



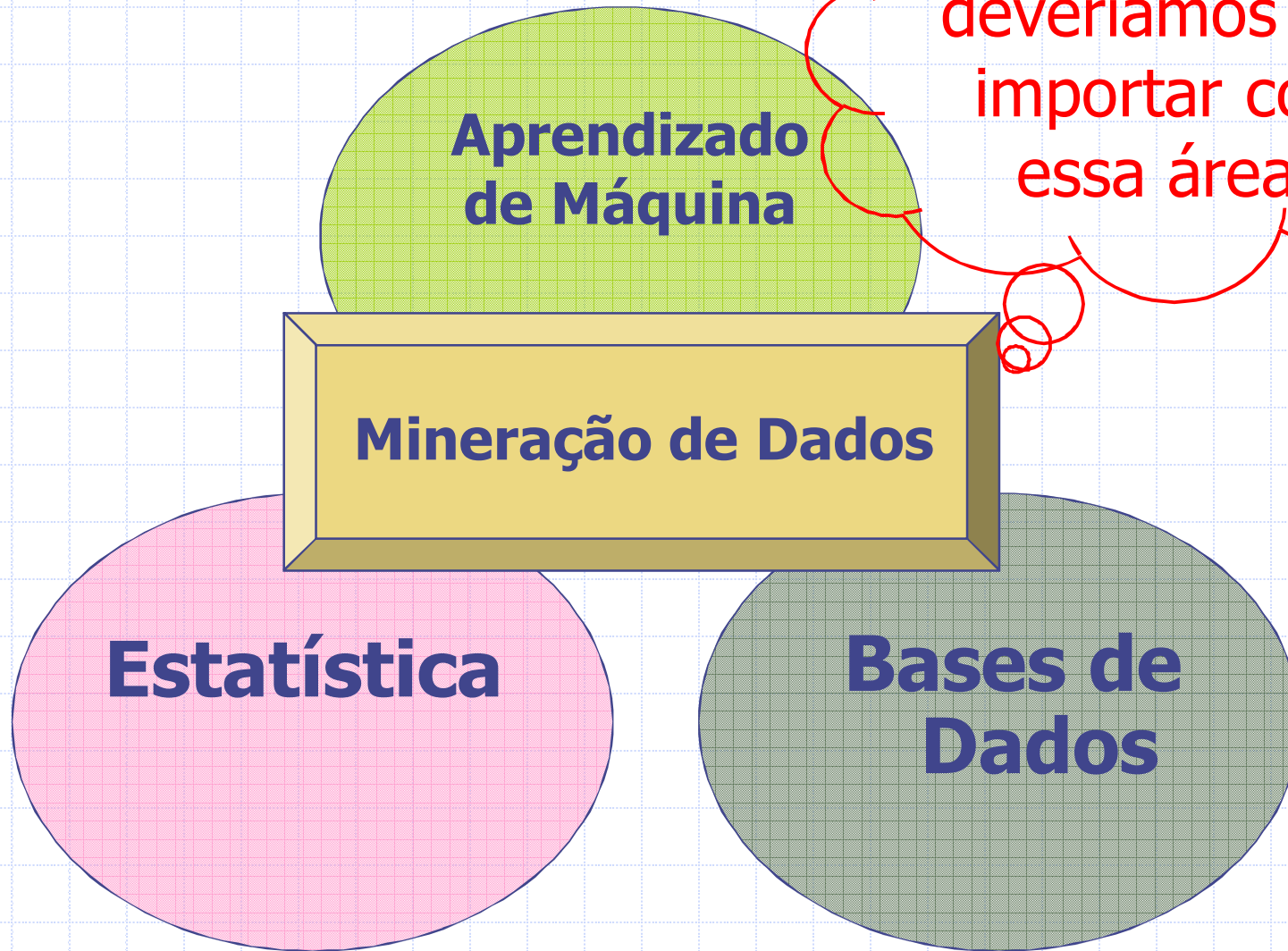
# **Mineração de Dados: Panorama, Aplicações e Tendências**

**Eduardo R. Hruschka**

# Mapa do Território

- 1) Visão Geral sobre Mineração de Dados (*Data Mining, Knowledge Discovery from Databases, Data Science, Predictive Analytics, Big Data*)
- 2) Agrupamento de Dados ("Clusterização")
- 3) Classificação
- 4) Sistemas de Recomendação (Regressão)
- 5) Tendências Futuras e Impactos na Sociedade

# 1. Visão Geral





# Alguns (poucos) exemplos...



- Taxas de *clicks* em torno de 0.05% - R\$ 12 bilhões (2013);
- 30% de *smartphones* e *tablets*;
- 1 bilhão de usuários (65 milhões no Brasil),  $>10^9$  posts/dia.

## Alguns (poucos) exemplos...



*Cortesia - Prof. Marko Grobelnik*

Comunicações entre diferentes usuários:

- > 100 bilhões de amizades (130 amigos em média);
- > 2 bilhões de "curtir/comentários" por dia;



# Como medir similaridades?

- Em geral, trata-se de um problema difícil:



- Abordagens matemáticas (e.g., cosseno entre dois documentos) são comumente adotadas;
- Técnicas de agrupamento de dados são úteis...

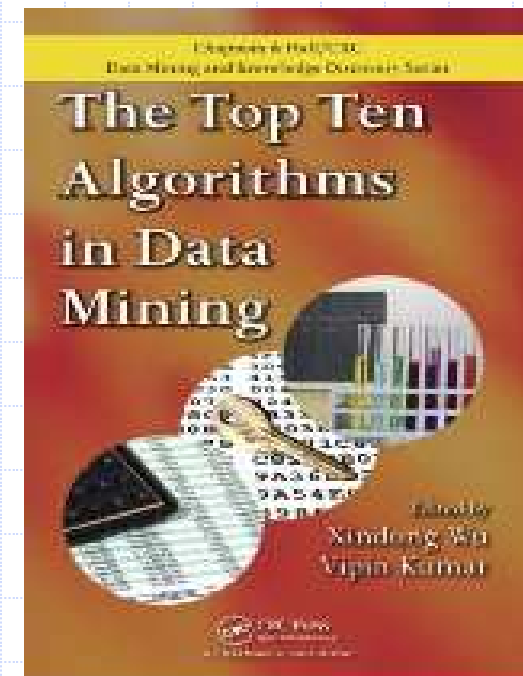


## 2. Agrupamento de Dados

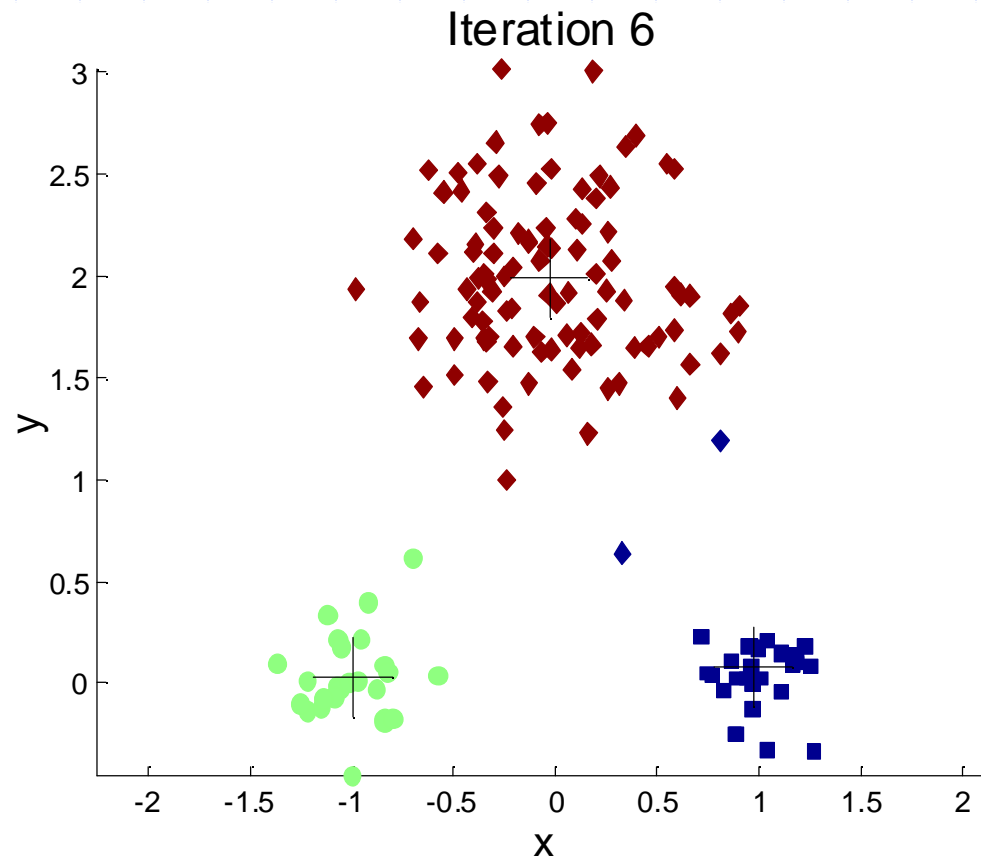
- Encontrar grupos (*clusters*) de dados similares;
- Diversas aplicações reais – análise exploratória de dados: mineração de textos, *marketing* (segmentação de clientes), recuperação de informação, reconhecimento de padrões, ...

Ideia geral e intuitiva por meio de um exemplo pedagógico:

- Algoritmo K-means (MacQueen, 1967; Kulis & Jordan, 2012)

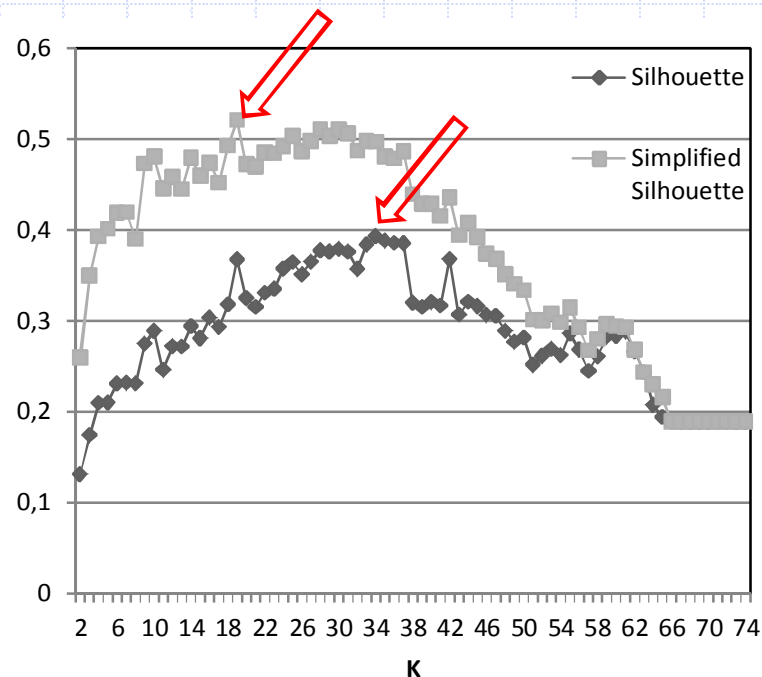


### Exemplo para K-means (K=3):



## 2. Agrupamento ...

- Otimização convexa para cada  $K$  : converge para ótimos locais com diversas medidas de distância...Bacana, mas:
  - Sensível à inicialização;
  - Como estimar  $K$  a partir dos dados?
- Rodar K-means várias vezes para diferentes valores de  $K$ ;
- Problema de otimização multi-modal:



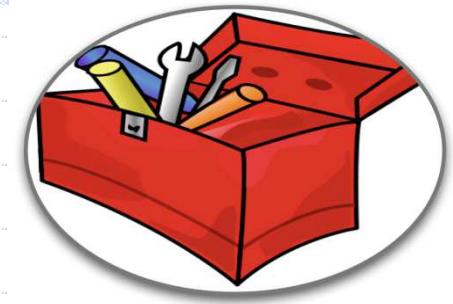
Número de grupos em torno de 20-30;  
Agrupamento de documentos para análise forense (Nassif & Hruschka, 2012):

- Computadores apreendidos pela PF;
- Facilitar análise do perito — inspecionar grupos de documentos.

## 2. Agrupamento ...

- Tarefa computacionalmente custosa;
- Algoritmos de busca probabilística (EAs, PSO, SA, ...);
- Abordagens que combinam diferentes algoritmos de otimização (Hruschka et al., 2009): busca local + busca global:
  - Idealmente dependentes de poucos parâmetros *não críticos* definidos pelo usuário.
- Extensões para Modelos de Misturas de Gaussianas:
  - *Expectation-Maximization* (EM) + complexidade do modelo.
- Modelos Gráficos Probabilísticos (Blei, 2012).

- Muitos algoritmos disponíveis;
- Caixa de ferramentas (compacta)?



- K-means, Bisecting K-Means, K-medoids;
- Índices para estimar K (e.g., silhueta);
- *Cluster ensembles*;
- Expectation-Maximization (EM) para modelos de misturas de Gaussianas.

### 3. Classificação

- Fraude: financeira, comércio eletrônico, seguros, ...
- Resulta em perdas de bilhões de Reais por ano;
- Como detectar automaticamente?



**Prevenção de fraude em tempo real:** esta transação é fraudulenta?

- a) Senha (ajuda em alguns casos)
- b) Sistema gera um escore baseado em fatores que qualificam fraude
- c) Poucos segundos para tomar decisão

- Erros de classificação custam caro
- Requer modelos estatísticos

### 3. Classificação ...

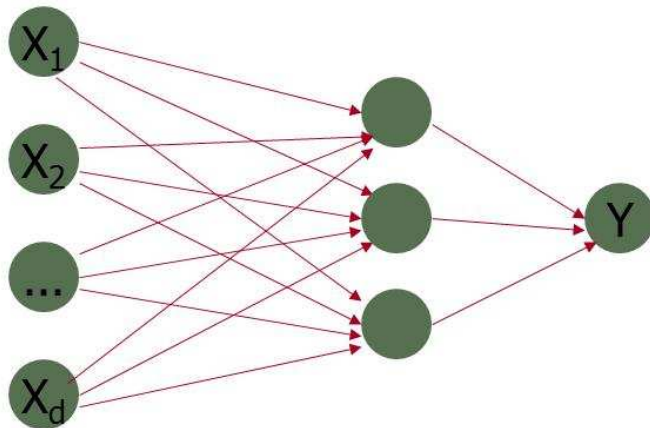
- Construindo classificadores automáticos  $Y=f(X_1, X_2, \dots, X_d)$ :

$X_1$	$X_2$	...	$X_d$	Y (classe)
...	...	...	...	...
...	...	...	...	...

$Y \in \{\text{fraude, normal}\}$  (menos do que 1% de transações fraudulentas)

$X = \{X_1, X_2, \dots, X_d\}$ : variáveis descrevendo as transações

- Diversos modelos – *e.g.*, redes neurais:



#### Passos principais:

- 1) Aprender/ajustar os parâmetros do modelo a partir de uma amostra de transações (algoritmos de otimização);
- 2) Predição de classes (Y) para novas transações baseando-se em  $\{X_1, X_2, \dots, X_d\}$

### 3. Classificação ...

- Na prática, usualmente vários modelos diferentes são combinados (*classifier ensembles*) tal como se formassem um comitê de especialistas;
  - Premissa: dados são *independent and identically distributed* (i.i.d.);
  - Violada no mundo real: amostras são polarizadas (formadas apenas por aqueles que foram *pegos*);
  - Distribuições de probabilidades de classe podem mudar (entre treinamento e uso dos modelos);
- Visão geral de um método que permite refinar distribuições de probabilidade de classe.

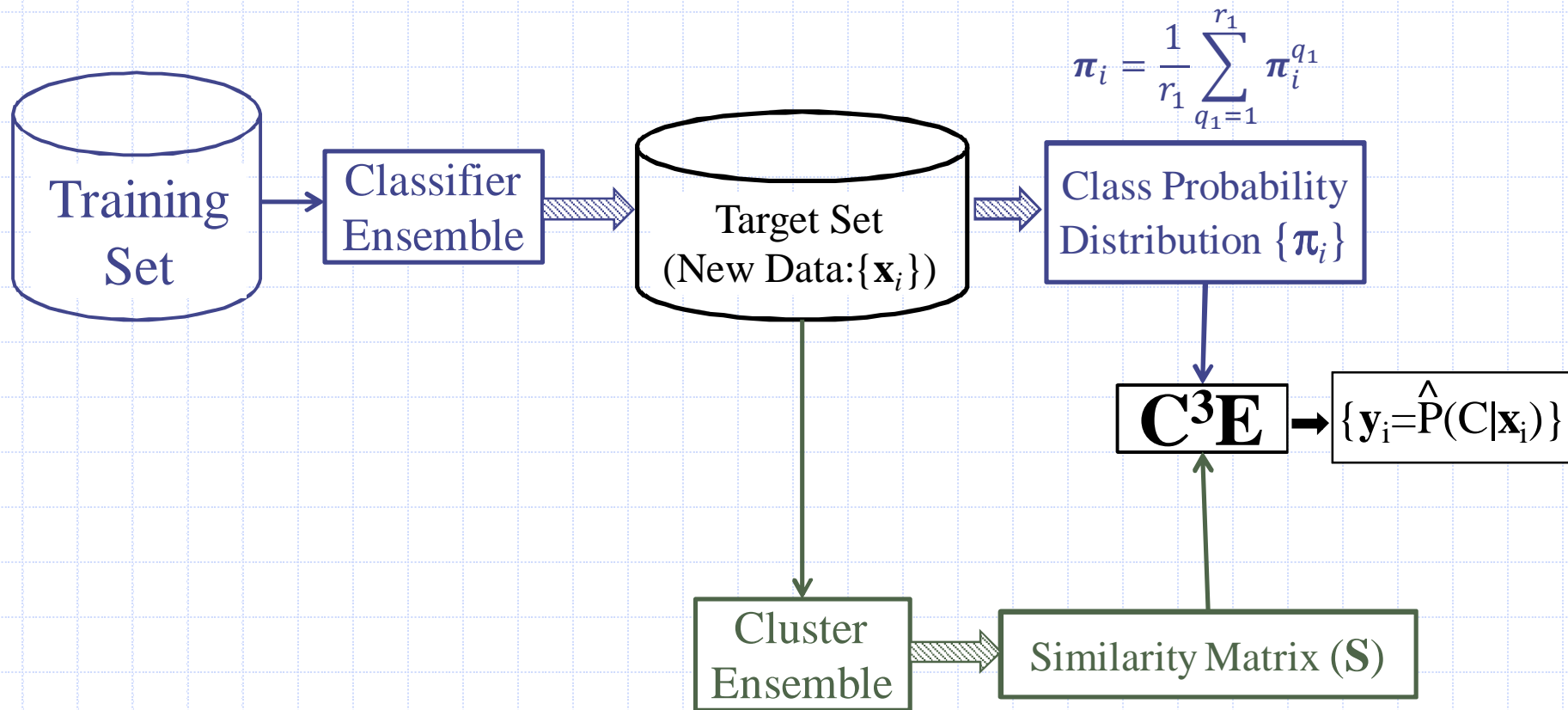


### 3. Classificação ...

- Combinar agregadores de classificadores e agrupadores:
  - Modelos não supervisionados podem fornecer restrições para classificação de novos dados:
    - ✓ Objetos semelhantes no conjunto alvo provavelmente são da mesma classe.
  - Melhorar acurácia preditiva, especialmente quando dados rotulados para treinamento são escassos.
  - Projetar métodos de aprendizado de máquina que sejam conscientes das possíveis diferenças entre distribuições de treinamento e do conjunto alvo.

### 3. Classificação ...

#### *Combining Classifier and Cluster Ensembles (C<sup>3</sup>E):*



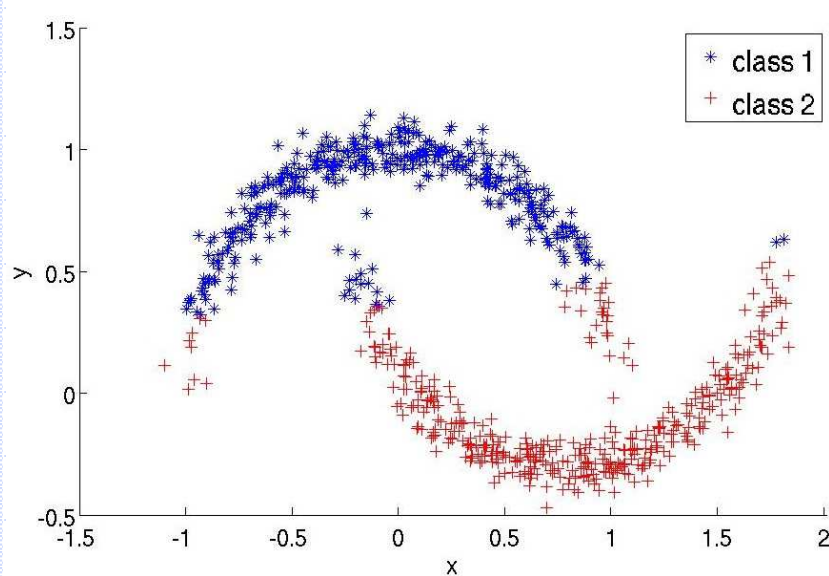
$$\pi_i = \frac{1}{r_1} \sum_{q_1=1}^{r_1} \pi_i^{q_1}$$

Algoritmo de otimização re-estima probabilidades de classes via similaridades entre novos objetos.

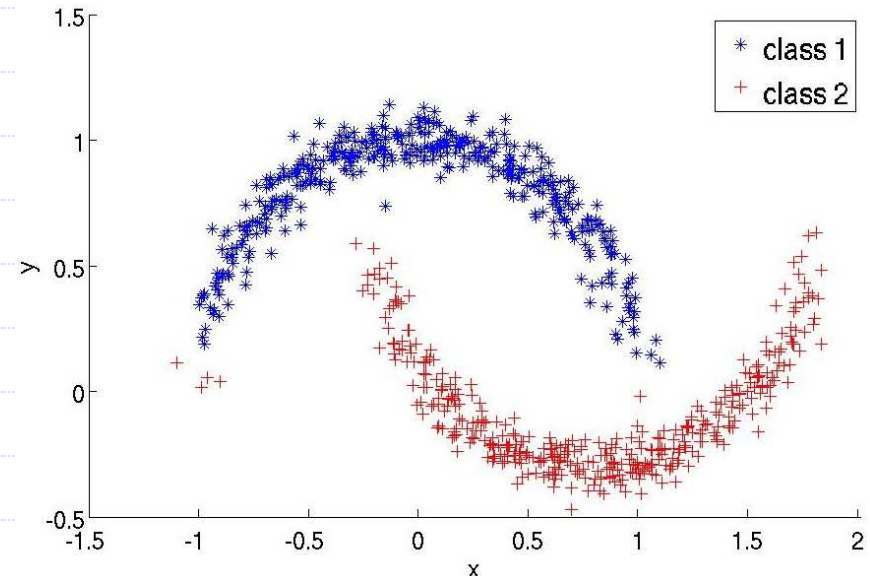
(e.g., proporção das  $r_2$  partições nas quais 2 objetos (transações) pertencem ao mesmo grupo (cluster))

### 3. Classificação ...

**Exemplo pedagógico** (somente 2% de objetos rotulados):



**Classifier Ensemble**



**C<sup>3</sup>E**

- Aplicações de sucesso em categorização de textos e classificação de imagens;
- Outras potenciais aplicações ?

### 3. Classificação ...

- Outras aplicações de classificação incluem:
  - *Churn prediction* (cliente abandona serviço/produto): telefonia, tv a cabo, assinatura de revista, cartão de crédito, etc.
  - Classificação de spam;
  - Recrutamento de profissionais (RH);
  - Abandono de emprego;
  - Análise de sentimentos (redes sociais);
  - Finanças (cliente irá cumprir contrato de financiamento?, cliente irá atrasar o pagamento do cartão de crédito?)
  - Bioinformática, Medicina, etc...
    - Requer dados de boa qualidade;
    - Diferentemente do trabalho típico de um estatístico mais tradicional, dados não foram especificamente coletados com o propósito de modelagem.

- Muitos algoritmos disponíveis;
- Caixa de ferramentas (compacta)?



- Naïve Bayes (*wrapper*);
- Árvores de Decisão e *Random Forests*;
- Regressão Logística;
- *Classifier Ensembles*;
- Engenharia de atributos (feature selection);
  - SVMs, redes neurais, etc.

# 4. Sistemas de Recomendação

The screenshot shows the Amazon.com product page for the book "Principles of Data Mining (Adaptive Computation and Machine Learning)" by David J. Hand, Rajski Mannil, and Padhraic Smyth. The page is viewed in a browser window with the address bar showing the product URL. The top navigation bar includes search, account, and cart options. The main content area features a book cover, pricing information, and a "Buy New" button. A sidebar on the right offers additional purchase options like Kindle Edition and gift cards. Below the main product information, there is a "Book Description" section, "Special Offers and Product Promotions", and a "Frequently Bought Together" section. At the bottom, a "Customers Who Bought This Item Also Bought" section displays a grid of related books.

**Principles of Data Mining (Adaptive Computation and Machine Learning) [Hardcover]**  
David J. Hand (Author), Rajski Mannil (Author), Padhraic Smyth (Author)  
37 customer reviews

**Buy New**  
\$64.54 & FREE Shipping. Details

**Rent**  
\$19.55

**Only 10 left in stock (more on the way).**  
Ships from and sold by Amazon.com. Gift-wrap available.

Want it Tuesday, Aug. 27? Order within 50 hrs 57 mins and choose One-Day Shipping at checkout. [Details](#)

32 new from \$40.00 38 used from \$22.00

**FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS**  
Eligible items only. [Learn more](#)

Format	Amazon Price	New from	Used from
Kindle Edition	\$29.00		
Hardcover	\$64.54	\$40.00	\$22.00

**SELL YOUR BOOKS**  
Sell Back Your Copy for \$9.00. No matter where you bought them, get up to 70% back when you sell your books at Amazon.com.

Used Price	Trade-In Price
\$22.00	\$9.00
\$13.00	\$13.00

**Book Description**  
Publication Date: August 1, 2001 | ISBN-10: 026208290X | ISBN-13: 978-0262082907

The growing interest in data mining is motivated by a common problem across disciplines: how does one store, access, model, and ultimately describe and understand very large data sets? Historically, different aspects of data mining have been addressed independently by different disciplines. This is the first truly interdisciplinary text on data mining, blending the contributions of information science, computer science, and statistics.

The book consists of three sections. The first, foundations, provides a tutorial overview of the principles underlying data mining algorithms and their application. The presentation emphasizes intuition rather than rigor. The second section, data mining algorithms, shows how algorithms are constructed to solve specific problems in a principled manner. The algorithms covered include trees and rules for classification and regression, association rules, belief networks, classical statistical models, nonlinear models such as neural networks, and local "memory-based" models. The third section shows how all of the preceding analysis fits together when applied to real-world data mining problems. Topics include the role of metadata, how to handle missing data, and data preprocessing.

**Special Offers and Product Promotions**  
Buy \$25 or more in Textbooks, get a \$3 Amazon MP3 Credit. [View details](#) (restrictions apply)

**Frequently Bought Together**  
Price for all three: **\$159.97**  
[Add all three to Cart](#) [Add all three to Wish List](#)  
[View availability and shipping details](#)

**This Item:** Principles of Data Mining (Adaptive Computation and Machine Learning) by David J. Hand. Hardcover. \$64.54

**Also Bought:**  
Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) by Ian H. Witten. Paperback. \$41.28  
Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han. Hardcover. \$94.17

**Customers Who Bought This Item Also Bought**

Book Title	Author	Price	Rating
The Elements of Statistical Learning	Trevor Hastie	\$35.00	★★★★ (25)
Data Mining: Practical Machine Learning Tools and Techniques	Ian H. Witten	\$41.28	★★★★★ (21)
Data Mining: Concepts and Techniques, Third Edition	Jiawei Han	\$94.17	★★★★★ (15)
Data Mining: Practical Machine Learning Tools and Techniques, Third Edition	Ian H. Witten	\$41.28	★★★★★ (21)
Data Mining with R: Learning with Case Studies	Lutz Fong	\$35.00	★★★★★ (11)
Pattern Recognition and Machine Learning	Christopher B. Bishop	\$45.00	★★★★★ (77)
The Elements of Statistical Learning	Trevor Hastie	\$35.00	★★★★★ (25)
R Cookbook (O'Reilly)	Robert I. Ikin	\$35.00	★★★★★ (28)
Python for Data Analysis	Wes McKinney	\$35.00	★★★★★ (12)
Learning From Data	Thomas G. Dietterich	\$35.00	★★★★★ (24)
Machine Learning: The Art and Science of Probabilistic Modeling	David Blei	\$35.00	★★★★★ (9)
Mastering Data Mining: The Art and Science of Data Mining	John P. Hughey	\$35.00	★★★★★ (15)
Machine Learning: A Probabilistic Perspective	John D. Lafferty	\$35.00	★★★★★ (15)

Page 1 of 6

Principles of Data M... Tese\_Livre\_Docenci... Roteiro para Entrevi... Microsoft PowerPoi...

EN 17:02

## 4. Sistemas de Recomendação ...

Suposto usuário



The screenshot shows the Netflix website interface. At the top, there is a navigation bar with the Netflix logo and links for Home, Just for Kids, Genres, and Taste Prof. A search bar is visible on the right. Below the navigation bar, a message reads: "Hey Eduardo, want some quick suggestions? Rate this title...". To the left of this message is a movie poster for "HOTEL FOR DOGS". To the right of the poster is a rating section with the text "What did you think?" and five stars. Below the stars are three large boxes, each containing a question mark, representing a rating scale. Below the rating section, there is a question: "How often do you watch Cats & Dogs?" with three radio button options: "Never", "Sometimes", and "Often". Below this, there are two sections: "Recently Watched" and "Popular on Netflix". The "Recently Watched" section shows the "HOTEL FOR DOGS" poster. The "Popular on Netflix" section shows a row of posters for "TUFF PUPPY", "TEAM UMIZOOMI", "VINNY CUNYON", "BUCKLE UP!", "JUSTIN TIME", "JACK AND THE BEANSTALK", "POWER RANGERS SUPER SAMURAI", and "Doc McStuffins". At the bottom of the screenshot, the Windows taskbar is visible, showing several open applications: "Netflix - Google Chr...", "Tese\_Livre\_Docenci...", "Roteiro para Entrevi...", "Microsoft PowerPoi...", and "Amazon\_Picture - P...". The system clock shows the time as 17:08.

### ➤ **Agrupamento e Regressão Simultâneos**

- Em problemas difíceis de classificação/regressão, frequentemente se segmenta a base de dados em grupos relativamente homogêneos;
- Posteriormente, constrói-se um modelo por grupo;
- Tal procedimento usualmente proporciona bons resultados com modelos mais simples e interpretáveis;
- SCOAL: *Simultaneous CO-clustering and Learning* (Deodhar and Ghosh, 2010);
- Permite modelagem preditiva de dados em grande escala;
- Aprender modelos locais a partir dos grupos:



## 4. Sistemas de Recomendação ...

Dados de usuários

Dados de filmes

	95- ação/aventura/thriller	94- comédia	95- policial/drama/thriller	89- romance/comédia	95- ação	94- comédia	95- policial/thriller	97- romance/drama	94- ação/thriller	94- comédia/romance	95- policial/thriller	95- romance/drama	95- ação/drama/thriller	88- comédia	94- policial/drama/rom./thriller	95- comédia/romance	96- ação/thriller	79- comédia	95- policial	96- romance/drama	
24 anos - M - técnico	1	5	2	1	?	5	1	?	?	4	2	?	?	4	2	?	?	4	?	1	?
53 anos - F - outros	2	4	2	?	1	4	?	?	1	5	?	?	1	5	?	?	1	4	?	?	2
23 anos - M - escritor	3	5	?	1	2	5	?	?	?	4	?	?	?	?	?	?	?	4	?	?	?
24 anos - M - técnico	4	5	?	?	4	5	?	?	5	5	?	?	5	5	?	?	?	4	?	?	4
33 anos - F - outros	5	5	?	?	?	5	?	?	5	4	?	?	5	5	?	?	?	4	?	?	4
42 anos - M - executivo	6	?	5	4	?	?	4	3	?	?	5	4	?	?	4	5	?	?	5	5	?
57 anos - M - admin	7	?	5	?	?	?	4	?	?	?	3	?	?	?	?	?	?	?	?	?	?
36 anos - M - admin	8	?	?	5	?	?	?	5	?	?	5	?	?	?	?	?	?	?	?	?	?
29 anos - M - estudante	9	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	5
53 anos - M - advogado	10	5	?	?	?	5	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

Nota/preferência do usuário pelo item (filme)

## 4. Sistemas de Recomendação ...

Escolher aleatoriamente alguns grupos de linhas e colunas

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grupo de Linha 1																				
Grupo de Linha 2																				

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5	2	1	?	5	1	?	?	4	2	?	?	4	2	?	?	4	?	1	?
2	4	?	?	1	4	?	?	1	5	?	?	1	5	?	?	1	4	?	?	2
3	5	?	?	?	?	?	?	?	4	?	?	?	?	?	?	?	4	?	?	?
4	5	?	?	4	5	?	?	5	5	?	?	5	5	?	?	?	4	?	?	4
5	5	?	?	?	?	?	?	5	4	?	?	5	?	?	?	?	4	?	?	4
6	?	5	4	?	?	4	3	?	?	5	4	?	?	4	5	?	?	5	5	?
7	?	5	?	?	?	4	?	?	?	3	?	?	?	?	?	?	?	?	?	?
8	?	?	5	?	?	?	5	?	?	?	5	?	?	?	?	?	?	?	?	?

- Treinar 4 modelos de regressão (um por grupo).

## 4. Sistemas de Recomendação ...

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5	2	1	?	5	1	?	?	4	2	?	?	4	2	?	?	4	?	1	?
2	4	2	?	1	4	?	?	1	5	?	?	1	5	?	?	1	4	?	?	2
3	5	?	1	?	5	?	?	?	4	?	?	?	?	?	?	?	4	?	?	?
4	5	?	?	?	?	?	?	5	5	?	?	5	5	?	?	?	4	?	?	4
5	5	?	?	?	5	?	?	5	4	?	?	5	5	?	?	?	4	?	?	4
6	?	5	4	?	?	4	3	?	?	5	4	?	?	4	5	?	?	5	5	?
7	?	?	?	?	?	4	?	?	?	3	?	?	?	?	?	?	?	?	?	?
8	?	?	?	?	?	?	5	?	?	?	5	?	?	?	?	?	?	?	?	?

➤ Um modelo por bicluster:

$$x_{ij} = [1, u, v]$$

$$\beta^T = [\beta^0, \beta_i^T, \beta_j^T]$$

$$\text{Compute } \hat{z}_{ij} = \beta^T x_{ij}$$

$$MSE = \sum_{ij} w_{ij} (z_{ij} - \hat{z}_{ij})^2$$

$$MSE: 0.2887$$

## 4. Sistemas de Recomendação ...

### Atualizar grupos de linhas:

Mover linhas para grupos que minimizam o erro de predição

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grupo 1																				
4	5	?	?	4	5	?	?	5	?	?	5	5	?	?	?	4	?	?	4	
5	5	?	?	?	5	?	?	5	4	?	?	5	5	?	?	?	4	?	?	4
6	?	5	4	?	?	4	3	?	?	5	4	?	?	4	5	?	?	5	5	?
Grupo 2																				
7	?	5	?	?	?	4	?	?	?	3	?	?	?	?	?	?	?	?	?	?
8	?	?	5	?	?	?	5	?	?	?	5	?	?	?	?	?	?	?	?	?

- Linha 3 obtém menor MSE ao ser predita pelos modelos do grupo 2:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grupo 1																				
1	5	2	1	?	5	1	?	?	4	2	?	?	4	2	?	?	4	?	1	?
2	4	2	?	1	4	?	?	1	5	?	?	1	5	?	?	1	4	?	?	2
3	5	?	1	2	5	?	?	?	?	?	?	?	?	?	?	4	?	?	?	?
Grupo 2																				
4	5	?	1	2	5	?	?	?	4	?	?	?	?	?	?	4	?	?	?	?

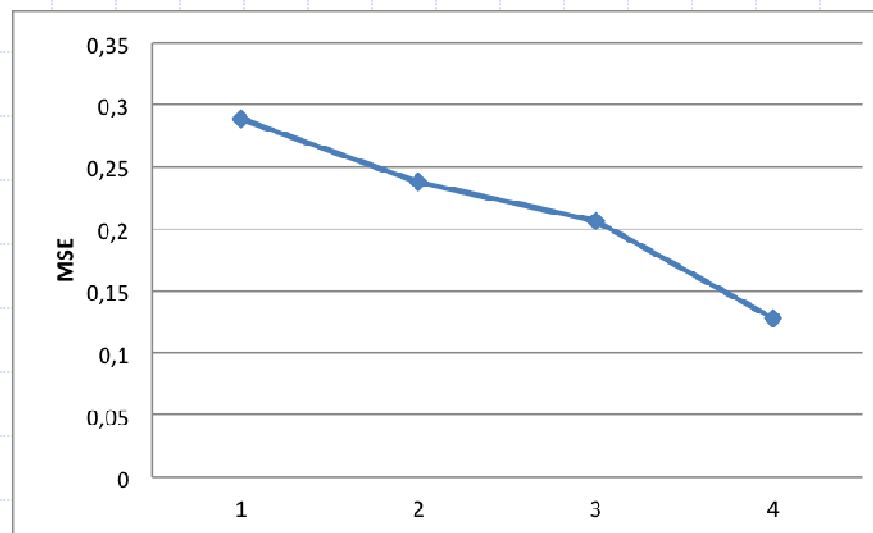
- Repetir o processo para cada linha/coluna...

Após o grande laço, re-estimar os modelos:

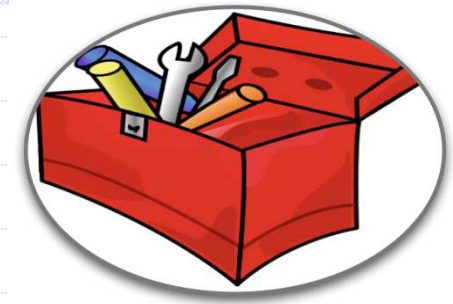
	1	2	3	4	5	7	9	15	19	20	8	10	11	12	13	14	16	17	19	8
1	5	2	1	?	5	?	4	?	1	?	?	2	?	?	4	2	?	4	?	1
2	4	2	?	1	4	?	5	?	?	2	1	?	?	1	5	?	1	4	?	?
4	5	?	?	?	?	?	?	?	?	4	5	?	?	?	?	?	?	?	?	?
5	5	?	?	?	5	?	4	?	?	4	5	?	?	5	?	?	?	4	?	?
6	?	5	4	?	?	?	?	5	5	?	?	5	4	?	?	4	?	?	5	4
7	?	5	?	?	?	?	?	?	?	?	?	3	?	?	?	?	?	?	?	4
8	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
3	5	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

MSE: 0.23758

- Mover linhas/colunas para os grupos que minimizam o erro;
- MSE global é (garantidamente) minimizado nas iterações:



- Muitos algoritmos disponíveis;
- Caixa de ferramentas (compacta)?



- Regressão linear;
- Árvores de regressão / *random forests* ;
- Modelos lineares generalizados;
- Outros modelos não lineares (e.g., redes neurais).

## 5. Tendências e Impactos

- Tendências no projeto de algoritmos para *big data*:
  - Combinar diferentes algoritmos de otimização;
  - Diminuir o número de parâmetros críticos definidos pelo usuário via ajuste automático (a partir dos dados);

*"Essentially, all models are wrong, but some are useful."*

(George E. P. Box, Professor Emeritus, University of Wisconsin)

- Crescente número de novas aplicações;
- Gartner (líder em TI) prevê que em 2015 a demanda por profissionais de *big data* será de 4.4 milhões.
- Questões éticas do uso de modelos automáticos.

# **Agradecimentos:**

## **USP**

Thiago F. Covões, Luiz F. S. Coletta, André P. Vizine

## **University of Texas at Austin**

Joydeep Ghosh, Ayan Acharya, Sreangsu Acharyya



# Referências

- Kulis & Jordan, Revisiting k-means: New Algorithms via Bayesian Nonparametrics, Proc. of ICML 2012, Edinburgh, Scotland.
- Nassif & Hruschka, Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection. IEEE Transactions on Information Forensics and Security, to appear.
- Hruschka et al., A Survey of Evolutionary Algorithms for Clustering. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews, 2009.
- Blei, Probabilistic topic models, Communications of the ACM, 2012.
- Oza & Tumer, Classifier ensembles: Select real-world applications, 2008.
- Wang et al., Bayesian cluster ensembles. Statistical Analysis and Data Mining, 2011.
- Acharya et al., C<sup>3</sup>E: A Framework for Combining Ensembles of Classifiers and Clusterers. Proc. of 10th Int. Workshop on Multiple Classifier Systems, 2011.
- Deodhar and Ghosh, SCOAL: A Framework for Simultaneous Co-Clustering and Learning from Complex Data, ACM Transactions on Knowledge Discovery from Data, 2010.