Industrial Collaborations and Entity Resolution

Dr. Lucas Nissenbaum
03 de Fevereiro de 2023

IMPA:

A research center and a school of graduate studies, founded by the federal government in 1952.

A private legal entity collaborating with the Ministry of Science & Technology and the Ministry of Education, since 2000.

Our mission is to:

- Carry state-of-the-art research in mathematics.
- Train reaserchers and teachers at all levels.
- Disseminate mathematical knowledge in society.
- Integrate mathematics to science and industry.

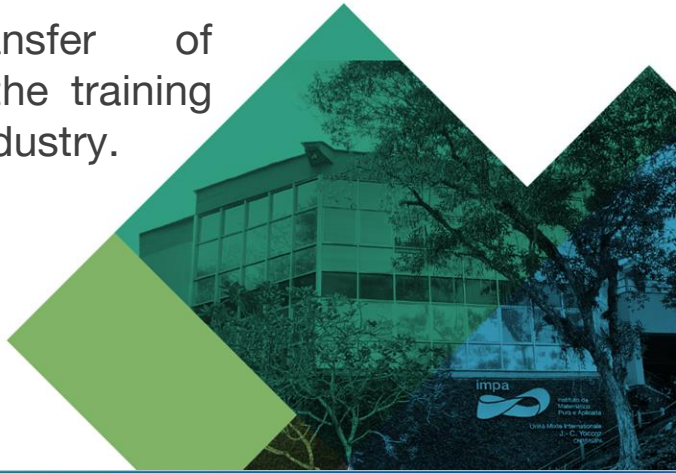
Innovation center in industrial mathematics

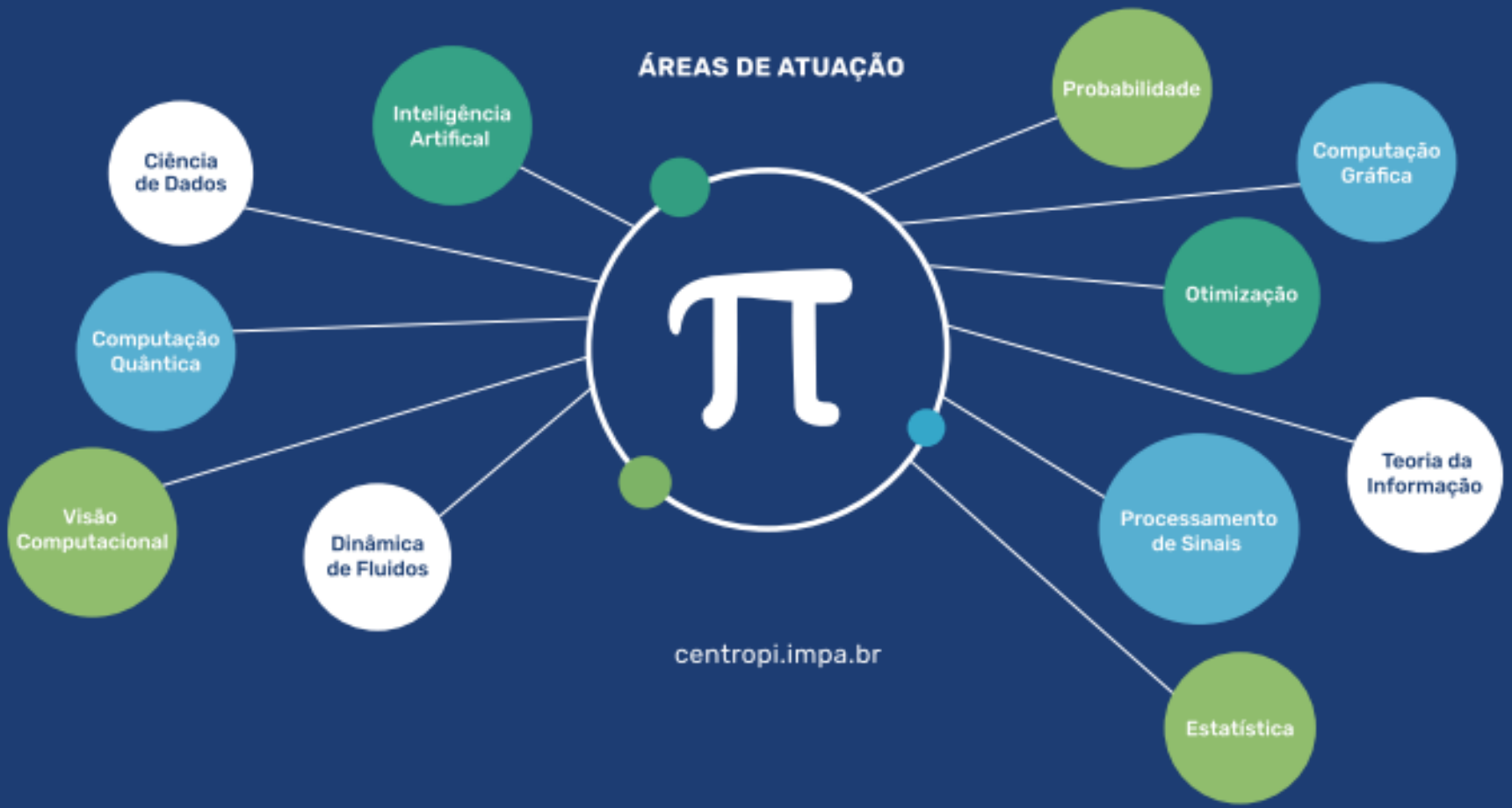
Centro Pi
Centro de Projetos
e Inovação IMPA

Solving concrete problems and developing projects that benefit from a strong contribution of mathematical sciences.

Contributing to the transfer of mathematical technology and the training of high-level professionals for industry.

centropi.impa.br




ÁREAS DE ATUAÇÃO
Ciência de Dados
Inteligência Artifical
Probabilidade
Computação Gráfica
Computação Quântica
Otimização
Visão Computacional
Dinâmica de Fluidos
Teoria da Informação
Processamento de Sinais
Estatística
centropi.impa.br


DASA


Big
Data


CARTESI


stone


VALE


McKinsey
& Company
shape


equinor


rumo


carteira
global


raízen
Energia que mobiliza


RADIX
Engenharia e Software


Centro Pi
Centro de Projetos
e Inovação IMPA


Tribunal
Superior
Eleitoral


Rio
PREFEITURA


GPP
Grupo de Políticas Públicas
USP - ESALQ


PETROS

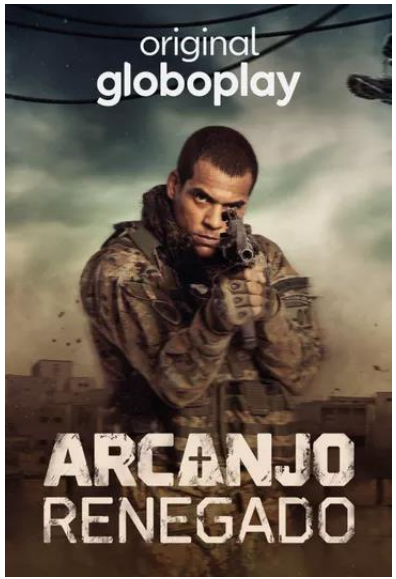

anos
impa
Instituto de
Matemática
Pura e Aplicada

Main goals of a collaboration:

1. Develop a product that solves an entity's challenge.

2. Progress mathematical research in areas associated with the problem.

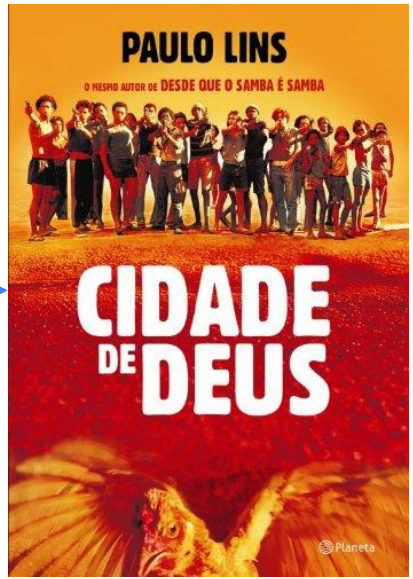
Metadata



- Title: Arcanjo Renegado
- Cast: Marcelo Mello Jr., Erika Januza
- Plot: Mikhael (Marcelo Mello Jr.) is the leader of BOPE's main squad. As one of his colleagues is hurt during a police operation, he opts to avenge him. The search for vengeance leads to a conflict with the political status quo.

For a recommendation based on meta-data

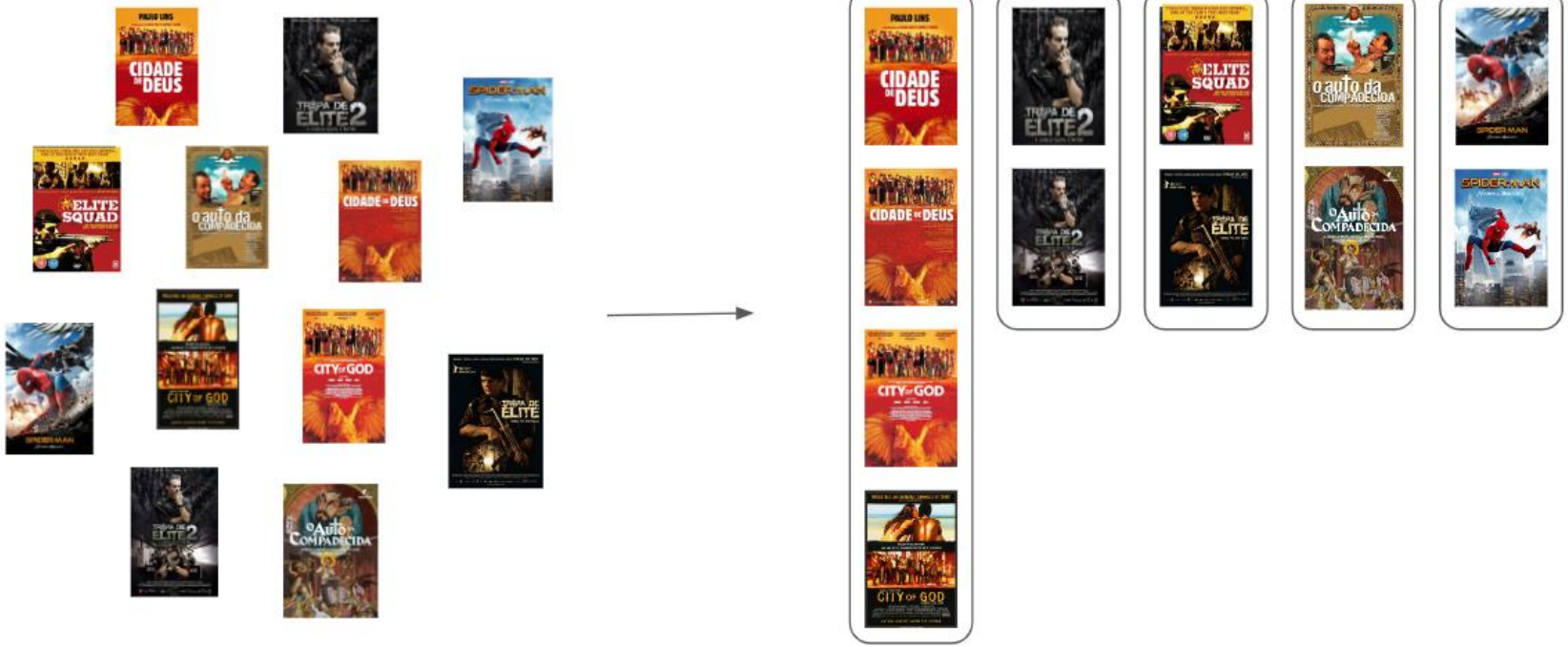
?



How do we build a recommendation system built on meta-data?

Step 1: Aggregation

Step 1: Aggregation



Step 2: Meta-data Extraction

Mikhael (Marcelo Mello Jr.) is the leader of BOPE's main squad. As one of his colleagues is hurt during a police operation, he opts to avenge him. The search for vengeance leads to a conflict with the political status quo.

Step 3: Recommendation

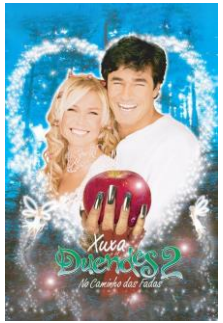
$$similarity = \rho($$  ,  $)$

Step 3: Recommendation

ρ(

, 

) > ρ(

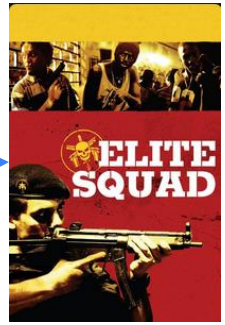
, 

)

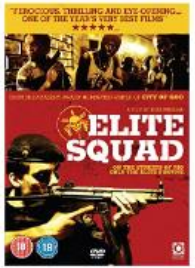
Step 3: Recommendation

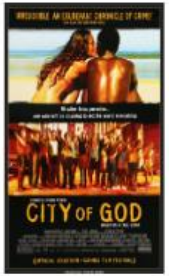
$$\begin{matrix} x_1 \\ \vdots \\ x_n \end{matrix} \qquad \rho(\vec{x}, \vec{y}) \qquad \begin{matrix} y_1 \\ \vdots \\ y_n \end{matrix}$$

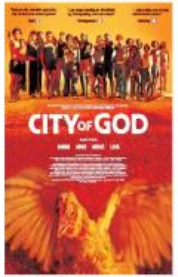
Step 3: Recommendation



Entity resolution



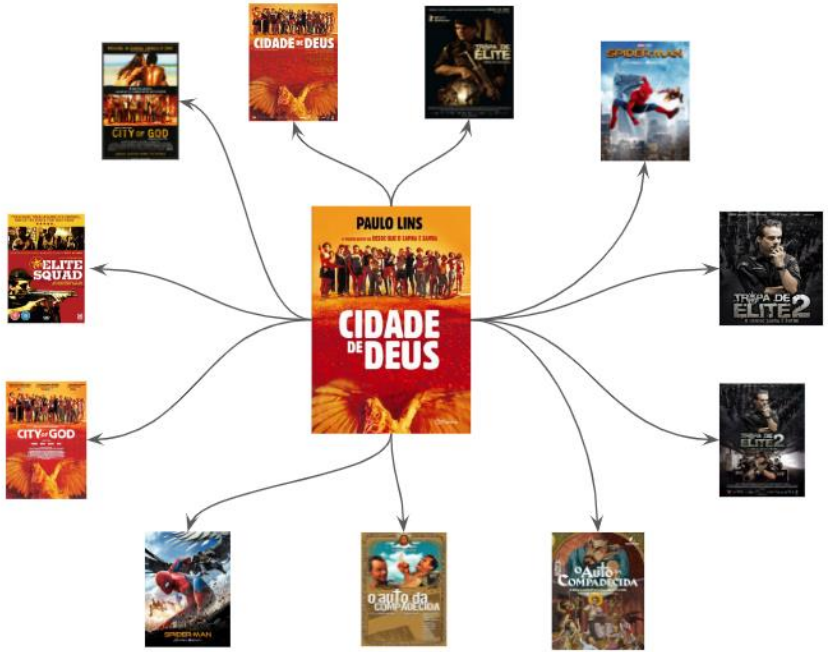




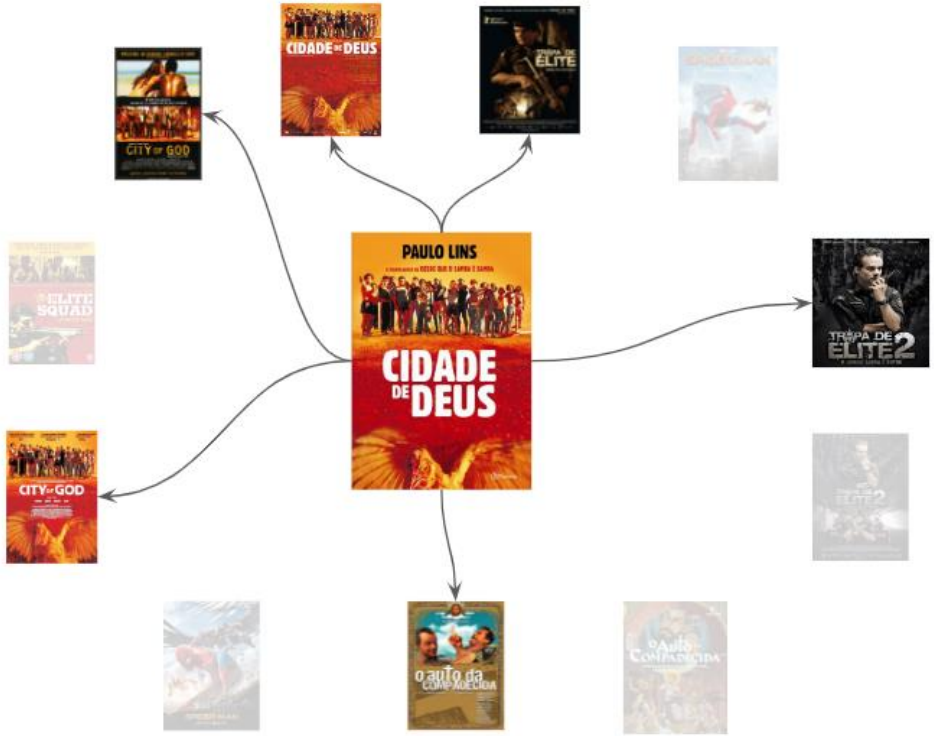
Applications

Name	Age	City
Lucas Nissembaum	32	Rio de Jameiro
Jorge Lopes	36	Rio de Janeiro
Jennifer Lopez	53	NY
Lucas Nissenbaum	32	Rio
Thiago Ramos	28	Araras

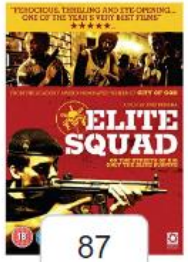
Entity resolution requires many comparisons



Do we need to try every pair?



Hashing



87

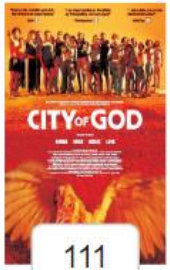
110


9

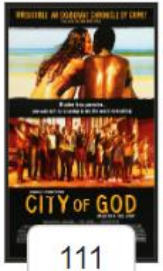
110

87

9



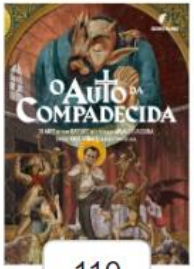
111



111

110

88



110

87

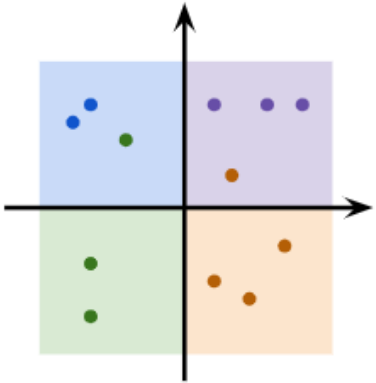
Hashes lead to fewer comparisons

Learn to hash

fixed hashes

Wide literature

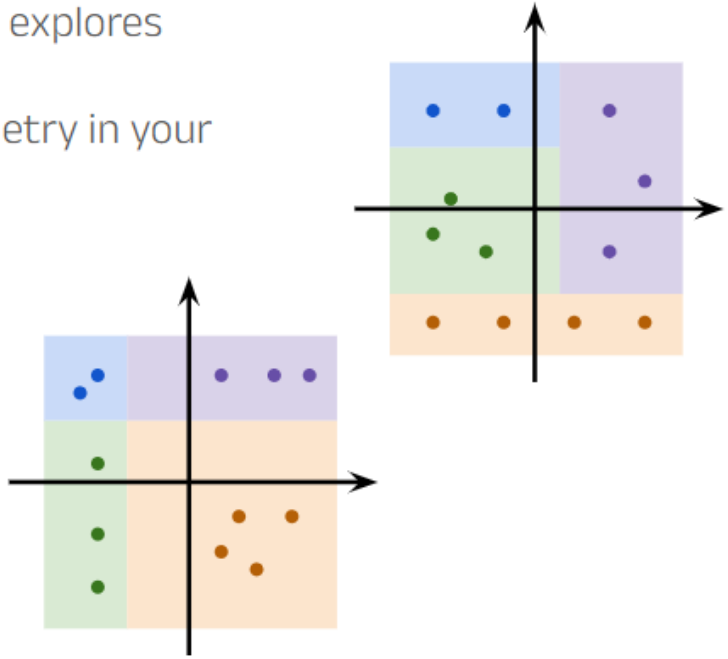
No training required



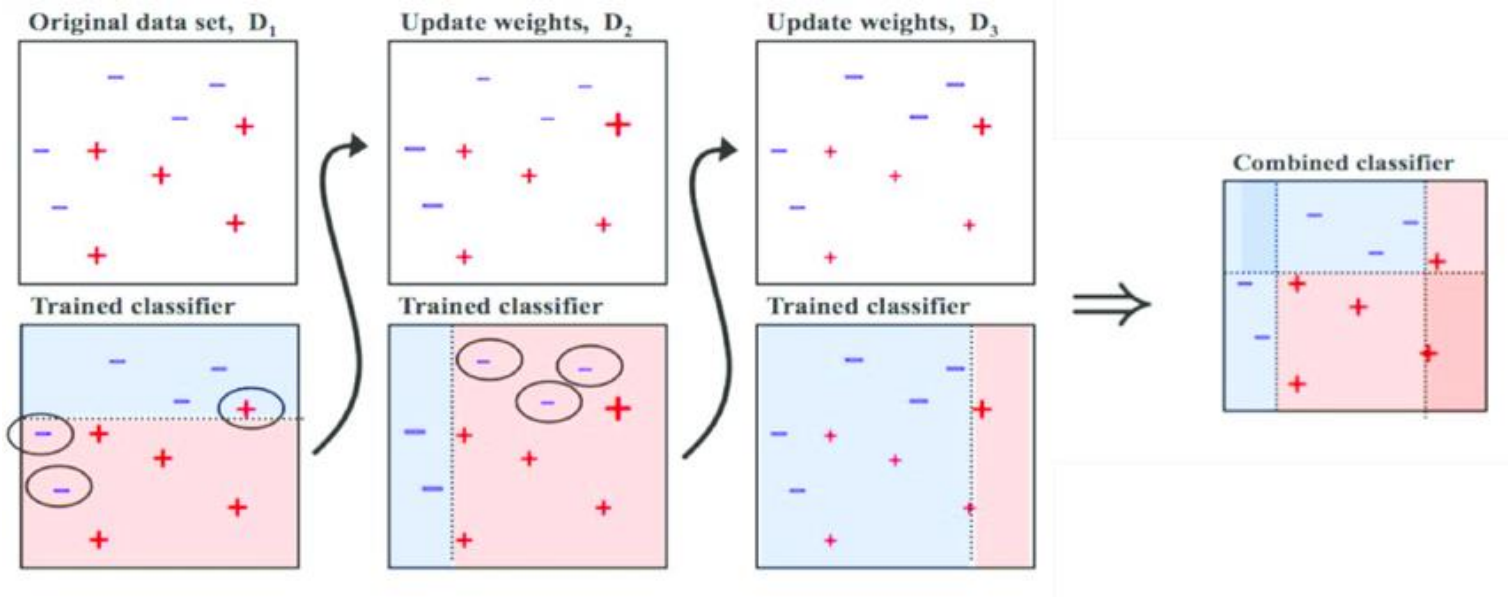
learned hashes

Recent and less explores

Learn the geometry in your feature space



Boosting



Fonte: https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851

Adapt boosting for hashing

Algorithm Algorithm to construct the hash codes

Require: $k, L \in \mathbb{N}$, convex weights$(\alpha_t)_{t=1}^{T}$, Rules $(\mathrm{Rule}_t)_{t=1}^{T}$

1: **for** $i \leftarrow 1$ to L **do**
2: **for** $j \leftarrow 1$ to k **do**
3: $g_{i,j} \leftarrow \mathrm{Rule}_t$ with probability α_t
4: **end for**
5: $g_i \leftarrow (g_{i,1}, \ldots, g_{i,k})$
6: **end for**
7: $g \leftarrow (g_1, \ldots, g_L)$
8: **return** g

Theoretical guarantees

Theorem 3.4. *Consider databases $\mathcal{A}$ and $\mathcal{B}$ such that $|\mathcal{A}| = N_{\mathcal{A}}$ and $|\mathcal{B}| = N_{\mathcal{B}}$. If Condition 1 holds for the output f^* of Algorithm 4 for a given $\theta > 0$, $\gamma \in (0,1)$ is given, and we set:*

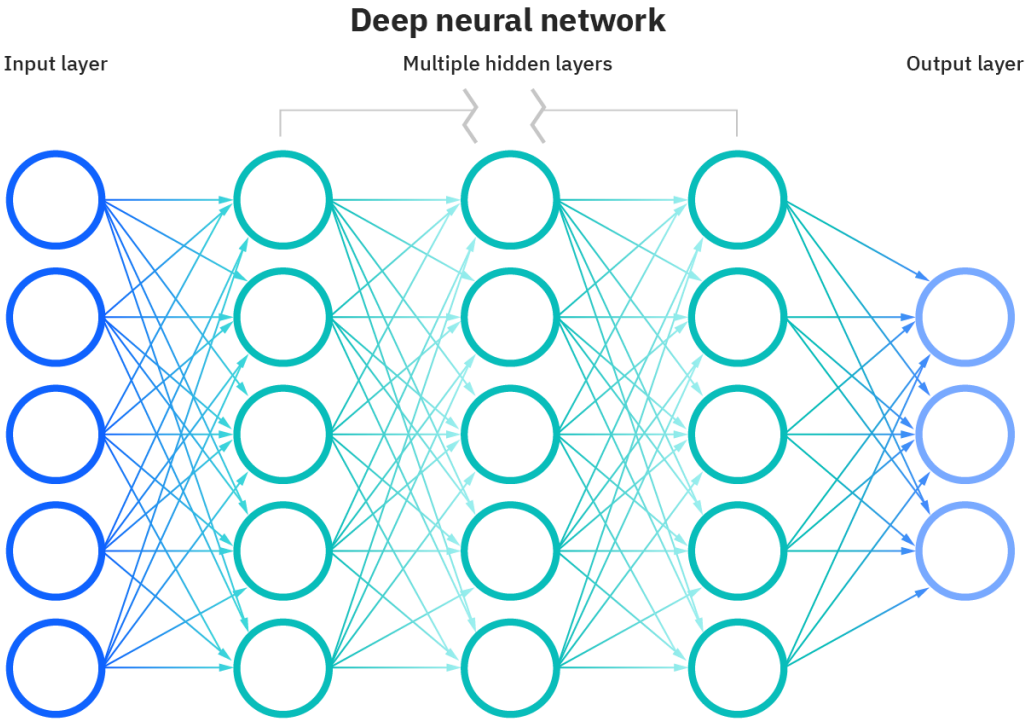
$$\rho := \frac{\log\left(\frac{2}{1+\theta}\right)}{\log\left(\frac{2}{1-\theta}\right)} \in [0,1),\ k := \left\lceil \log_{\frac{2}{1+\theta}} N_{\mathcal{A}} \cdot N_{\mathcal{B}} \right\rceil \text{ and } L := \left\lceil \frac{2(N_{\mathcal{A}} \cdot N_{\mathcal{B}})^{\rho} \log(1/\gamma)}{1+\theta} \right\rceil,$$

then Algorithm 5 achieves the following expected values for the Recall and RR metrics defined in (3.2) and (3.3):

$$\mathbb{E}\left[\text{Recall}\right] \geq (1-\gamma)(1-\varepsilon)$$
$$\mathbb{E}\left[\text{RR}\right] \geq \left(1 - \frac{|\mathcal{M}| + L}{N_{\mathcal{A}} \cdot N_{\mathcal{B}}}\right)(1-\varepsilon).$$

Both expectations are with respect to the randomness in the hash code.

Neural networks



Fonte: https://www.ibm.com/cloud/learn/neural-networks

Estrada Dona Castorina, 110 Jardim Botânico
22460-320, Rio de Janeiro, RJ – Brasil
(21) 2529-5000 | impa.br