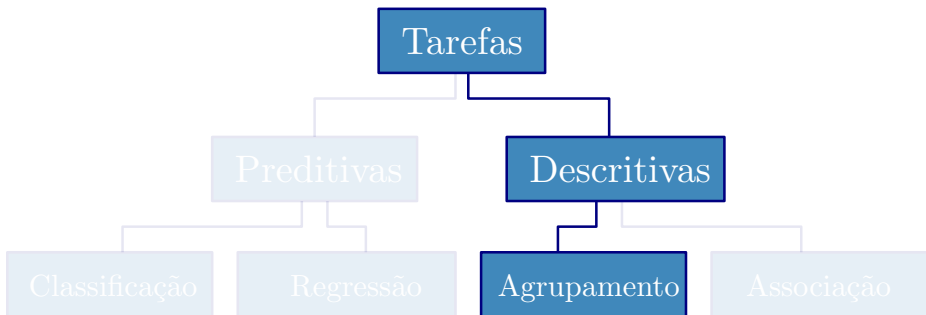
# Tarefas da MD - Descritivas



As **tarefas descritivas** tem o objetivo de encontrar padrões (correlações, tendências, agrupamentos, anomalias) nos dados. Nestas tarefas não existe um atributo "alvo" / *target*, como nas tarefas preditivas.

# Agrupamento/Clusterização

# Tarefas da MD - Descritivas



A **análise de agrupamento (clusterização)** busca grupos de informações que estão "próximos" uns dos outros, de acordo com alguma medida de similaridade.

# Agrupamento/Clusterização

Considere um banco de dados de pontos no plano cartesiano:

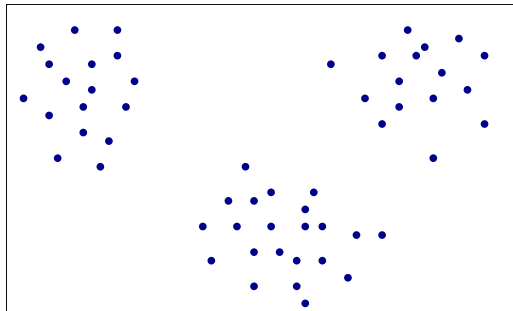
x	y
1	2
10	20
4	54
8	20
...	...

# Agrupamento/Clusterização

Considere um banco de dados de pontos no plano cartesiano:

x	y
1	2
10	20
4	54
8	20
...	...

Podemos representar os dados graficamente:



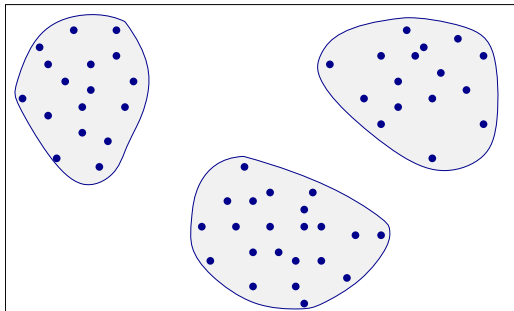


# Agrupamento/Clusterização

Considere um banco de dados de pontos no plano cartesiano:

x	y
1	2
10	20
4	54
8	20
...	...

O agrupamento encontraria grupos de pontos que estão próximos, no caso abaixo, 3 grupos.



## Agrupamento/Clusterização

Mas como isso é usado na prática? Considere que os dados anteriores não são coordenadas, mas sim valores que um conjunto de clientes gasta nas lojas x, y e z de um *shopping center*:

x	y	z
1	2	200
10	20	150
4	54	0
8	20	450
...	...	...

## Agrupamento/Clusterização

Mas como isso é usado na prática? Considere que os dados anteriores não são coordenadas, mas sim valores que um conjunto de clientes gasta nas lojas x, y e z de um *shopping center*:

x	y	z
1	2	200
10	20	150
4	54	0
8	20	450
...	...	...

Ao encontrarmos *clusters* ou grupos de clientes por meio da clusterização, uma **campanha de marketing direcionada** pode ser feita: *e-mails* ou propagandas com produtos semelhantes são enviados para clientes de grupos iguais (clientes mais ricos, que compram nas mesmas lojas, com perfis de compra parecidos).

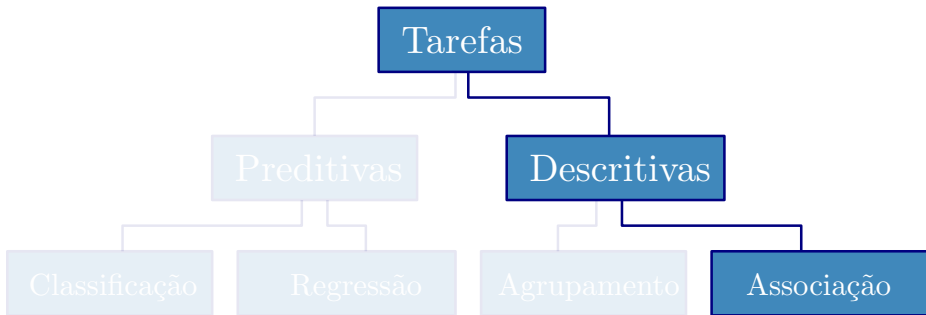
# Agrupamento/Clusterização

Algumas aplicações de agrupamento envolvem:

1. Segmentação de clientes/mercados
2. Agrupamento de textos (*text-mining* / *sentiment analysis*)

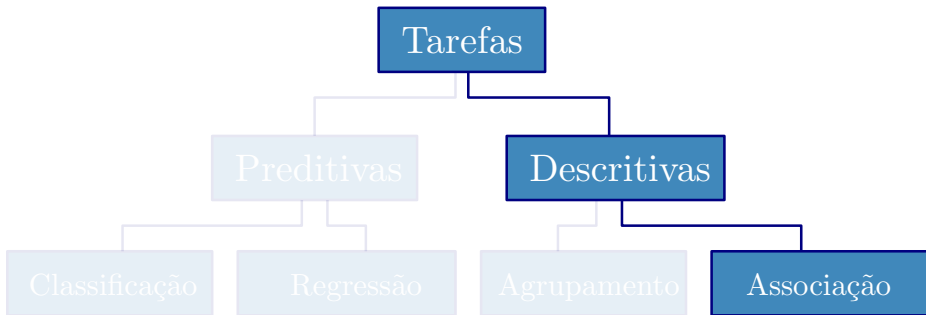
# Associação

## Associação (regras de associação)



Finalmente temos a tarefa de **associação, ou geração de regras de associação**. A análise de associação é utilizada para descobrir fortes associações entre os atributos dos dados.

## Associação (regras de associação)

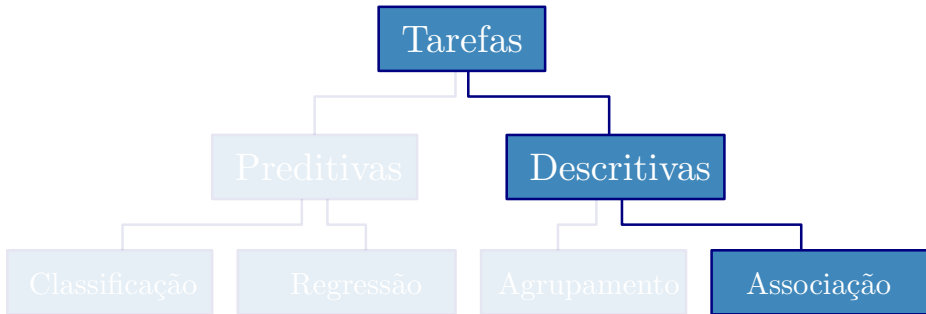


Geralmente o conhecimento extraído nesta tarefa é representado na forma de regras de associação do tipo:

$$A \rightarrow B$$

A informação é lida da seguinte forma: **em muitos casos em que A ocorre, B também ocorre.**

# Associação (regras de associação)



OBS: A regra não implica causalidade, ou seja, não podemos ler: se A então B!



## Associação (regras de associação)

Considere um banco de dados com as informações de transações em um supermercado (os itens que cada cliente comprou, pela nota fiscal):

Transações	Itens comprados
1	pão, leite, manteiga, carne
2	café, manteiga, carne
3	pão, leite
4	churrasqueira, leite ,ovo
...	...

## Associação (regras de associação)

Considere um banco de dados com as informações de transações em um supermercado (os itens que cada cliente comprou, pela nota fiscal):

Transações	Itens comprados
1	pão, leite, manteiga, carne
2	café, manteiga, carne
3	pão, leite
4	churrasqueira, leite ,ovo
...	...

Com esse banco de dados, uma regra extraída poderia ser:

café  $\rightarrow$  pão

E com essa informação a equipe de vendas poderia trabalhar em "combos" para os itens, ou mesmo alocar os dois mais próximos um do outro.

## Associação (regras de associação)

Duas histórias sobre associação (uma sem usar algoritmos) são bem famosas no meio da mineração de dados:

1. O caso da descoberta de gravidez antes da grávida ([link](#))
2. O caso da relação fraldas/cerveja ([link](#))

## Retomando...

**EXERCÍCIO:** Considere o seguinte banco de dados, com informações a respeito da quantidades produzidas de um determinado produto.

ID ordem	$T_i$	$T_f$	Funcionário	Produto	Maquina	Qtde.	Refugo.
O1	28-02-2019: 11:05	28-02-2019: 11:50	Dwight	A	1	20	0
O1	28-02-2019: 12:10	28-02-2019: 12:30	Dwight	A	2	20	0
O1	28-02-2019: 12:35	28-02-2019: 13:55	Angela	A	3	20	1
O2	28-02-2019: 14:00	28-02-2019: 16:30	Oscar	B	1	30	1
O2	28-02-2019: 17:00	28-02-2019: 19:00	DeAngelo	B	2	30	4
O2	28-02-2019: 19:30	28-02-2019: 22:30	Oscar	B	3	30	0
O2	28-02-2019: 22:30	29-02-2019: 02:30	Michael	B	4	30	0
O3	29-02-2019: 03:00	29-02-2019: 03:22	Michael	A	1	10	4
O4	29-02-2019: 18:25	29-02-2019: 18:45	Jan	B	5	5	3

**O que poderíamos fazer com estes dados? (agora que conhecemos as tarefas da mineração de dados) Que informações poderiam ser investigadas?**

## Tarefa extra - mineração de processos

# Mineração de processos

Embora não esteja nas aplicações clássicas de mineração de dados, pois surgiu depois, a **mineração de processos** também é uma forma de mineração de dados. A MP precisa de um banco de dados especial, chamado de **log de eventos**, que contém atividades e datas ou horários que estas atividades foram realizadas. O algoritmos de MP extraem o processo gerador dos dados (em diversos formatos, BPMN, redes de Petri, etc...).

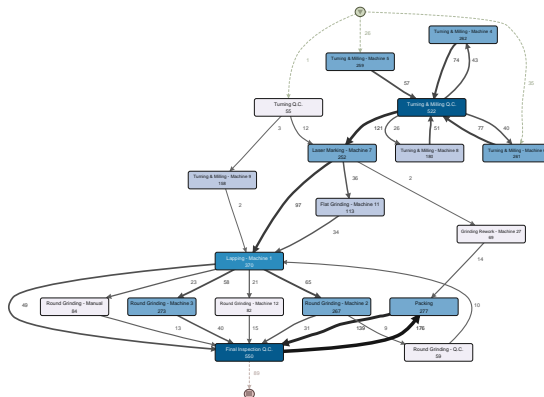
## Mineração de processos

Os dados abaixo são um **log de eventos** de um ambiente produtivo:

ID ordem	$T_i$	$T_f$	Produto	Maquina	Qtde.
O1	28-02-2019: 11:05	28-02-2019: 11:50	A	1	20
O1	28-02-2019: 12:10	28-02-2019: 12:30	A	2	20
O1	28-02-2019: 12:35	28-02-2019: 13:55	A	3	20
O2	28-02-2019: 14:00	28-02-2019: 16:30	B	1	30
O2	28-02-2019: 17:00	28-02-2019: 19:00	B	2	30
O2	28-02-2019: 19:30	28-02-2019: 22:30	B	3	30
O2	28-02-2019: 22:30	29-02-2019: 02:30	B	4	30
O3	29-02-2019: 03:00	29-02-2019: 03:22	A	1	10
O4	29-02-2019: 18:25	29-02-2019: 18:45	B	5	5

## Mineração de processos

O processo extraído do banco é o seguinte:

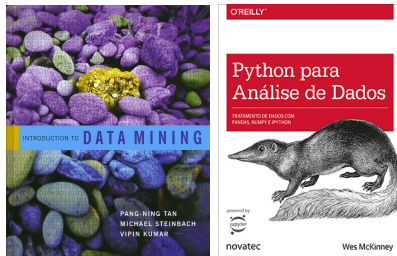




## Sobre a disciplina

## Sobre a disciplina - Referências

Existem diversas ferramentas que podem ser usadas para a MD, nós usaremos a linguagem de programação Python. Cerca de 70% das aulas serão sobre Python, e somente 30% sobre MD. A principal referência para mineração de dados é o livro de Tan, Michael Steinbach e Vipin Kumar - **"Introduction to Data Mining"**. E para Python **"Python para Análise de Dados"** de Wes Mckinney.



## Sobre a disciplina - Notas

As notas serão compostas dos seguintes termos:

1. **PROVA (3.0):** Uma prova escrita sobre os conteúdos (mineração de dados e Python).
2. **LISTAS DE EXERCÍCIOS (3.0):** Resolução das listas de exercícios de Python (todas no site).
3. **TRABALHO FINAL EM GRUPO (4.0):** Os alunos deverão criar alguma aplicação (qualquer que seja), usando pelo menos um pacote que não foi passado em aula. Explicar o pacote em sala e mostrar o código.

**OBS:** A prova de exame final será resolver um estudo de caso