Aprendizado de Máquina

Aula 1 - Introdução

Eduardo R. Hruschka

Agenda:

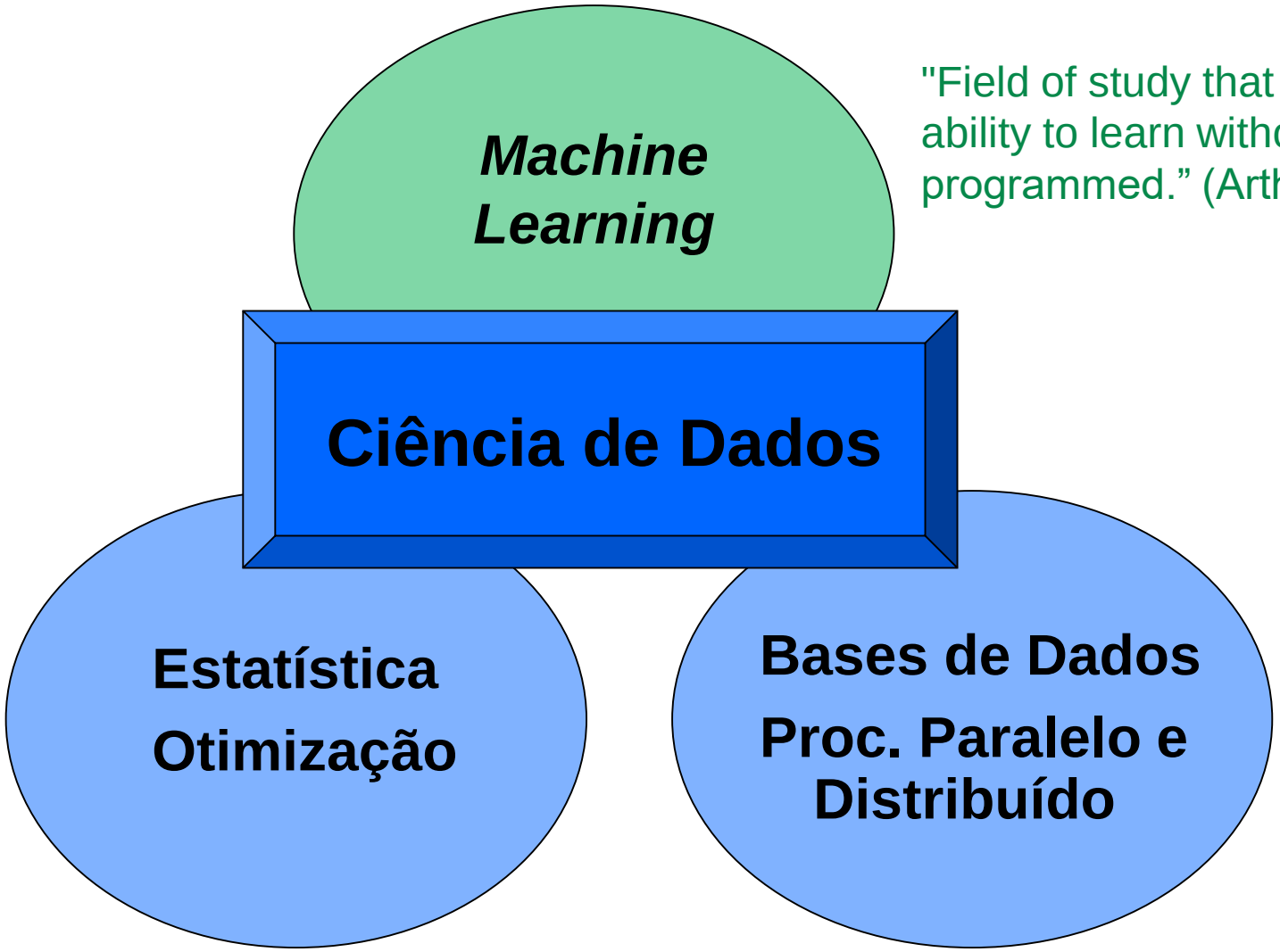
- **Aprendizado de Máquina e Ciência de Dados**
- **Noções sobre Agrupamento**
- **Noções sobre Classificação**
- **Noções sobre Regressão**
- **Tendências e Desafios**

Visão sobre Ciência de Dados

Big Data Analytics = Ciência de Dados

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Tom Mitchell ← Alan Turing,1950)

"Field of study that gives computers the ability to learn without being explicitly programmed." (Arthur Samuel,1959)



Foco no negócio

Técnicas e Aplicações I: Agrupamento de Dados

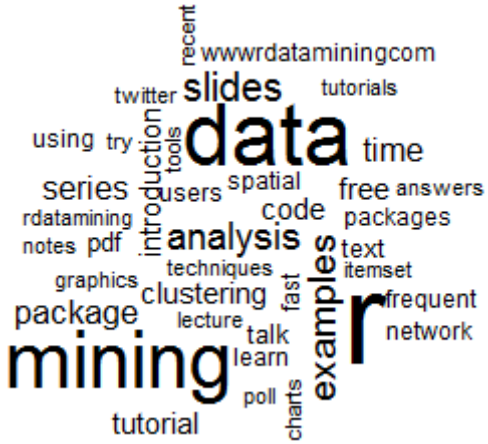
Técnicas e Aplicações I (agrupamento de dados)

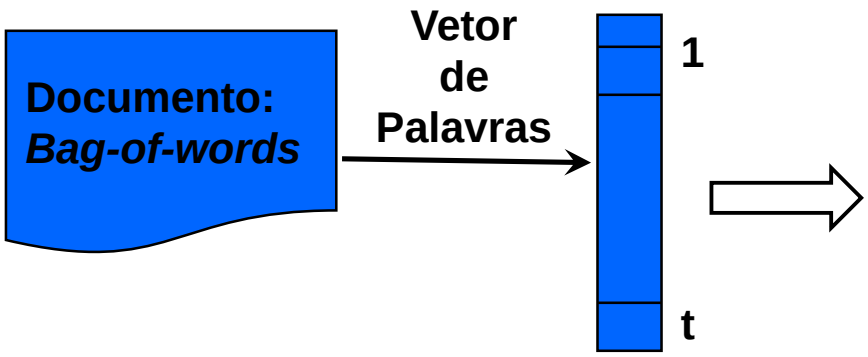
- Agrupar dados semelhantes/parecidos. Como definir semelhança?
- Em geral, trata-se de um problema difícil:

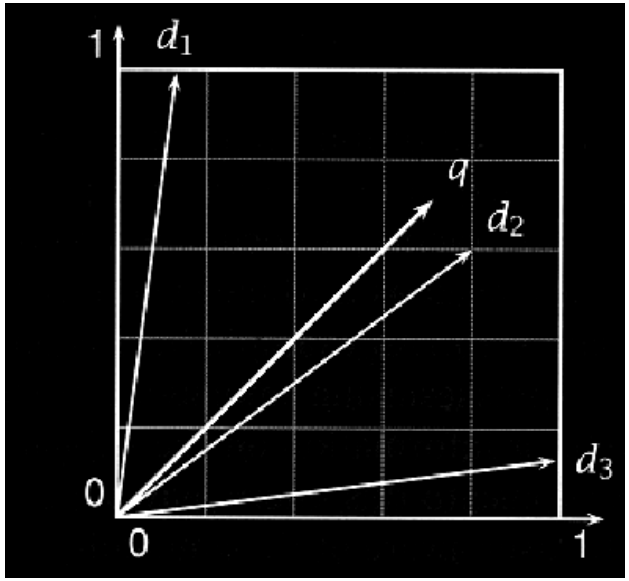
- Abordagens matemáticas são comumente adotadas.
- Vejamos a ideia básica para agrupar textos semelhantes…

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

- Selecionar palavras de interesse (eliminar preprosições, artigos etc.);
- Para cada palavra, computar sua frequencia por *texto;*





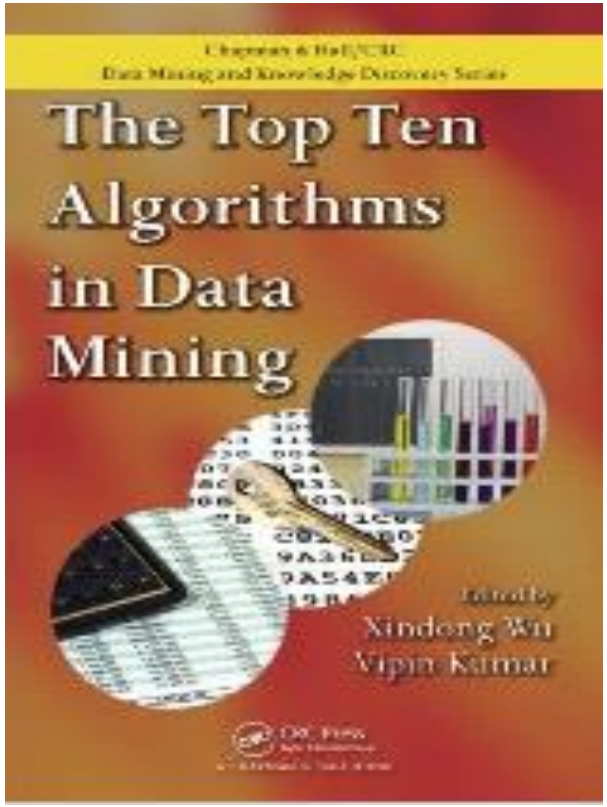


- Cada texto é representado por um vetor;
- Como encontrar *textos* semelhantes?

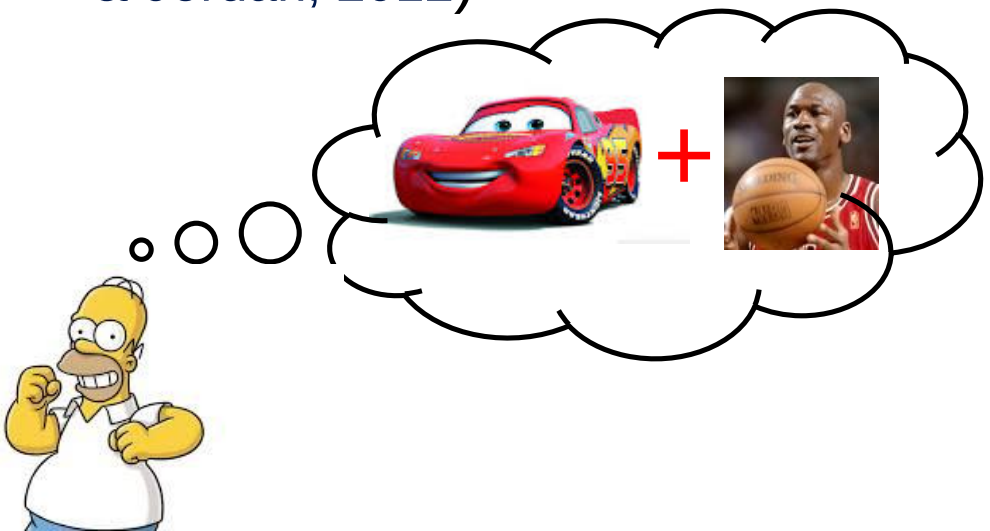
Técnicas e Aplicações I (agrupamento de dados)

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

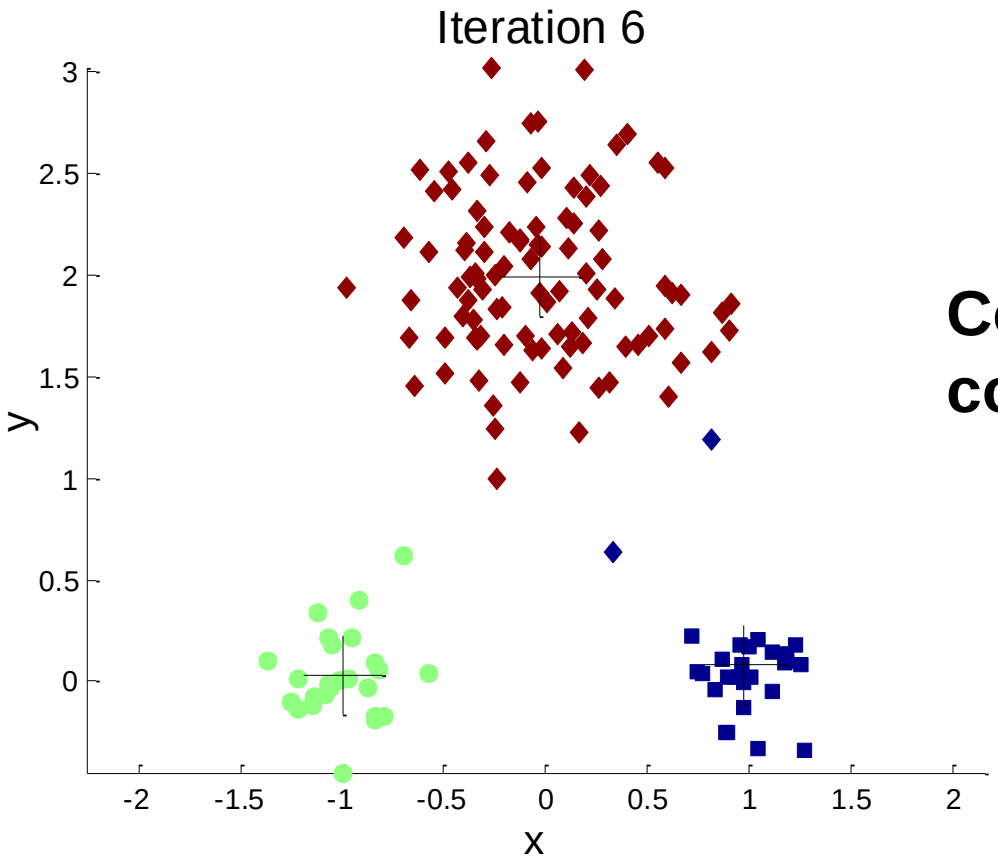
Ideia geral e intuitiva por meio de um exemplo ilustrativo:



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



Rodando K-means (K=3):



Complexidade computacional?

Tan et al., Introduction to Data Mining, 2004.

Custo Computacional

- **Complexidade (assintótica) de tempo**:

$$O(i \cdot K \cdot N \cdot n)$$

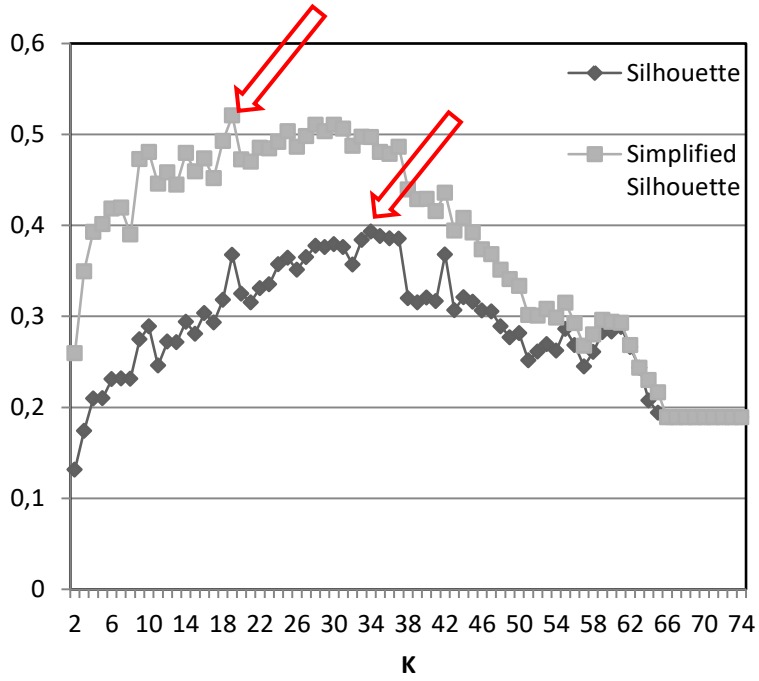
- O que isso significa?

O que dizer sobre a constante de tempo?
→ Computar Distância Euclidiana via aproximações sucessivas (Newton-Raphson) custa caro.

Se também tenho problema de espaço em memória...
→ Solução aproximada (*sampling*);
→ Paralelizar (mesmo computador) ou distribuir (e.g., *map-reduce*) o processamento.

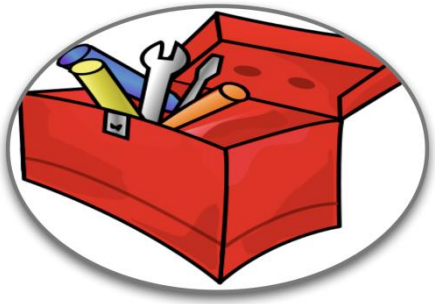
i = número de iterações, *K* = número de clusters, *N* = número de objetos, *n* = número de atributos

- Otimização convexa para cada *K*: converge para ótimos locais com diversas medidas de distância, 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 (e.g., computação forense):



- K* em torno de 20-30

Caixa de ferramentas (compacta)?



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

 **Don´t try this at home**

Técnicas e Aplicações II: 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

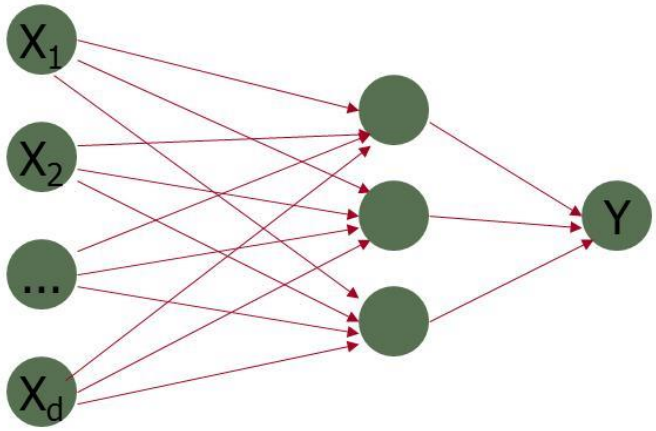
Técnicas e Aplicações II (classificação)

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

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

$Y \in \{\text{fraude, normal}\}$ (menos do que 1% de transações fraudulentas);
$X = \{X_1, X_2,\ldots,X_d\}$: variáveis descrevendo as transações;

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



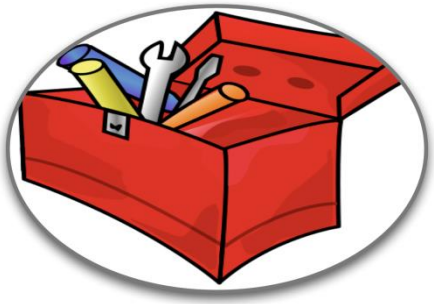
Passos principais:

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 $\{X_1, X_2,\ldots,X_d\}$.

Outras aplicações de classificação incluem:

- *Churn prediction* (cliente abandona serviço/produto): cartão de crédito, conta corrente etc.
- Finanças (cliente irá cumprir contrato de financiamento?, cliente irá pagar a fatura do cartão de crédito?);
- Alinhar atendente a cliente (telemarketing, SAC etc.);
- Recrutamento de profissionais;
- Abandono de posto de trabalho;
- Análise de sentimentos sobre produtos/serviços (redes sociais);
    - Requer dados de boa qualidade;
    - Diferentemente do trabalho típico de um “estatístico mais tradicional”, dados não foram especificamente coletados com o propósito de modelagem.

Caixa de ferramentas (compacta)?



- Regressão Logística
- Logistic LASSO (*Least Absolute Shrinkage and Selection Operator*)
- Naïve Bayes (*wrapper*)
- Árvores de Decisão e *Random Forests*
- *Classifier Ensembles* 
- Engenharia de atributos (feature selection) 
  - SVMs, redes neurais (*deep learning*) etc. 

Momento de reflexão:

- Cuidado com generalizações baseadas em pequenas amostras;
- Se *tentei a técnica X* uma vez no passado e não funcionou* então *nunca irá funcionar*. Fica frio, Wolpert & Macready vão te salvar!

- Vendo a luz: não há um algoritmo universal, mas para cada problema há um campeão!
- Teste sempre diferentes algoritmos, começando pelos mais simples e <u>otimizando seus parâmetros</u>.

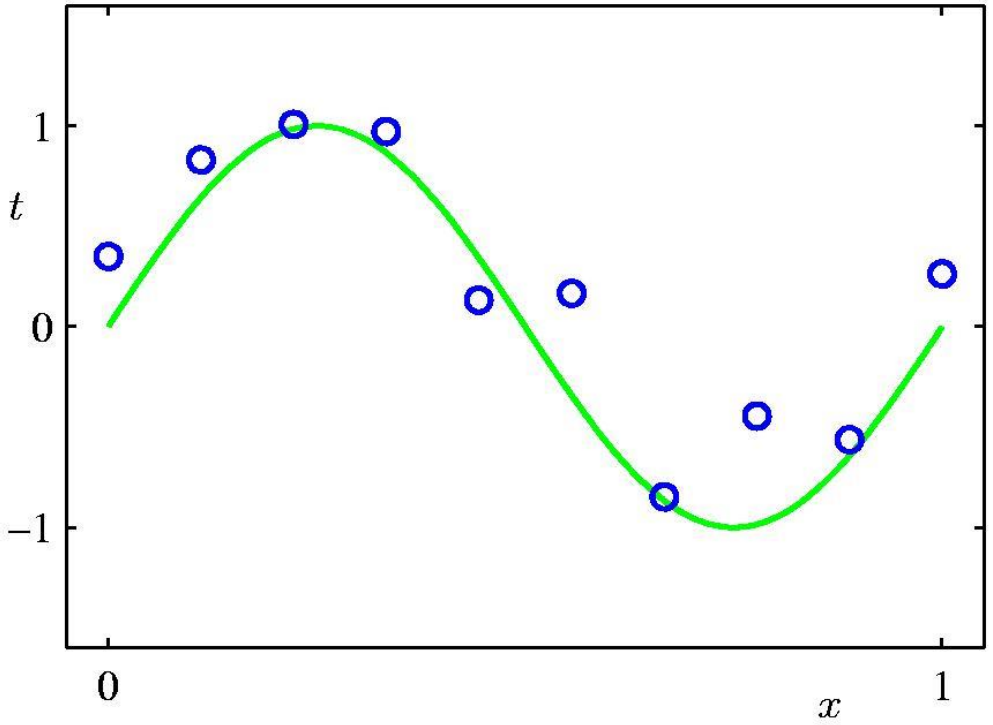
➢ Happy boss!

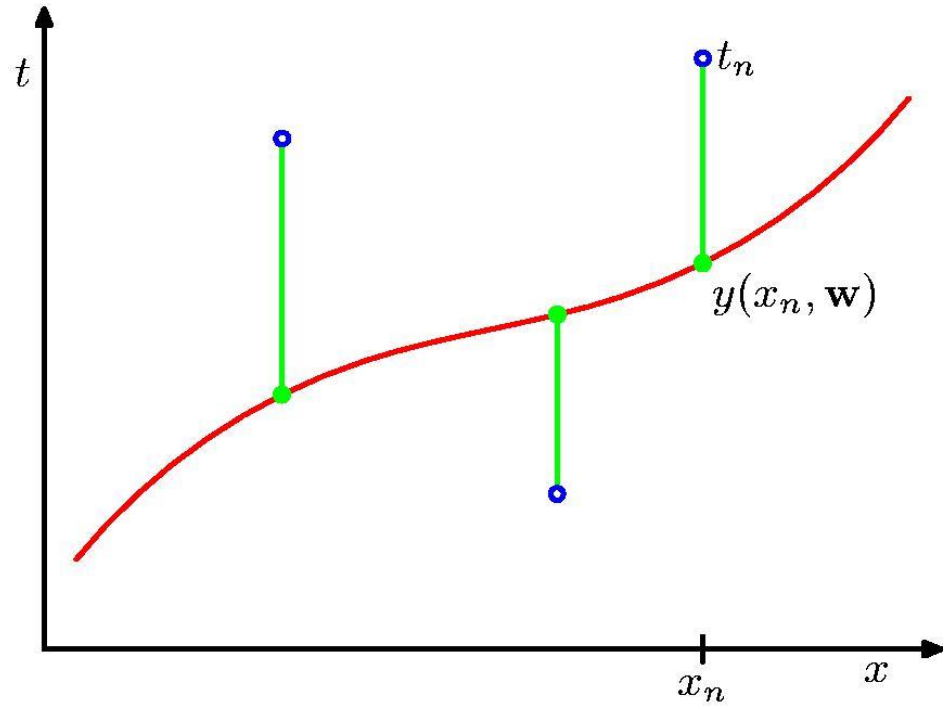
"*The reasonable man adapts himself to the world; the unreasonable one persists in trying to adapt the world to himself. Therefore all progress depends on the unreasonable man*." (G.B. Shaw, 1903)

Técnicas e Aplicações III: Regressão

- Onipresente na Estatística; pouco interesse na Computação
- Enorme sucesso na prática (renda, crédito, séries temporais etc.)
- Nosso foco será em sistemas de recomendação
- No contexto de big data: regressão multivariada, mas antes façamos uma breve digressão $Y=f(X_1)$

Aprender um polinômio que se ajusta bem aos dados:





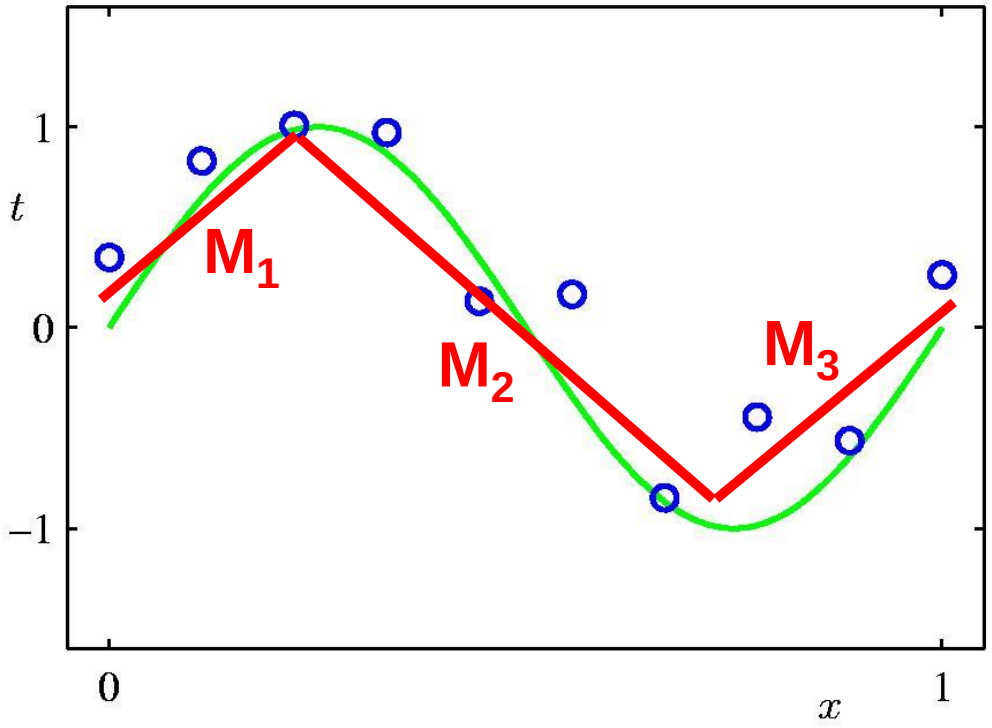
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

- Ideia simples e bem estudada: induzir modelos locais (mais simples) que aproximem suficientemente bem o polinômio.

1) $y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M=1} w_j x^j$

2) Usar vários modelos simples:



Desafio: como aprender de maneira automática e rápida?

Rápida = *O*(.) + constante

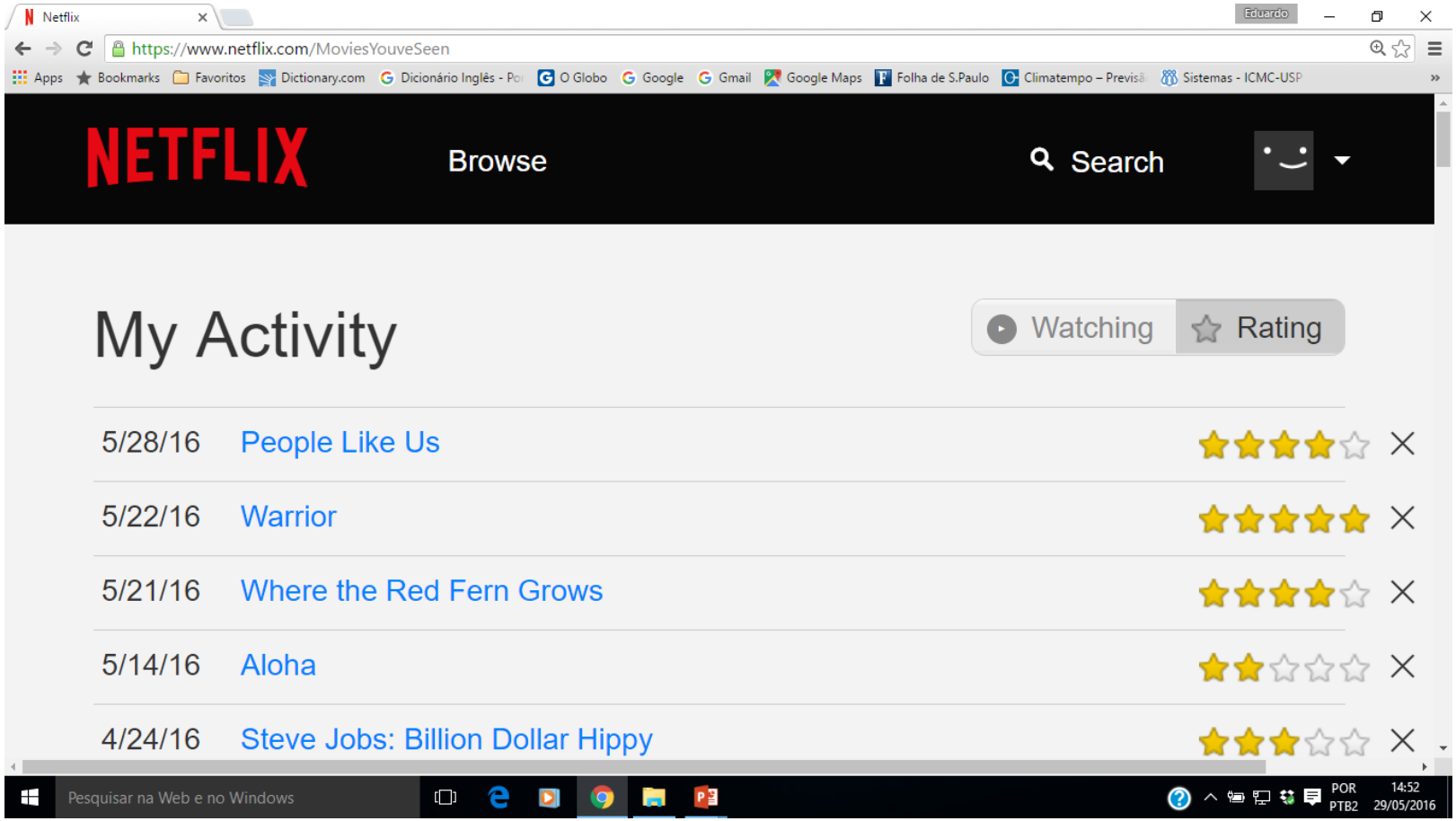
3) Alternativa principiada: SCOAL

➢ Agrupamento e Regressão Simultâneos

- Em problemas difíceis de regressão, frequentemente se segmenta a base de dados em grupos homogêneos e constrói-se um modelo por grupo;
- 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;
- Vejamos um exemplo ilustrativo sobre aprendizado automático de modelos locais a partir dos grupos:

Técnicas e Aplicações III (regressão)

Consideremos um cenário de recomendação:



Seja Y= f(gênero, duração, atores, diretor, ano, prêmios, ...) a avaliação (1,2,3,4,5) feita por um usuário para determinado filme.

Técnicas e Aplicações III (regressão)

Nota do filme

Dados de filmes

Dados de usuários

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

Algumas alternativas de modelagem:

- Construir um modelo para cada usuário

Y = f(atributos do filme)

- Construir um modelo para todos os usuários

Y = f(atributos do filme e do usuário)

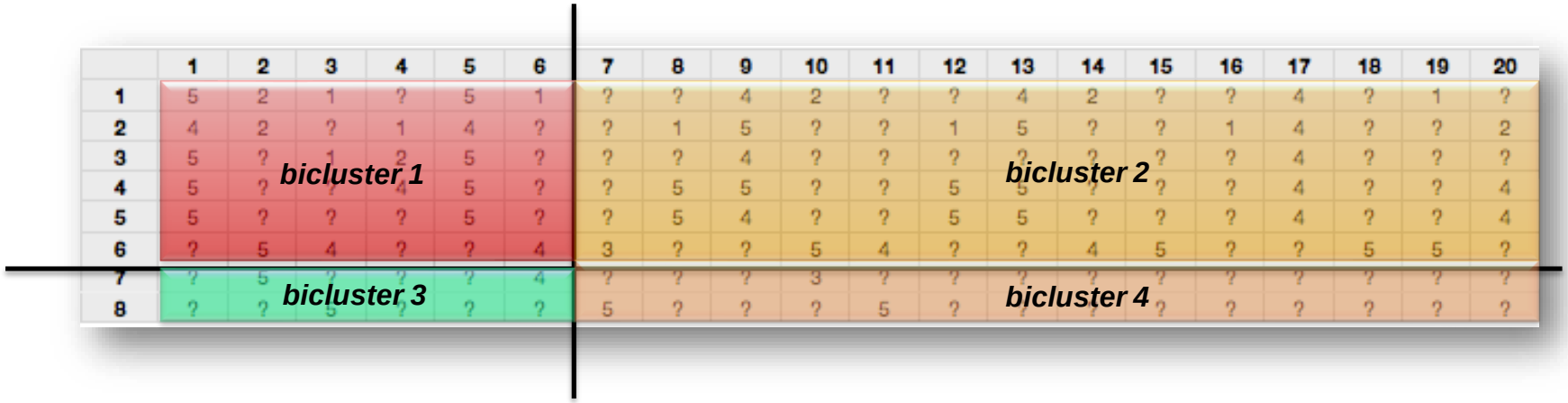
- Construir modelos para grupos de usuários e filmes

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
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	?	?	?	?	?	?	?	4	?	?	?
4	5	?	?	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	?	5	?	?	?	4	?	?	?	3	?	?	?	?	?	?	?	?	?	?
8	?	?	5	?	?	?	5	?	?	?	5	?	?	?	?	?	?	?	?	?

- Treinar 4 modelos de regressão (um por grupo)

Técnicas e Aplicações III (regressão)



- Um modelo por bicluster:

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

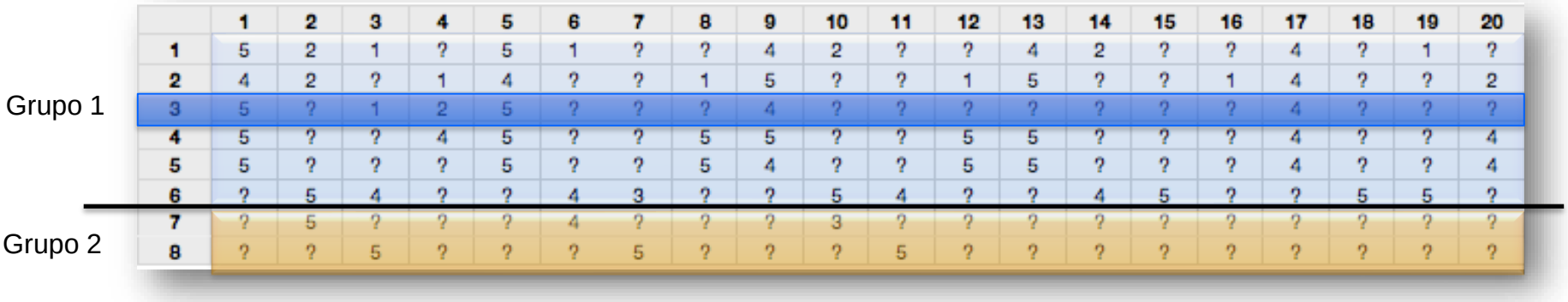
$$\beta^T = [\beta^0, \beta_i^T, \beta_j^T]$$

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

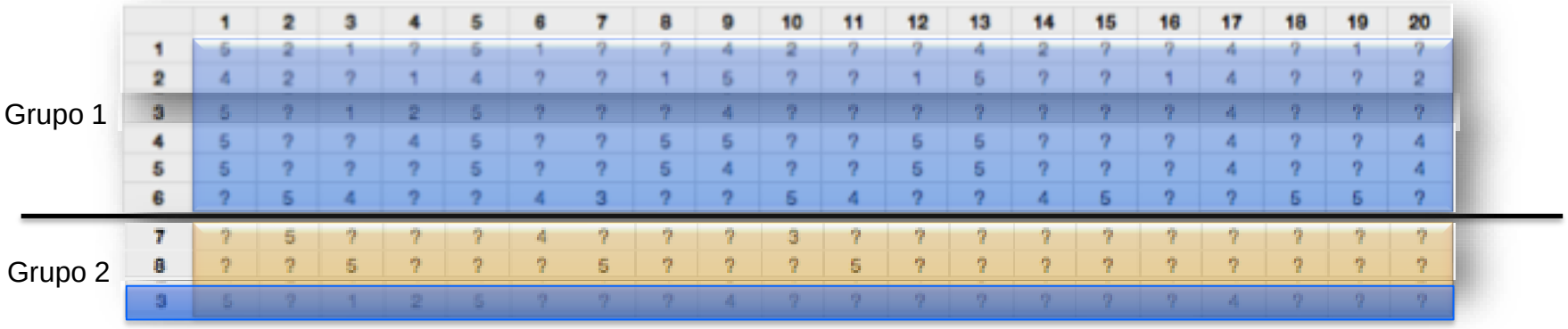
$$MSE = \sum_{ij} (z_{ij} - \hat{z}_{ij})^2$$

Técnicas e Aplicações III (regressão)

- Atualizar grupos de linhas, movendo-as para grupos que minimizam o MSE:

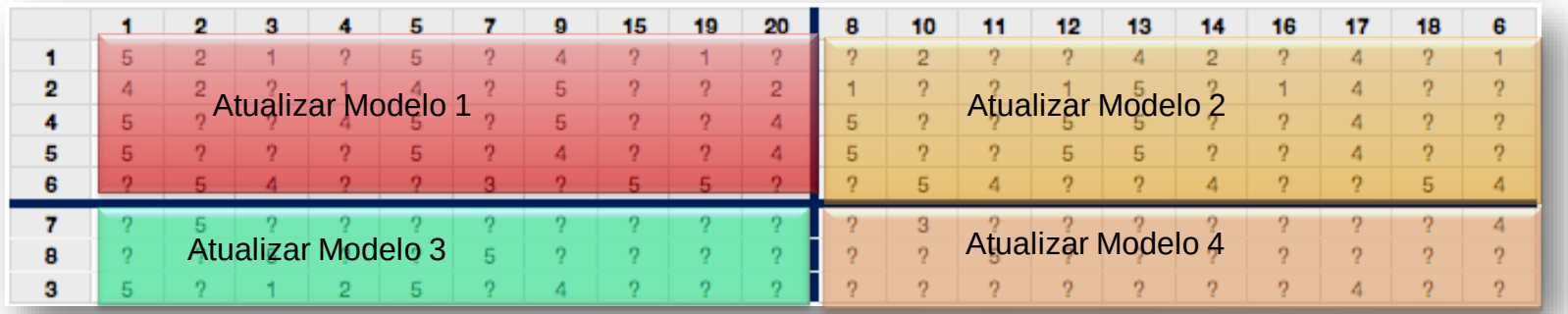


- Linha 3 obtém menor MSE ao ser predita pelos modelos do grupo 2:

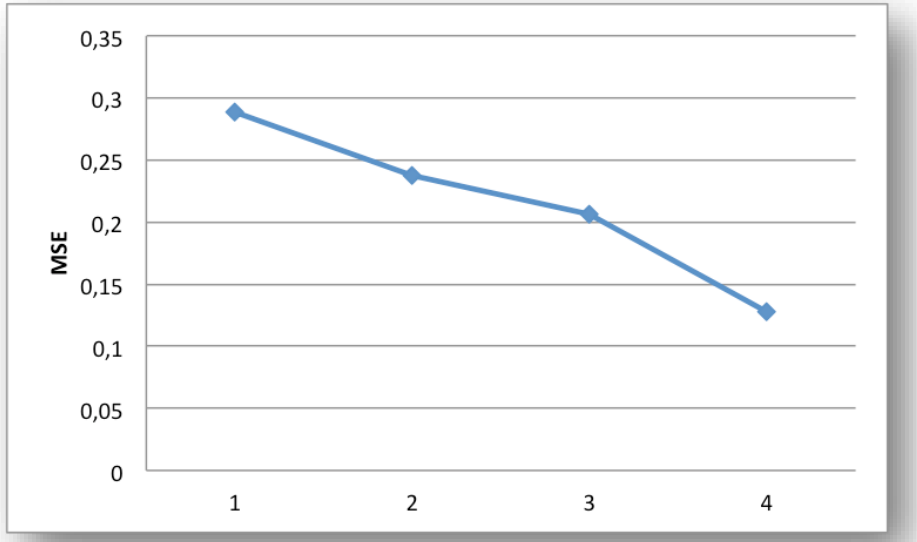


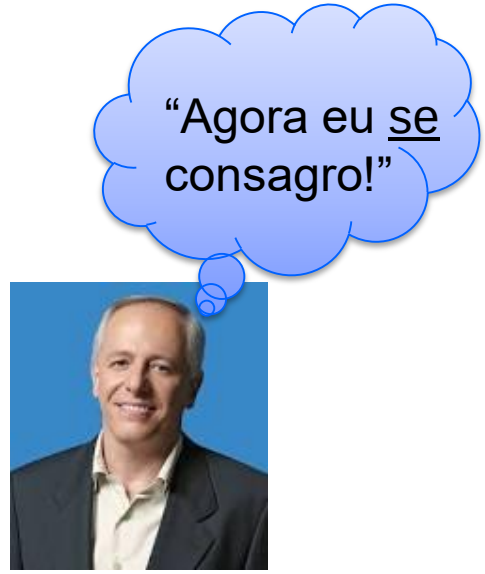
- Repetir o processo para cada linha/coluna...

Após o grande laço, estimar os modelos:

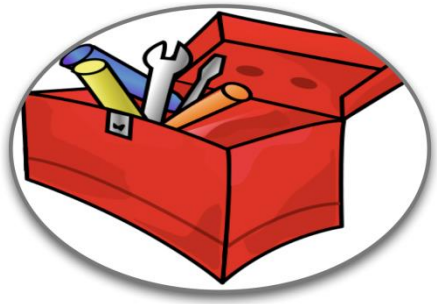


- Mover linhas/colunas para os grupos que minimizam o erro;
- MSE global é (garantidamente) minimizado nas iterações:





Caixa de ferramentas (compacta)?



- LASSO (Regressão linear é caso particular)
- Árvores de regressão / *random forests*
- Modelos lineares generalizados
- k-Nearest Neighbors 
- Redes neurais, SVMs etc. 

Tendências e Desafios

Tendências e desafios

- Combinar diferentes algoritmos de otimização;
- Diminuir o número de parâmetros críticos definidos pelo usuário via ajuste automático (a partir dos dados);
- Crescente número de novas aplicações;
- Questões éticas do uso de modelos automáticos.

”Essentially, all models are wrong, but some are useful.”
(George E. P. Box, Professor Emeritus, University of Wisconsin)

“We don´t have better algorithms, we just have more data.”
“More data beats clever algorithms, but better data beats more data.”
(Peter Norvig, Director of Research, Google)

Próximas aulas

- Classificação
- Regressão
- *Clustering*
- Preparação de Dados

➢ Curso do Andrew Ng (Coursera)

Pra virar "profissa":

Cenários: aprendizado semi-supervisionado, aprendizado ativo, fluxos de dados, mineração de textos, redes complexas etc.

Técnicas: modelos gráficos probabilísticos, deep learning, algoritmos evolutivos etc.