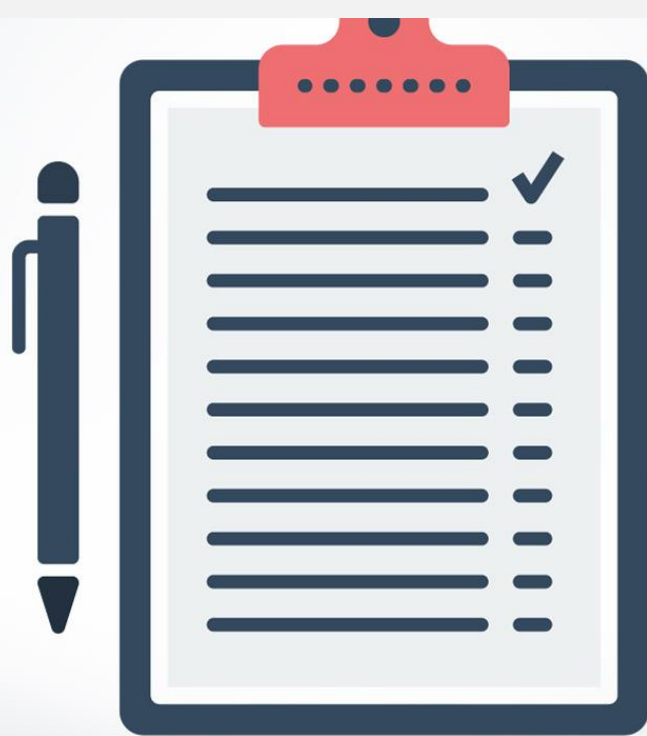


# Mineração de Dados

introdução

Eduardo Silva – easilva91@gmail.com

```
31 def __init__(self, settings):
32     self.file = None
33     self.fingerprints = set()
34     self.logdups = True
35     self.debug = debug
36     self.logger = logging.getLogger(__name__)
37     if path:
38         self.file = open(os.path.join(path, 'requests.json'),
39                           'w')
40         self.file.seek(0)
41         self.fingerprints.update(retrieved)
42
43 @classmethod
44 def from_settings(cls, settings):
45     debug = settings.getbool('SUPERDEBUG')
46     return cls(job_dir(settings), debug)
47
48 def request_seen(self, request):
49     fp = self.request_fingerprint(request)
50     if fp in self.fingerprints:
51         return True
52     self.fingerprints.add(fp)
53     if self.file:
54         self.file.write(fp + os.linesep)
55
56 def request_fingerprint(self, request):
57     return request_fingerprint(request)
```



# Agenda

## Mineração de Dados

### 1. Introdução

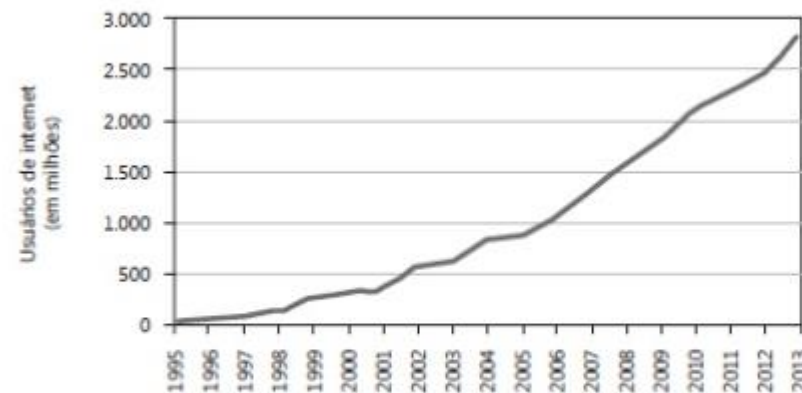
- I. O que é Mineração de dados?
- II. Principais Tarefas da Mineração de Dados.
- III. As Diferentes Nomenclaturas.
- IV. Falácias sobre Mineração de dados.
- V. Exemplos de Uso.

# Introdução

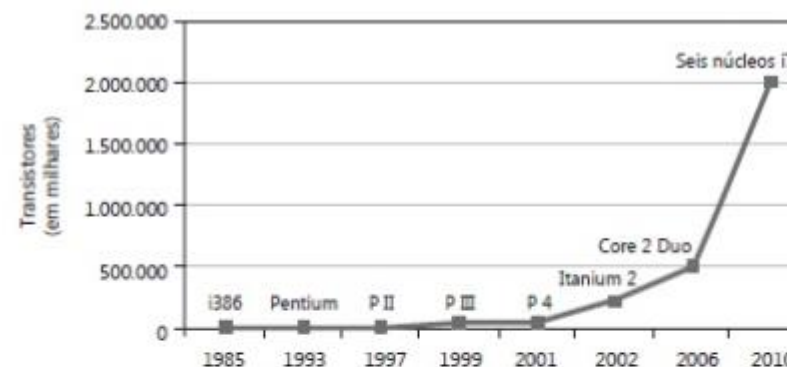
Nós últimos anos temos presenciado uma crescente em relação do uso da internet, assim como na capacidade de processamento de chips de computadores.

- Big Data
- Cloud Computing
- Baixo custo de recursos computacionais

Figura 1.1 Tempos exponenciais



(a)



(b)

(a) Crescimento do número de usuários de Internet no mundo (Fonte: Internetworldstats.com).

(b) Lei de Moore: crescimento do número de transistores em um circuito integrado (Fonte: Intel.com).

L. N. de Castro & D. G. Ferrari, Introdução à  
Mineração de Dados: Conceitos Básicos,  
Algoritmos e Aplicações, 1ª. Ed., Saraiva, 2016.

# Introdução

## O que é mineração de dados?

O termo mineração é proveniente da mineração de minérios. O termo mineração de dados faz uma alusão ao termo de origem, uma vez que se explora uma base de dados (mina) usando algoritmos (ferramentas) adequadas para se obter conhecimento (minerais preciosos).

Figura 1.2 Exemplo da diferença entre dados, informação e conhecimento





# Introdução

O que é mineração de dados?

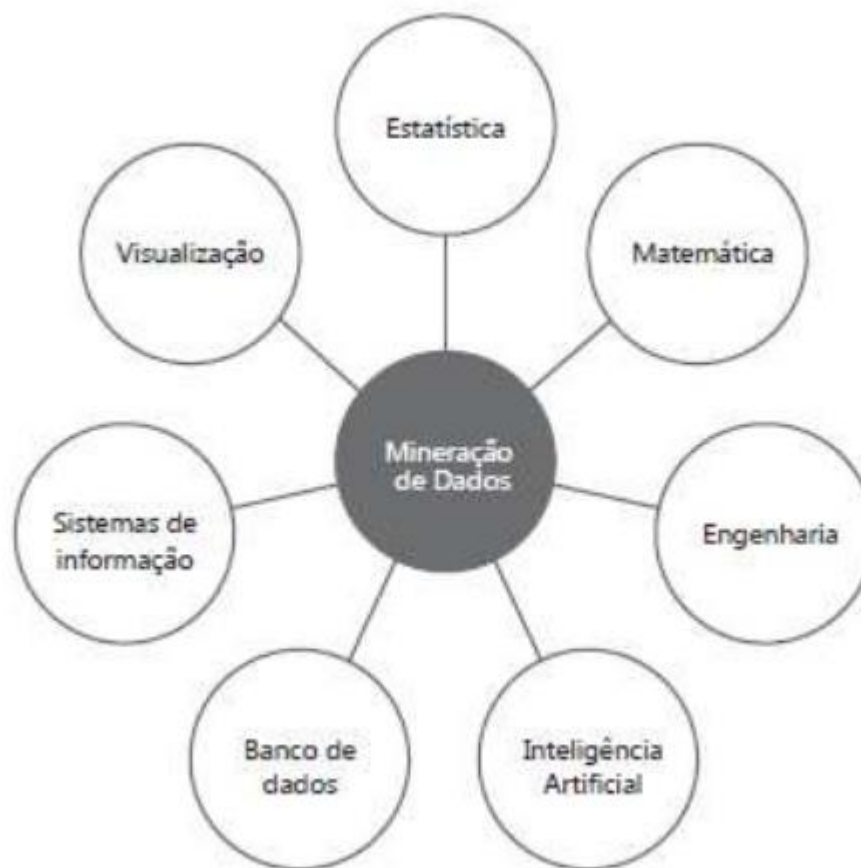
Figura 1.3 Processo de descoberta de conhecimento em bases de dados



# Introdução

O que é mineração de dados?

Figura 1.4 Multidisciplinaridade da mineração de dados



# Introdução

Principais tarefas da mineração de dados?

## 1 – Análise descritiva dos dados

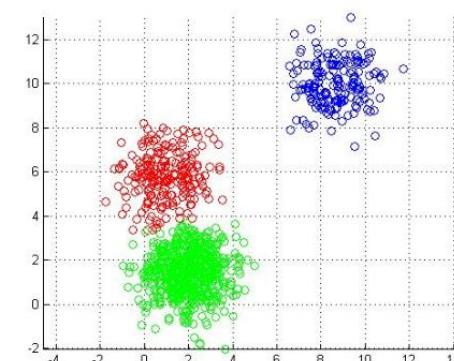
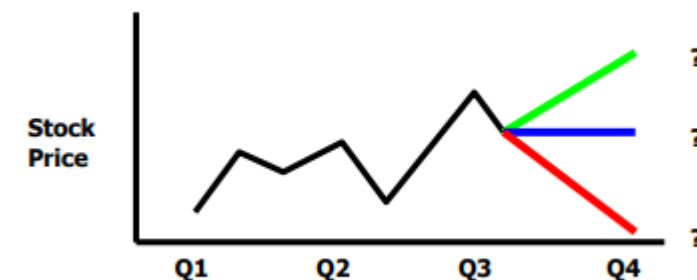
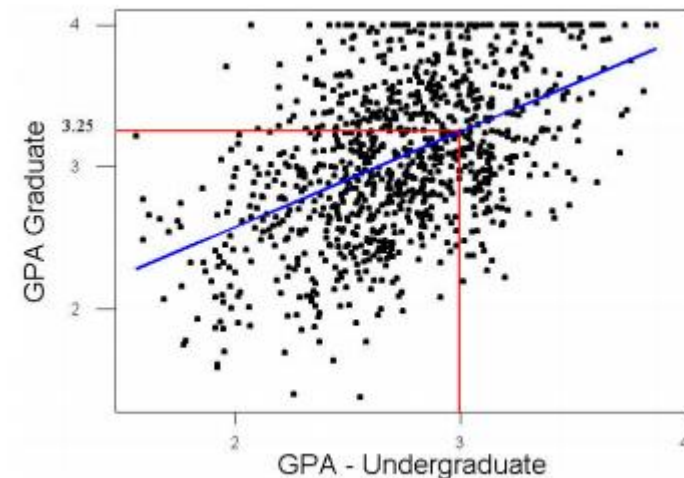
Os algoritmos de aprendizagem de máquina são ferramentas poderosas para a descoberta de conhecimentos em bases de dados. Entretanto, uma etapa inicial do processo de mineração que não requer elevado nível de sofisticação é a análise descritiva dos dados.

## 2 – Predição: classificação e estimação

*Predição* é uma terminologia usada para se referir à construção e ao uso de um modelo para avaliar a classe de um objeto não rotulado ou para estimar o valor de um ou mais atributos de dado objeto. No primeiro caso, denominamos a tarefa de *classificação* e, no segundo, denominamos de *regressão* (em estatística) ou simplesmente *estimação*. Sob essa perspectiva, classificação e estimação constituem os dois principais tipos de problemas de predição, sendo que a classificação é usada para prever *valores discretos*, ao passo que a estimação é usada para prever *valores contínuos*.

## 3 – Análise de grupos

Agrupamento (clustering) é o nome dado ao processo de separar (particionar ou segmentar) um conjunto de objetos em grupos (do inglês clusters) de objetos similares. Diferente da classificação os dados que irão gerar os grupos, normalmente não tem uma classe, o algoritmo em si irá definir a que grupo cada item pertence.



# Introdução

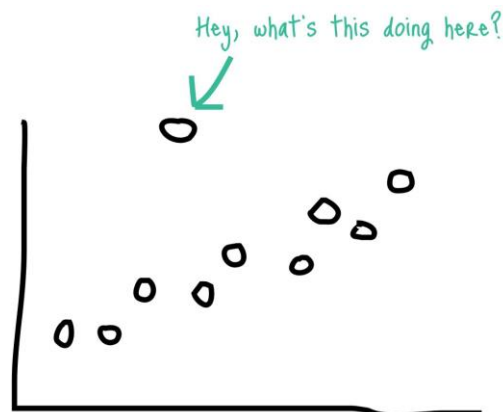
Principais tarefas da mineração de dados?

## 4 – Associação

Nas análises de grupos e preditivas, o objetivo em geral é encontrar relações (grupos, classes ou estimativas) entre os objetos da base. Entretanto, há diversas aplicações práticas nas quais o objetivo é encontrar relações entre os atributos (ou variáveis), e não entre os objetos. Para ilustrar esse caso, vamos considerar uma aplicação típica em marketing: a análise de carrinho de supermercado. Nesse tipo de análise, há um conjunto de transações (pedidos ou compras), e o objetivo é encontrar itens (produtos) que são comprados em conjunto; nesse sentido, as transações correspondem aos objetos da base e os itens, aos atributos.

## 5 – Detecção de anomalias (outliers)

Uma base de dados pode conter objetos que não seguem o comportamento ou não possuem a característica comum dos dados ou de um modelo que o representa. Esses dados são reconhecidos como anomalias ou valores discrepantes (outliers).





# Introdução



## As diferentes Nomenclaturas

### 1 – Inteligência Artificial Clássica

Ciência e Engenharia de máquinas inteligentes.

### 2 – Inteligência Computacional

Área criada para englobar as redes neurais artificiais, os algoritmos evolutivos e os sistemas fuzzy.

### 3 – Aprendizado de Máquina

Visa desenvolver programas computacionais capazes de melhorar seu desempenho por meio da experiência.

#### 3.1 – Aprendizado supervisionado

É baseado em um conjunto de objetos para os quais as saídas desejadas são conhecidas, ou em algum outro tipo de informação que represente o comportamento que deve ser apresentado pelo sistema.

#### 3.2 – Aprendizado não supervisionado

É baseado apenas nos objetos da base, cujos rótulos são desconhecidos. Basicamente, o algoritmo deve aprender a “categorizar” ou rotular os objetos.

# Introdução

- Four Fallacies of Data Mining (Louie, Nautilus Systems, Inc.)

Falácias		Realidade
1	<ul style="list-style-type: none"><li>• Um conjunto de ferramentas pode ser solto em repositórios de dados.</li><li>• Encontra respostas para todos os problemas de negócios</li></ul>	<ul style="list-style-type: none"><li>• Nenhuma ferramenta automática de mineração de dados resolve problemas</li><li>• Em vez disso, a mineração de dados é um processo</li><li>• Integra-se aos objetivos gerais de negócios</li></ul>
2	<ul style="list-style-type: none"><li>• O processo de Mineração de Dados é autónomo.</li><li>• Requer pouca supervisão.</li></ul>	<ul style="list-style-type: none"><li>• Requer intervenção significativa durante todas as fases</li><li>• Após a implantação do modelo, novos modelos exigem atualizações</li><li>• Medidas avaliativas contínuas monitoradas por analistas</li></ul>
3	<ul style="list-style-type: none"><li>• A mineração de dados se paga rapidamente.</li></ul>	<ul style="list-style-type: none"><li>• Taxas de retorno variam, dependendo da inicialização, pessoal, preparação de dados custos, etc.</li></ul>
4	<ul style="list-style-type: none"><li>• Software de mineração de dados fácil de usar.</li></ul>	<ul style="list-style-type: none"><li>• Facilidade de uso varia entre projetos</li><li>• Os analistas devem combinar o conhecimento do assunto com domínio de problema específico</li></ul>

# Introdução

## Exemplos

TECH | 2/16/2012 @ 11:02AM | 2,316,057 views

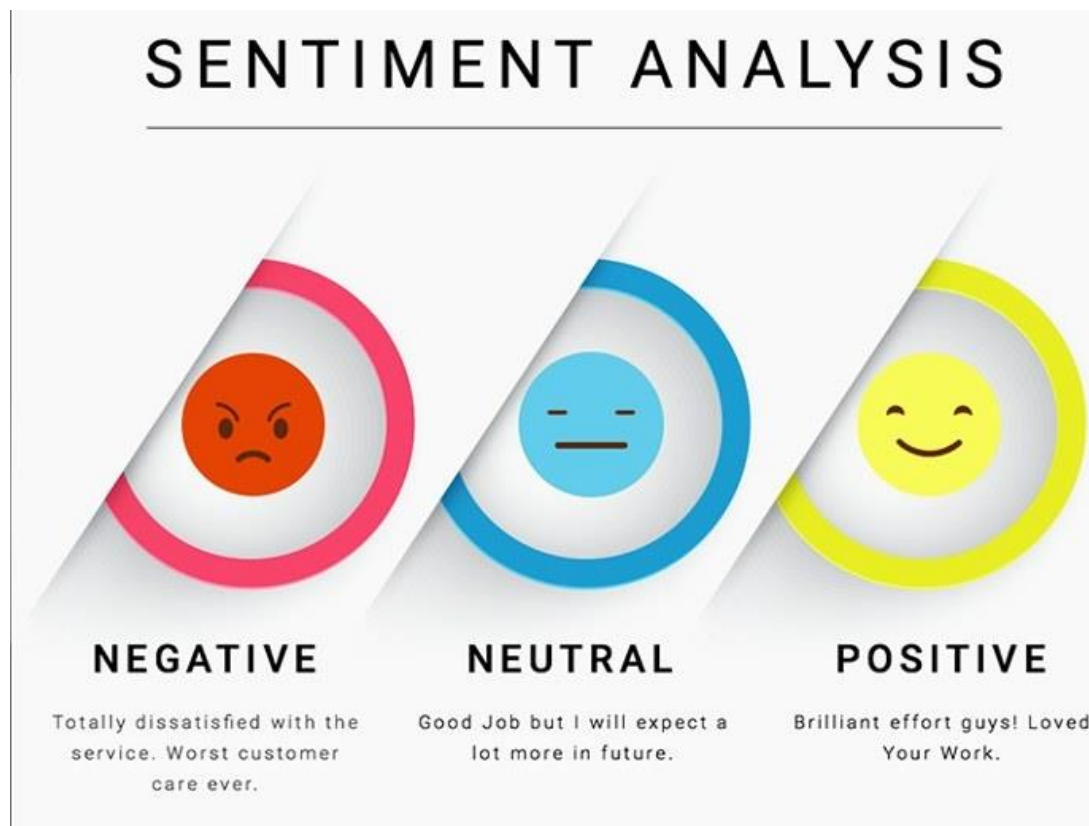
## How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



<https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

# Introdução

## Exemplos



# Introdução

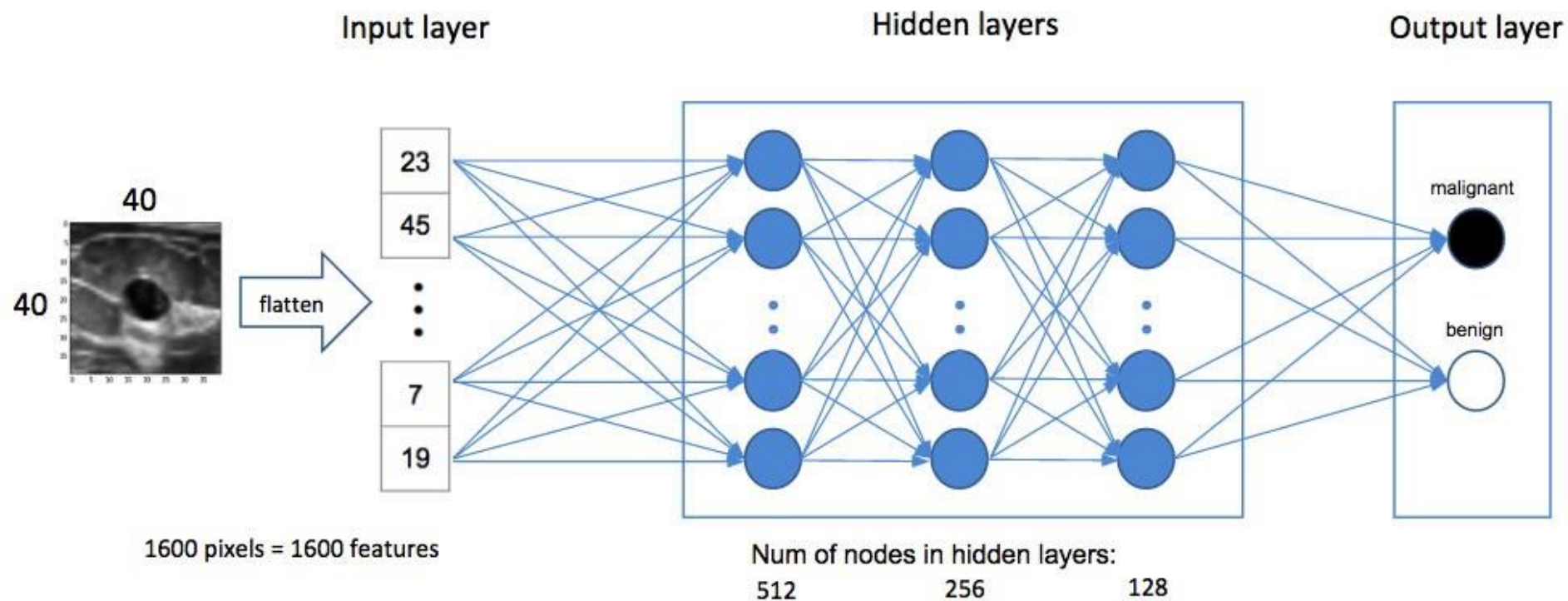
Exemplos





# Introdução

Exemplos (deep learning)



# Referências

Introdução à Mineração de Dados. Conceitos Básicos, Algoritmos e Aplicações

por [Leandro Nunes de Castro Silva](#) (Autor), [Daniel Gomes Ferrar](#) (Autor), [Patricia Quero](#) (Editor)

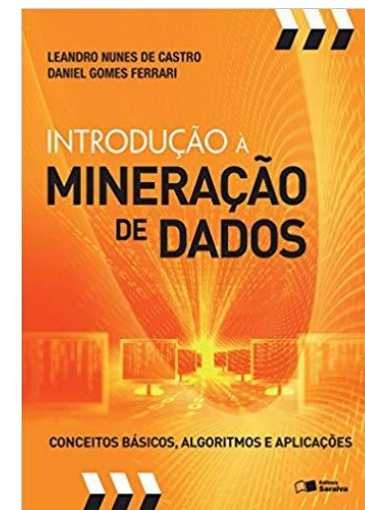
Dean Abbott (2014) “Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst,” Wiley

Daniel T. Larose, Chantal D. Larose (2015) “Data Mining and Predictive Analytics,” 2nd Edition, Wiley

Mitchell, T., (1997) “Machine Learning,” McGraw Hill.

Hand, D. J., Mannila, H., Smyth, P. (2001) “Principles of Data Mining (Adaptive Computation and Machine Learning),” MIT Press.

Python Programming: An Introduction to Computer Science, 1st Edition





Real Python