
QCon
RIO DE JANEIRO


Algoritmos de
Recomendação

Felipe Cruz

@felipejcruz

github.com/felipecruz

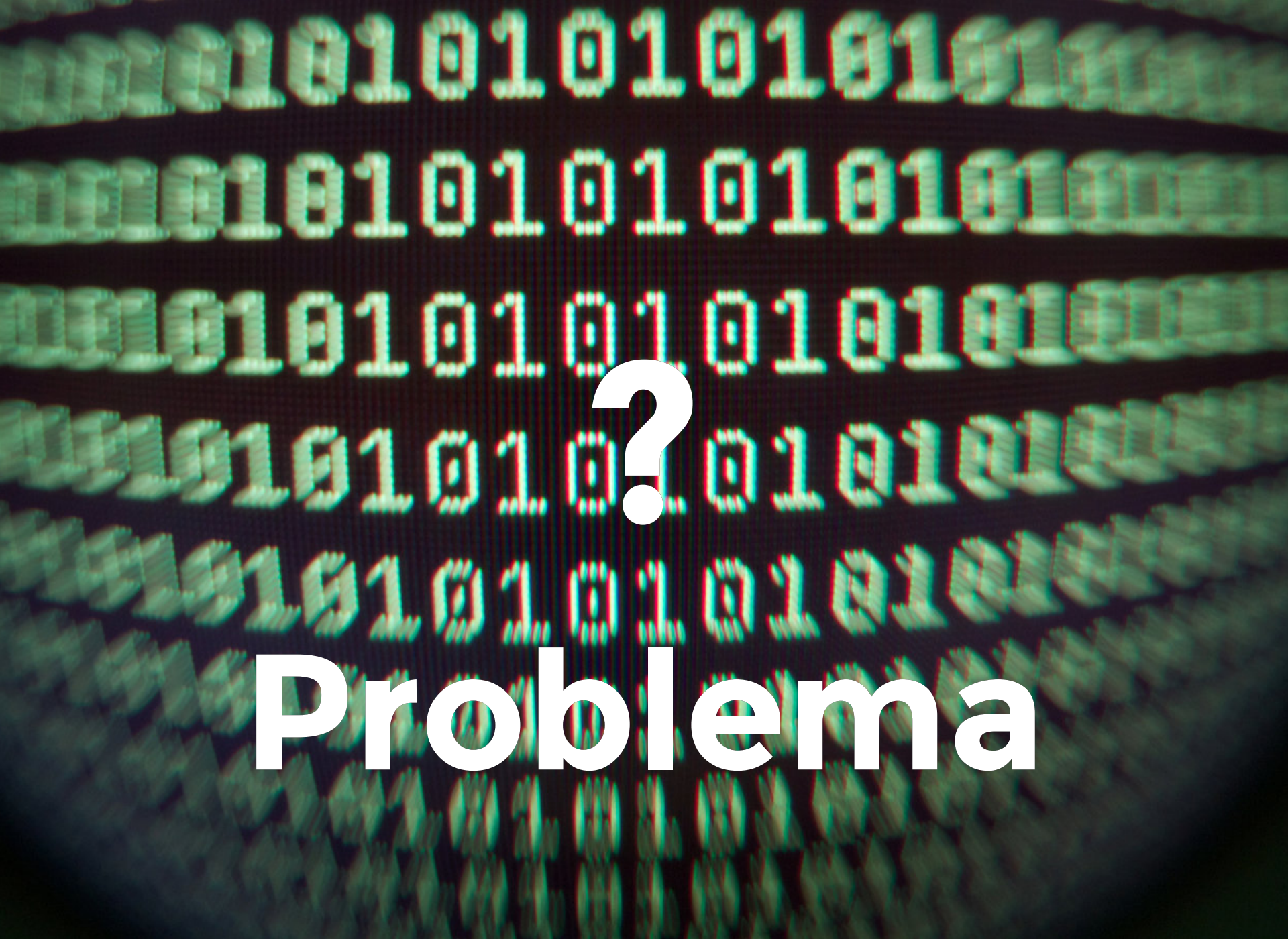
Problema

Algoritmos

CF & RBMs

Motivação?

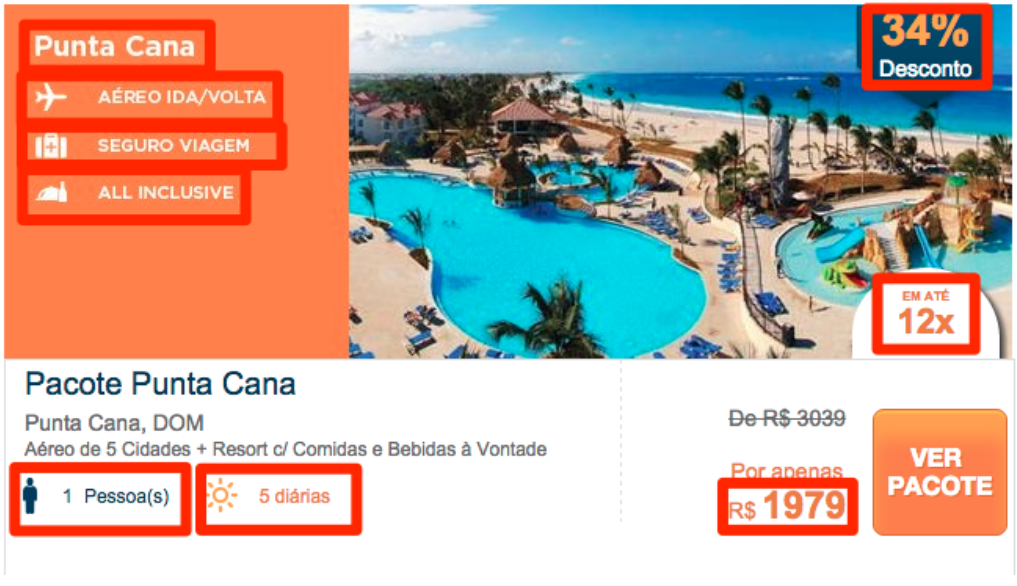
relevância
eficiência
personalização

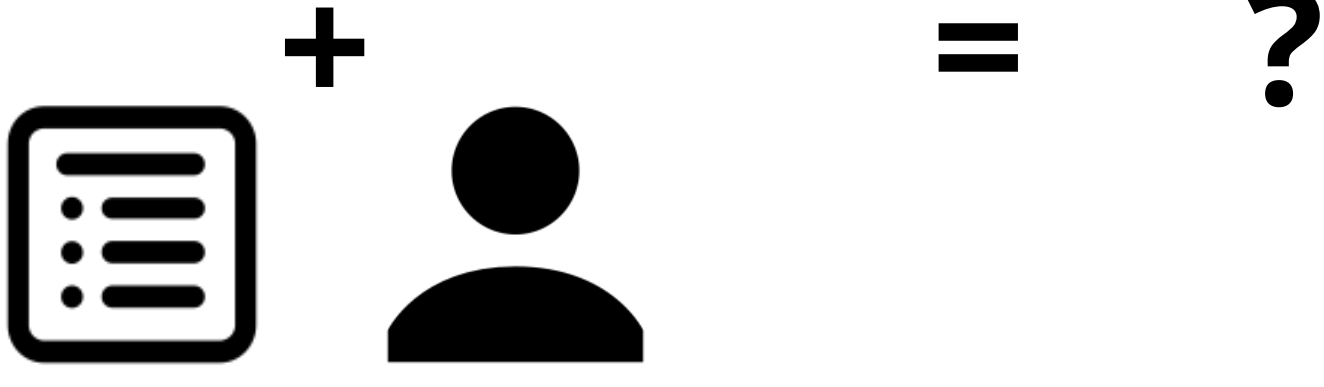

?
Problema

items

users

	w	x	y	z
a	4	3	?	?
b	?	4	?	1
c	?	?	3	4
d	2	4	?	?


Punta Cana
AÉREO IDA/VOLTA
SEGURO VIAGEM
ALL INCLUSIVE
34%
Desconto
EM ATÉ
12x
Pacote Punta Cana
Punta Cana, DOM
Aéreo de 5 Cidades + Resort c/ Comidas e Bebidas à Vontade
1 Pessoa(s)
5 diárias
De R$ 3039
Por apenas
R$ 1979
VER PACOTE




Quantos
usuários?


Quantos
produtos?

Recomendação não serve apenas para "Big Data", só que..

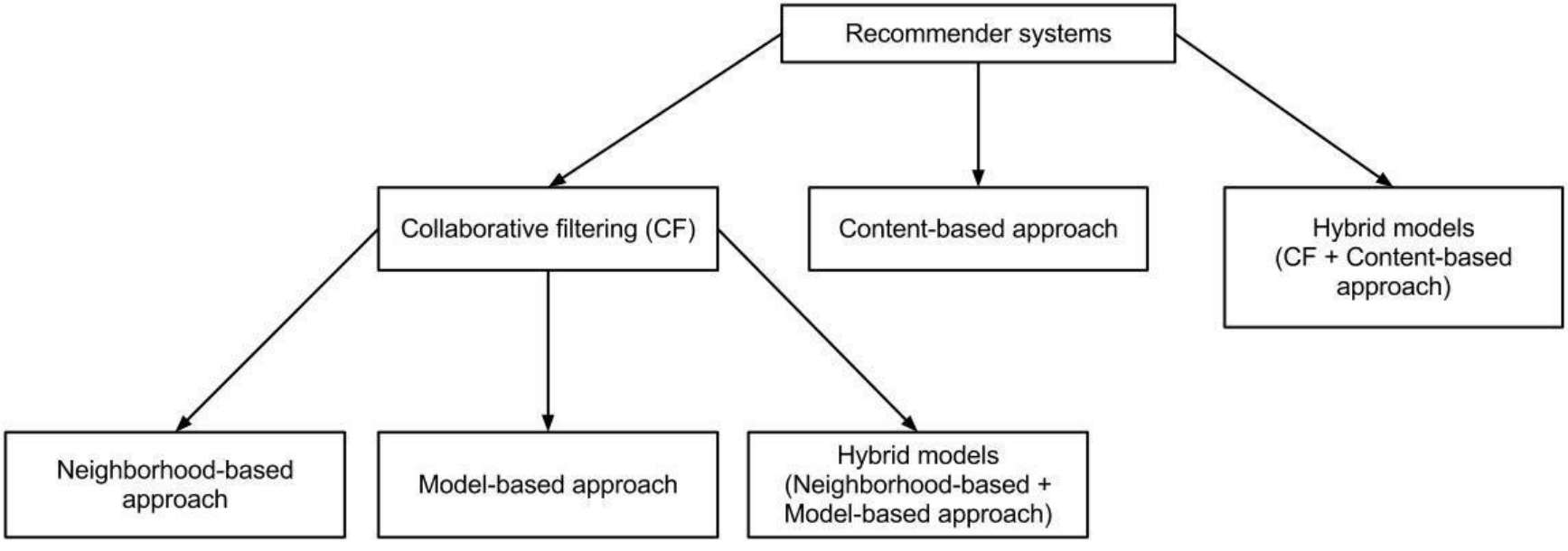
quanto maior for a matriz, mais recursos são necessários

Problema

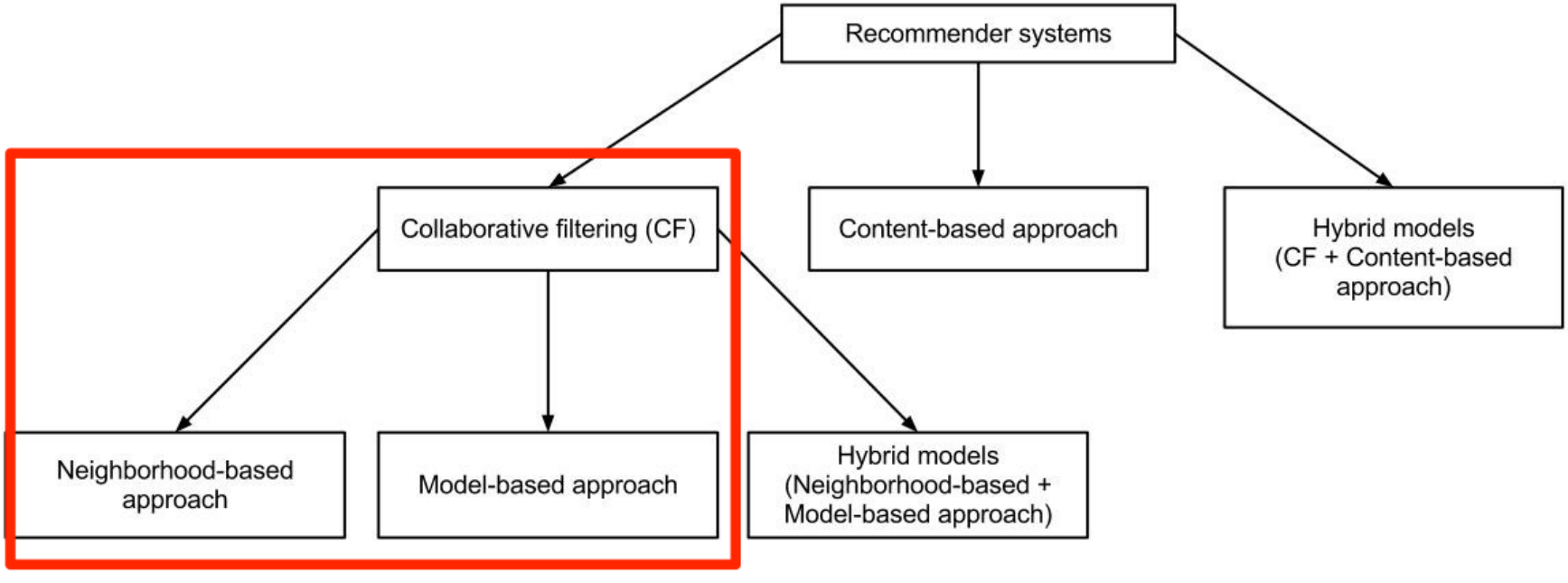
Algoritmos

CF & RBMs

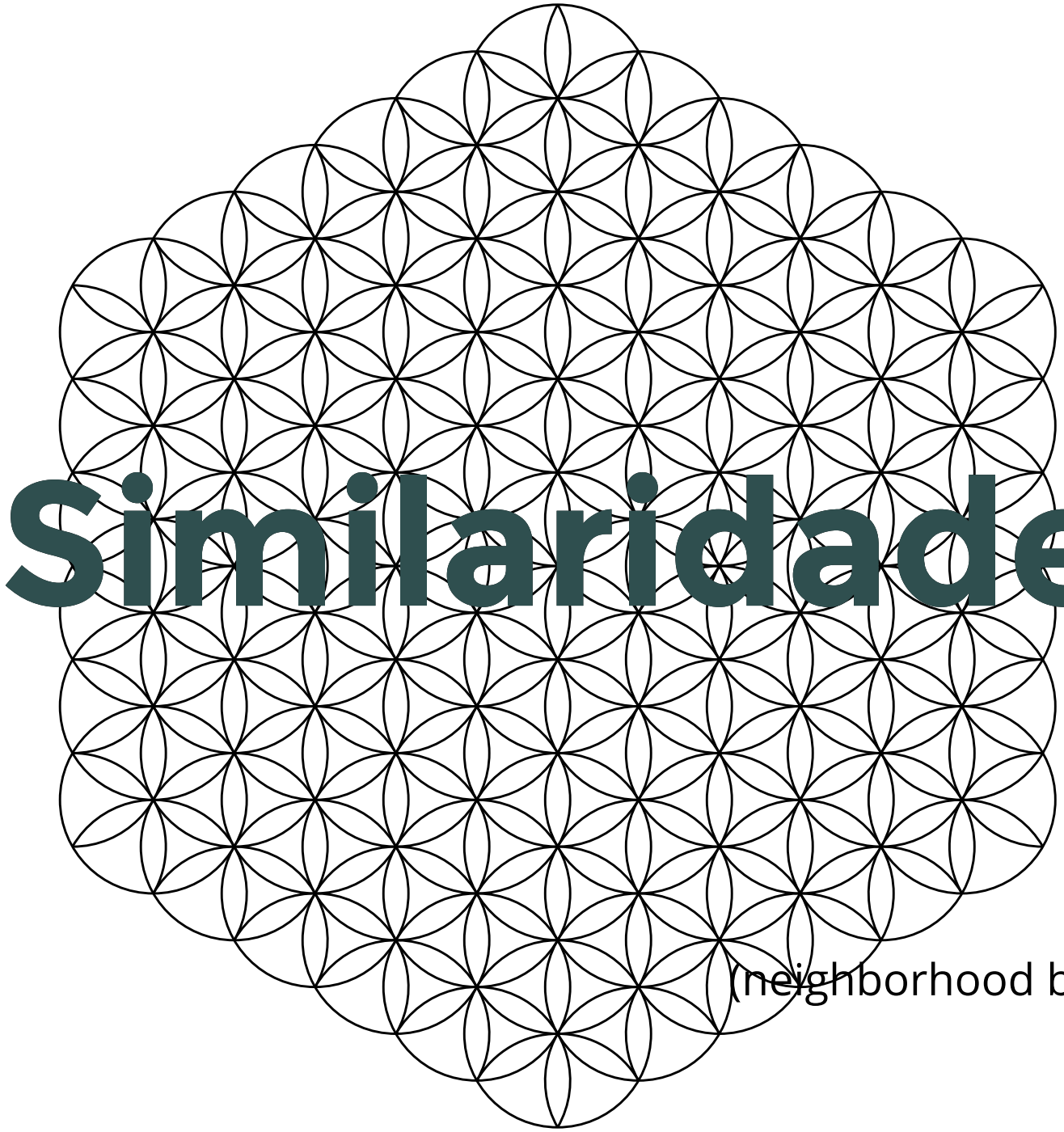
Sistemas de Recomendação



Sistemas de Recomendação



Filtragem Colaborativa



Similaridade

(neighborhood based)

Procedimento de treino

Para cada usuário

Descobre usuários que compraram produtos em comum

Procedimento de treino

Procedimento de treino

agora podemos calcular a similaridade

f é a função de similaridade

$$f\left(\text{  } \right)$$

$$f(a,u,I) = \frac{\sum_{i \in I}(r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})}{\sqrt{\sum_{i \in I}(r_{a,i} - \overline{r_a})^2 \sum_{i \in I}(r_{u,i} - \overline{r_u})^2}}$$

Pearson correlation

$$\sum_{i \in I} (r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})$$

$$I = \{ \text{  }, \text{  } \}$$

nota dada pelo usuário a ao item i

menos média das notas do usuário a

$$\sum_{i \in I}(r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})$$

$$[(4.5 - 4)(4 - 3.5)] + [(3.5 - 4)(3 - 3.5)]$$

$$[0.5 \times 0.5] + [-0.5 \times -0.5] = 0.5$$

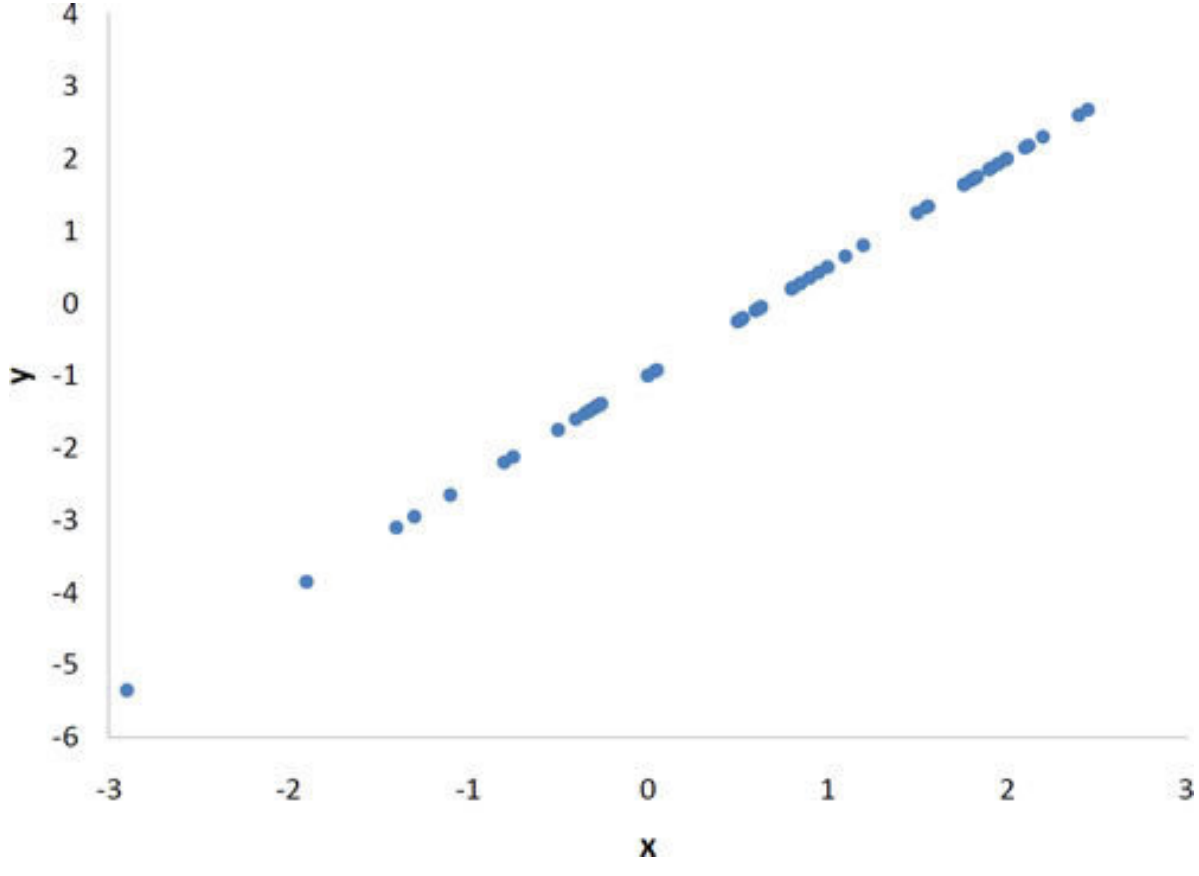
$$\sqrt{\sum_{i \in I}(r_{a,i} - \overline{r_a})^2 \sum_{i \in I}(r_{u,i} - \overline{r_u})^2}$$

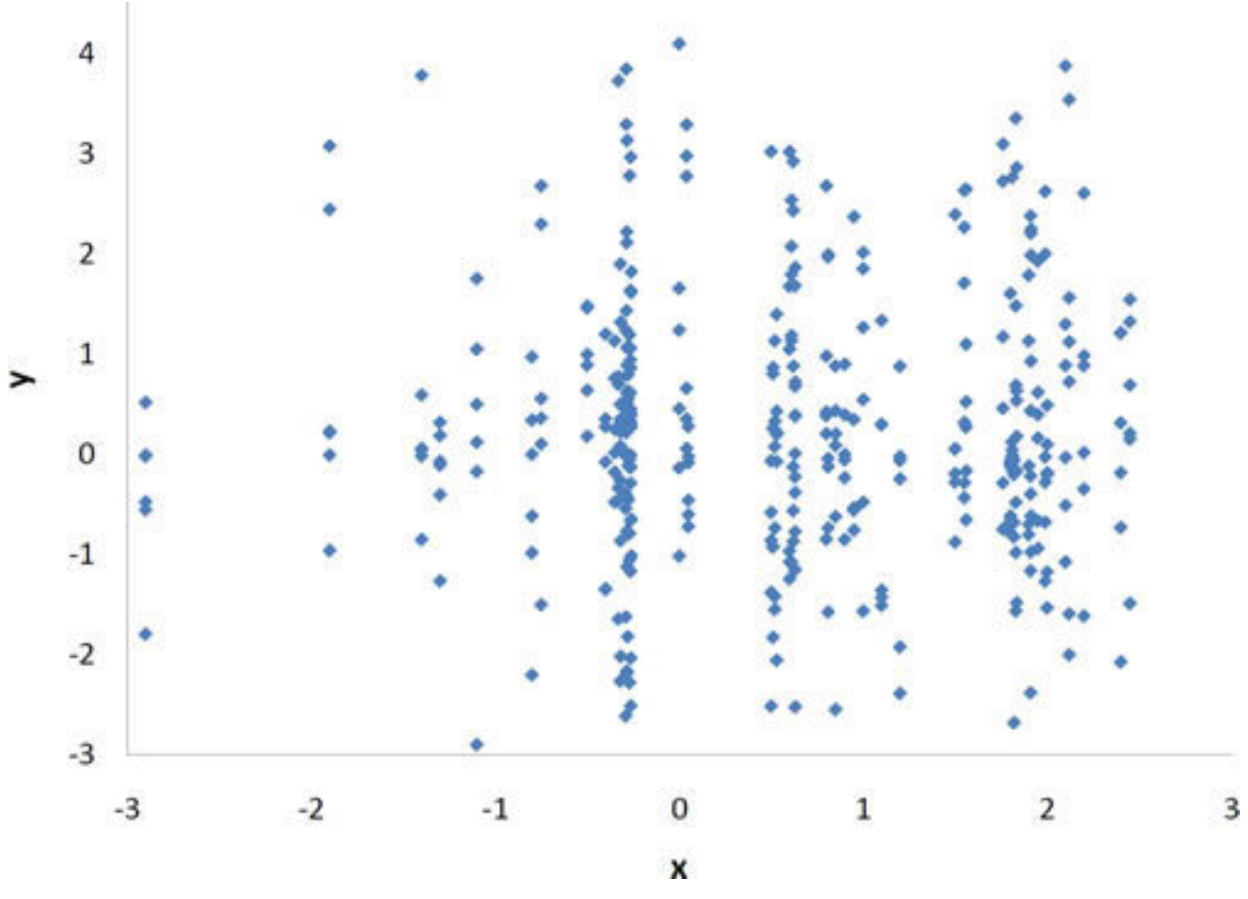
$$[(4.5 - 4)^2 + (3.5 - 4)^2] \times [(4 - 3.5)^2 + (3 - 3.5)^2]$$



$$[(0.5)^2 + (-0.5)^2] \times [(0.5)^2 + (-0.5)^2] = \sqrt{0.5^2} = 0.5$$

$$\frac{0.5}{0.5} = 1$$


4
3
2
1
0
-1
-2
-3
-4
-5
-6
y
-3
-2
-1
0
1
2
3
x


4
3
2
1
0
-1
-2
-3
y
-3
-2
-1
0
1
2
3
x

Predição?

Predição

$$p_{a,i} = \overline{r_a} + \frac{\sum_{u \in K} (r_{u,i} - \overline{r_u}) \times w_{a,u}}{\sum_{u \in K} w_{a,u}}$$

Predição

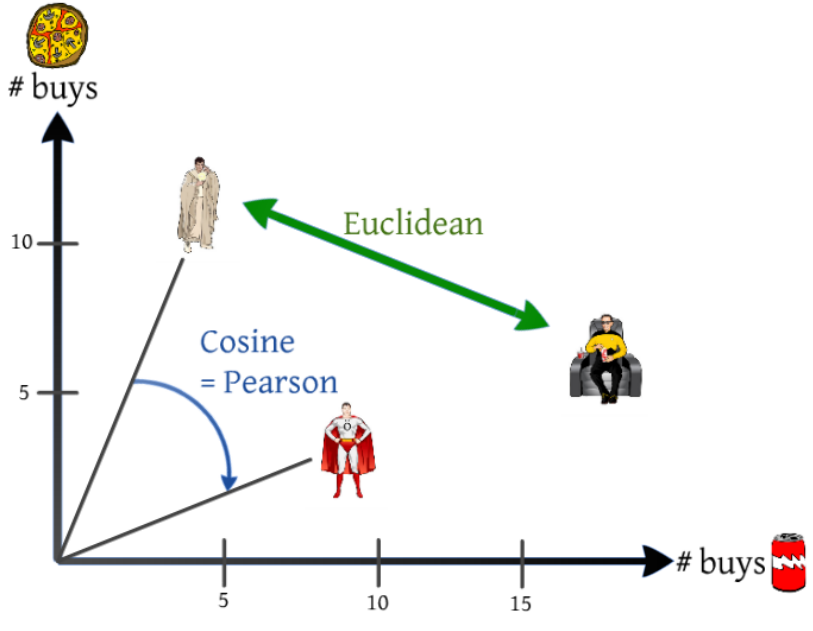
$$p_{a,i} = 4 + \frac{5 - 3.5 \times 1}{1} = 5.5$$

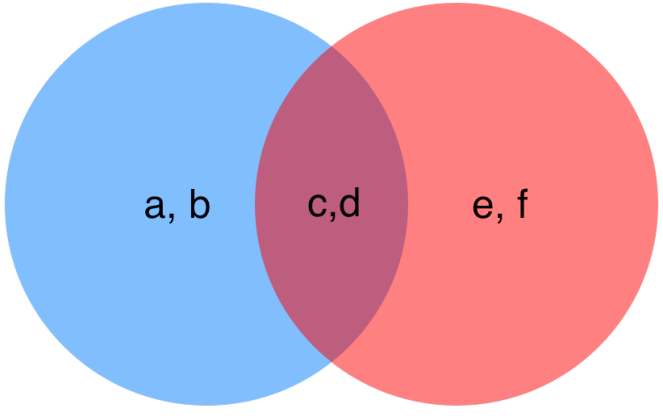

4.5
4
3.5
3
5
5.5

Template method - treino

- for u in usuarios:
  - similares = mesmos_produtos(R, u)
  - for s in similares:
    - S[(u,s)] = similaridade(u, s)

Outras medidas de similaridade





Jaccard

pausa...

Fator latente

(model based)

Capturar informação implícita

Capturar **informação** implícita

#comofaz

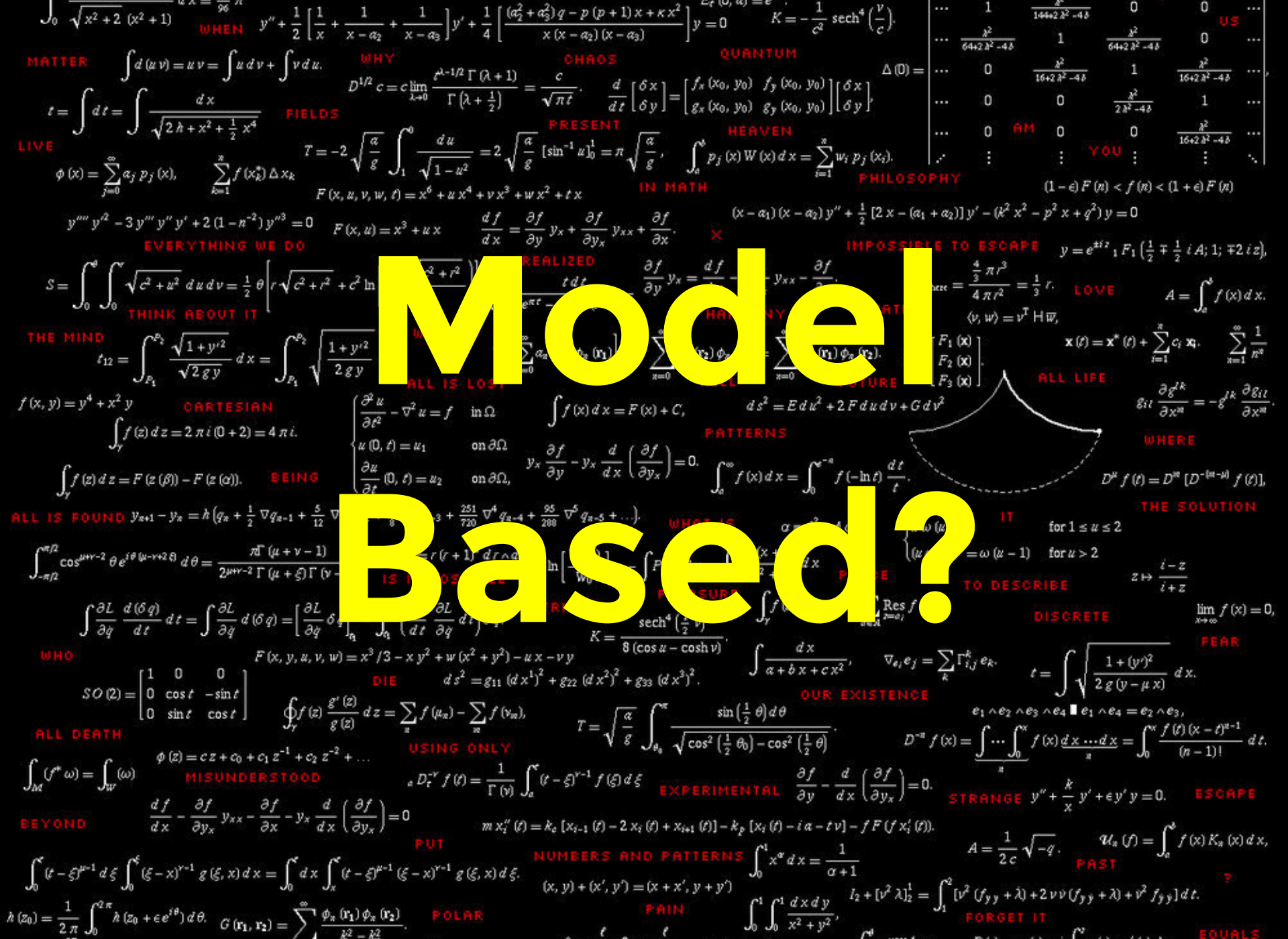
items

users

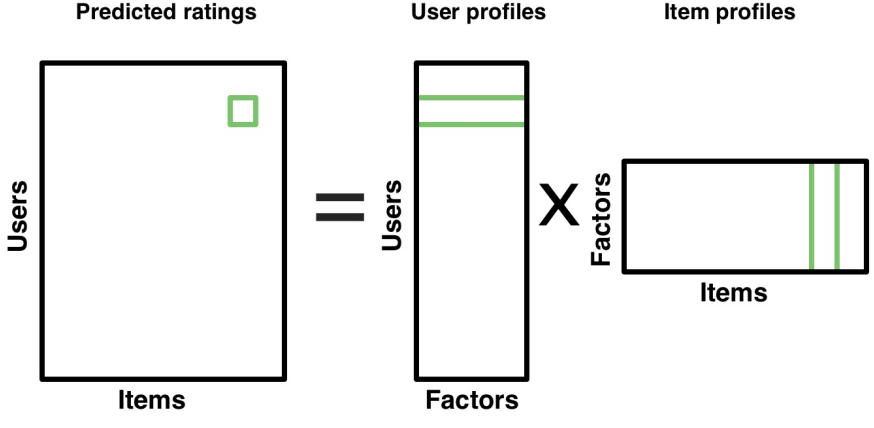
	w	x	y	z
a	4	3	?	?
b	?	4	?	1
c	?	?	3	4
d	2	4	?	?

Vamos dar um nome aquela matriz:

$$R_{ij} \in \mathbb{N}\,(u \times p)$$

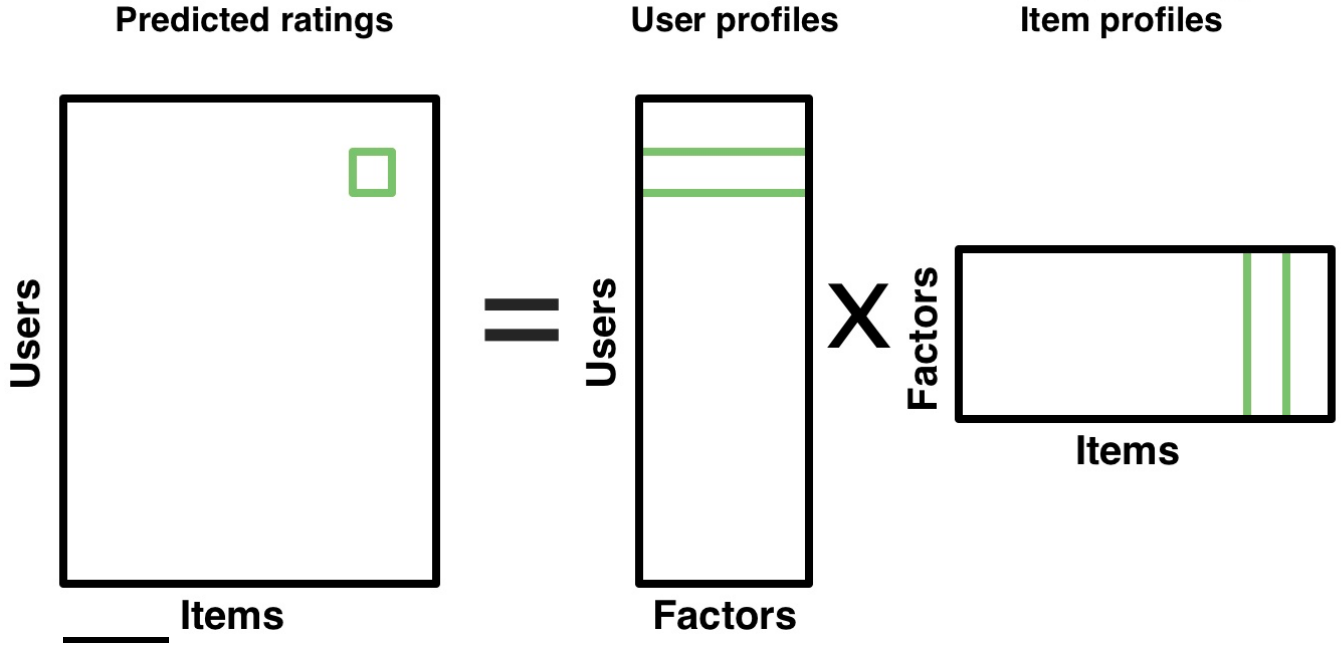

Model
Based?

Modelo



$$\langle u_i, p_j \rangle \approx R_{ij}$$

Fatorar matrizes



$$\overline{R}_{(u \times p)} = U_{(u \times k)} P_{(k \times p)}$$

Formalização do Problema

$$\min_{U,P} \sum_{i,j \in L} \left(r_{i,j} - u_i^T p_j\right)^2 \quad + \lambda \left(\sum_i n_{u_i} ||u_i||^2 + \sum_j n_{v_j} ||v_j||^2 \right)$$

$$\langle u_i, p_j \rangle$$

Formalização do Problema

$$\min_{U,P} \sum_{i,j \in L} (r_{i,j} - u_i^T p_j)^2 + \lambda \left(\sum_i n_{u_i} ||u_i||^2 + \sum_j n_{v_j} ||v_j||^2 \right)$$

Regularização contra *overfit*

Minimizar o erro

$$\min_{U,P} \sum_{i,j \in L} (r_{i,j} - u_i^T p_j)^2$$

$$E = \begin{pmatrix} 2 & 4 \\ 3 & 6 \end{pmatrix} - \begin{pmatrix} \langle u_1, p_1 \rangle & \langle u_1, p_2 \rangle \\ \langle u_2, p_1 \rangle & \langle u_2, p_2 \rangle \end{pmatrix}$$

R_{ij} $\overline{R}_{ij}$

Algoritmo ALS

1. Inicializa U com valores randômicos
2. Fixa U, e para cada linha de P resolvemos um problema de regressão linear para achar valores ótimos para P
3. Fixa P, e para cada linha de P resolvemos um problema de regressão linear para achar valores ótimos para U


ALS
Ajusta traseira
Ajusta frente
Ajusta traseira
Ajusta frente

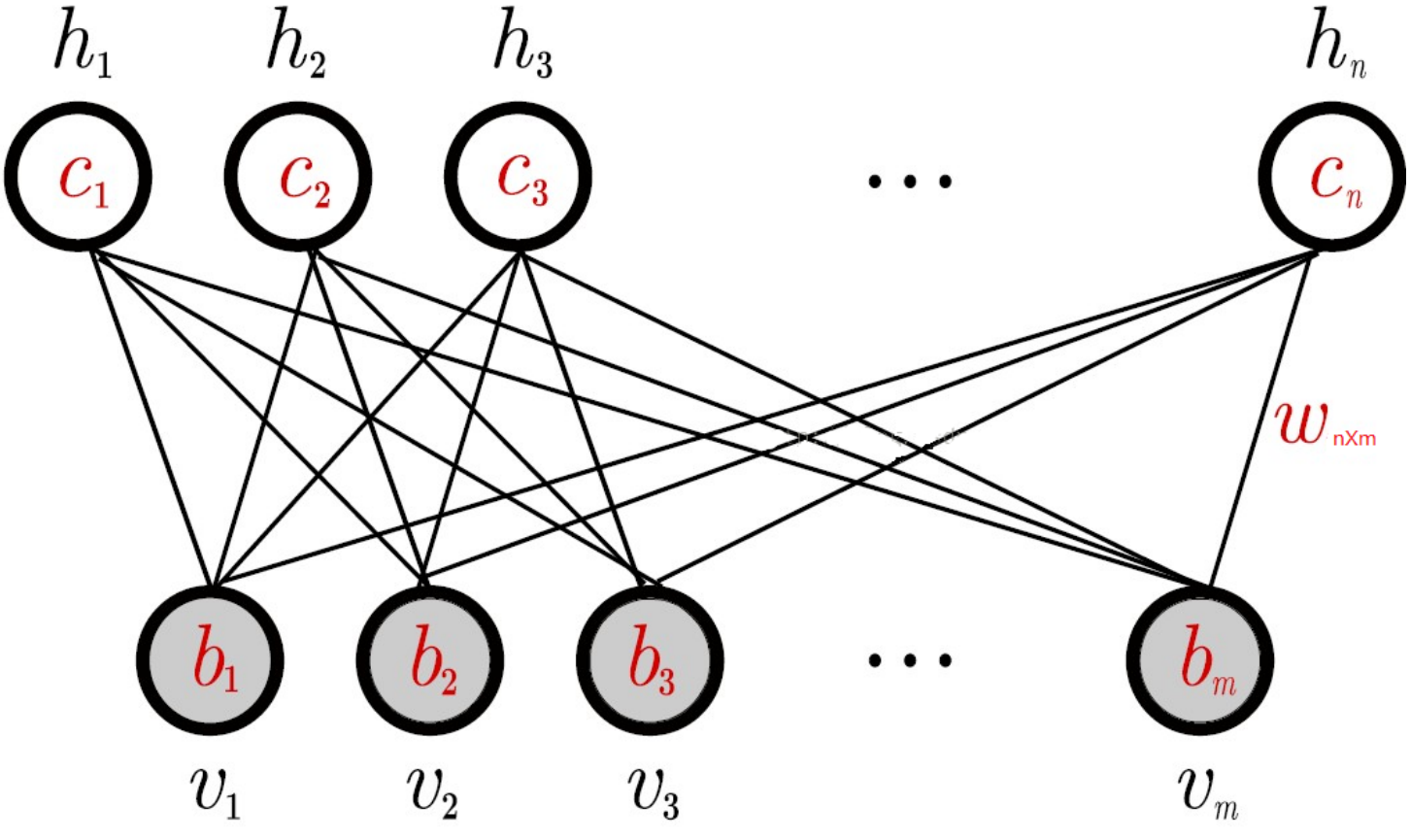
Problema

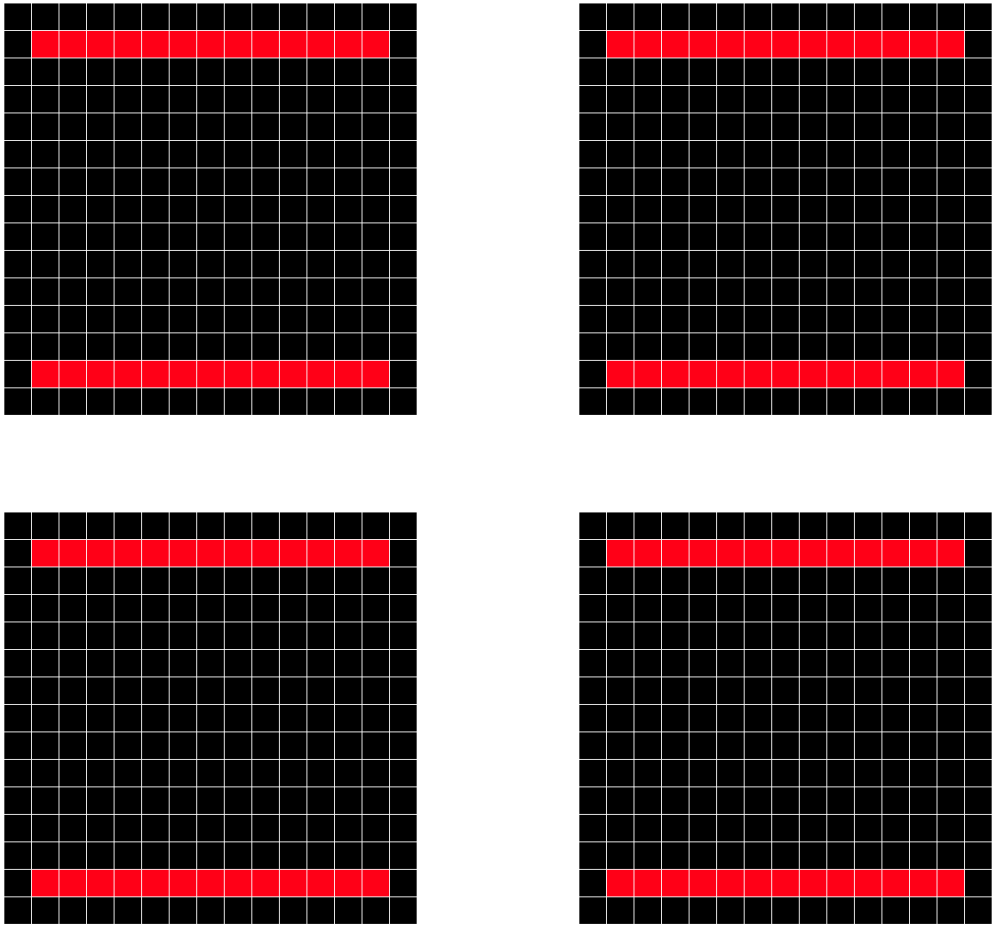
Algoritmos

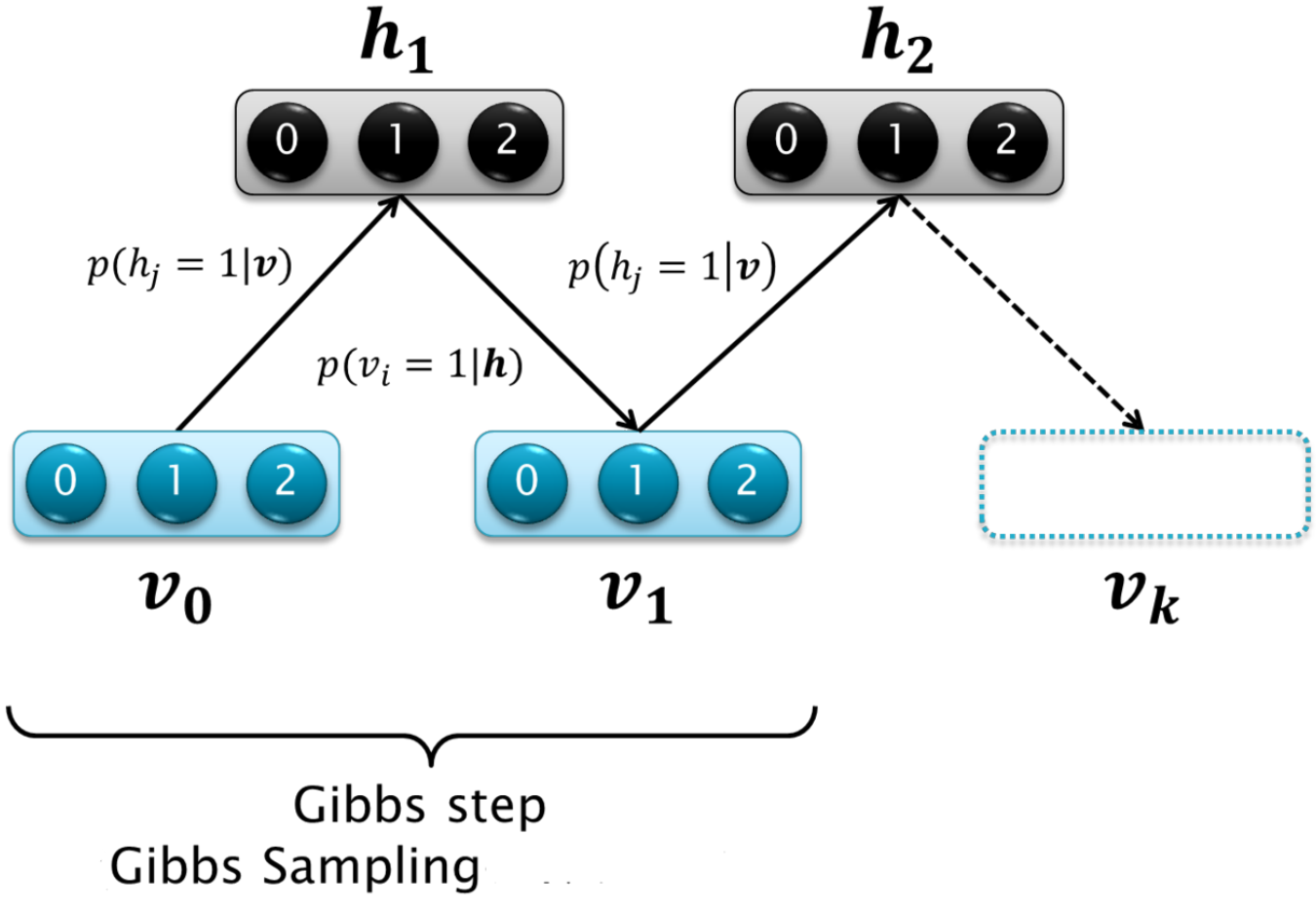
CF & RBMs

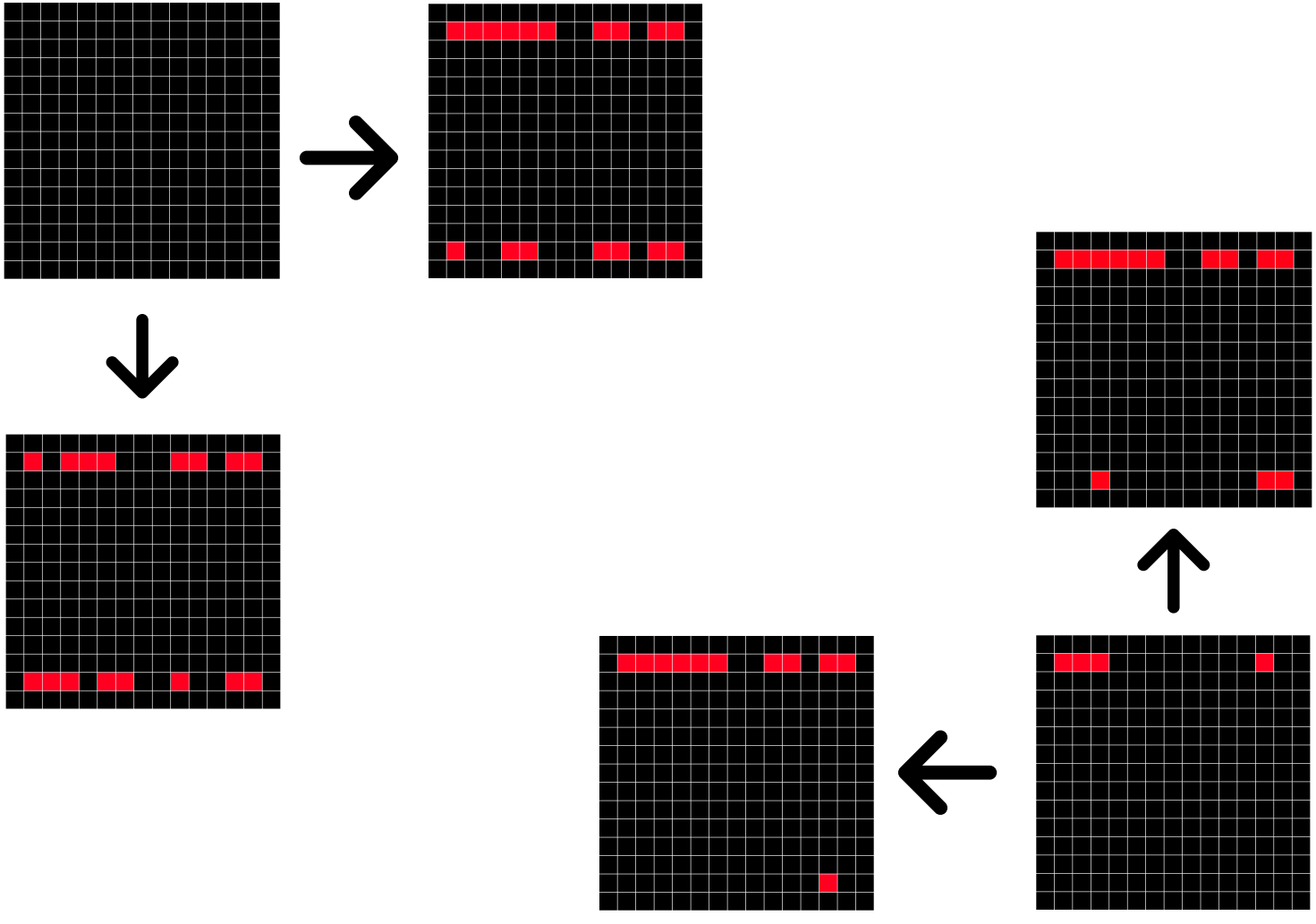
RBM

- Modelo generativo
- Reconhece e completa padrões
- Algoritmo de aprendizado eficiente


h_1
h_2
h_3
h_n
c_1
c_2
c_3
c_n
w_{nXm}
b_1
b_2
b_3
b_m
v_1
v_2
v_3
v_m




h_1
h_2
0 1 2
0 1 2
$p(h_j = 1|v)$
$p(h_j = 1|v)$
$p(v_i = 1|h)$
0 1 2
0 1 2
v_0
v_1
v_k
Gibbs step
Gibbs Sampling



RBMs & CF

Modelo

- Cada destino é uma unidade visível
  - 1 se comprou
  - 0 se não comprou

Também é possível usar atributos do usuário no modelo. Ex: tem filhos? Só compra pacotes de luxo?

Predição - input

profile = [0, 0, 0, 1, 0, 0, 0, 1]

(buzios e ilha grande)

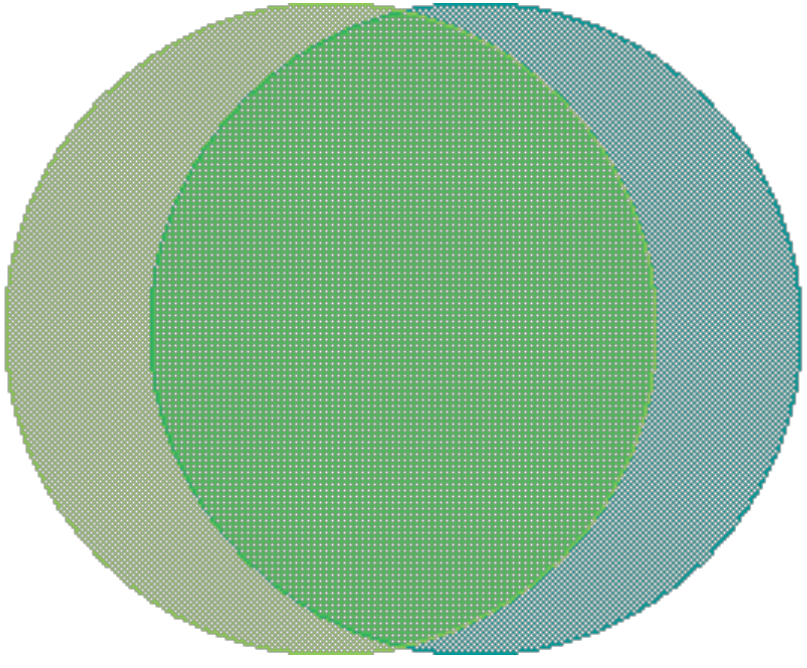
Predição - output

pred = [0, 1, 0, 1, 0, 0, 1, 1]

(buzios, ilha grande, **angra, balneário camboriú**)

RBM

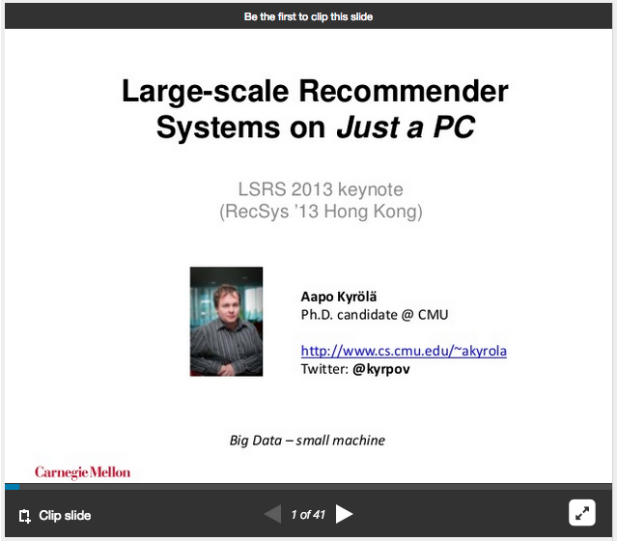
ALS



Na prática

- Funciona* bem com amostras do dataset
- Perfis vazios voltam com destinos **mais populares**
- Predição **não determinística**
- Pode ser usada antes de um modelo de Rating pra diminuir a quantidade de predições necessárias
- Ensemble!!!

http://www.slideshare.net/akyrola/largescale-recommendation-systems-on-just-a-pc



Perguntas?

@felipejcruz

github.com/felipecruz