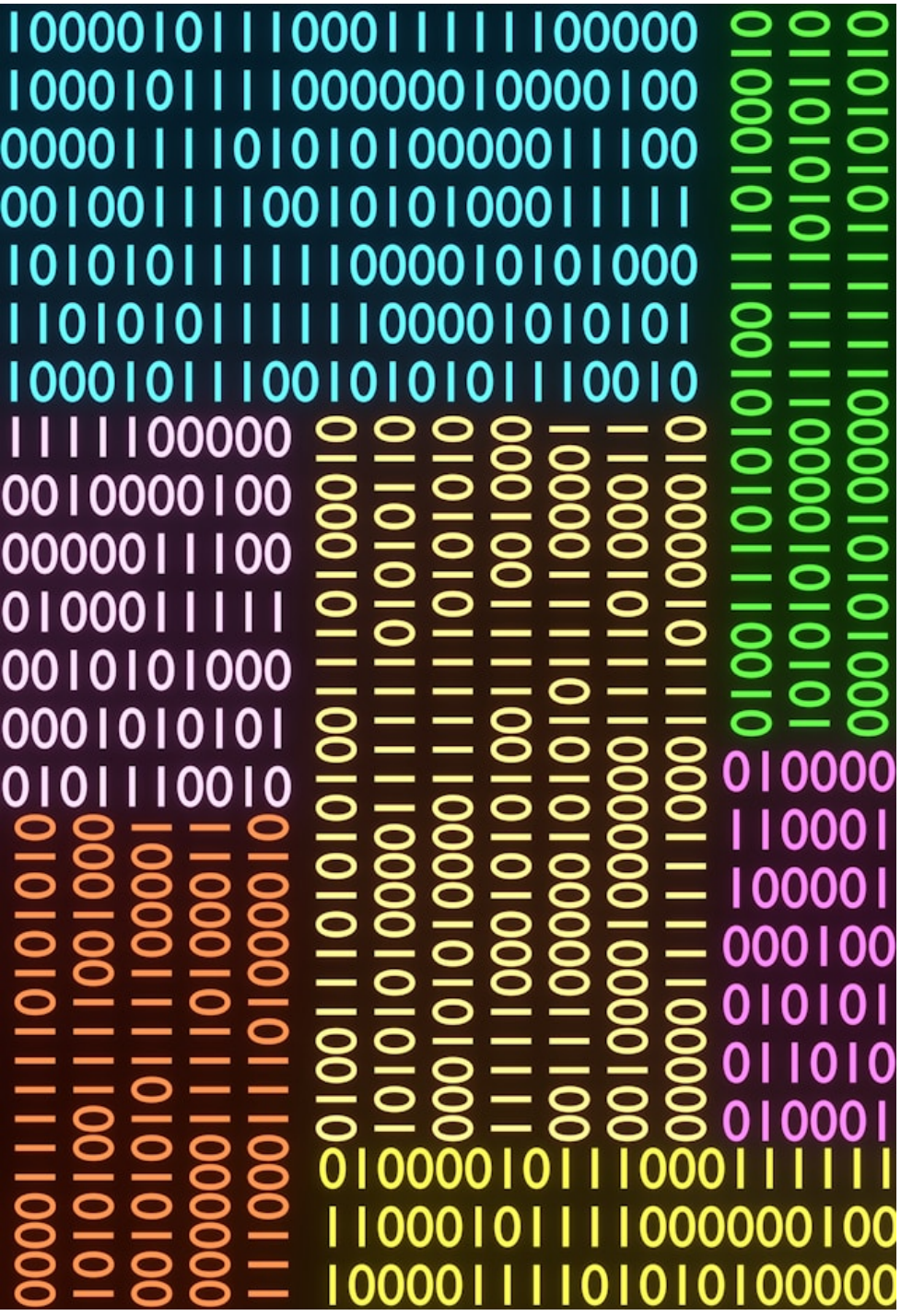


Ecossistemas de Dados Alimentam IA

Prof. Dr. Denis M. L. Martins
DCM | FFCLRP | USP

Sobre mim

- Engenheiro da Computação pela Universidade de Pernambuco
- Doutorado na Uni Münster
- Eng. de Software na Stefanini
- Pesquisador Sênior em IA na Samsung Research Brazil
- Professor na PUC-Campinas
- **Professor no DCM-USP**

> **Especialidade: Dados e IA**

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

"Data is the new Oil"

(Dados são o novo petróleo)

Forbes

INNOVATION

Data Is The New Oil -- And That's A Good Thing

By Kiran Bhageshpur, Former Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Published Nov 15, 2019, 08:15am EST, Updated Apr 14, 2022, 02:04pm EDT

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

Forbes

INNOVATION > AI

Data as The New Oil Is Not Enough: Four Principles For Avoiding Data Fires

By Nisha Talagala, Contributor. ⓘ Entrepreneur and technologist in AI and AI L... Follow Author

Published Mar 02, 2022, 05:48pm EST, Updated Mar 04, 2022, 05:05am EST

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

Forbes

INNOVATION

Stop Thinking Of Data As The New Oil

By Juan Carlos Santiago, Forbes Councils Member.

for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Published Jul 09, 2025, 09:45am EDT

Forbes

SMALL BUSINESS

If Data Is The New Oil, Decision Science Is The New Refinery

By Chris Chambers, MBA, Forbes Councils Member.

for Forbes Business Council, COUNCIL POST | Membership (fee-based)

Published Oct 07, 2025, 08:15am EDT

Dados

"Data is the new Oil".

- Frase de Clive Humby no início da década de 2000.
- Ficou popular na década de 2010.

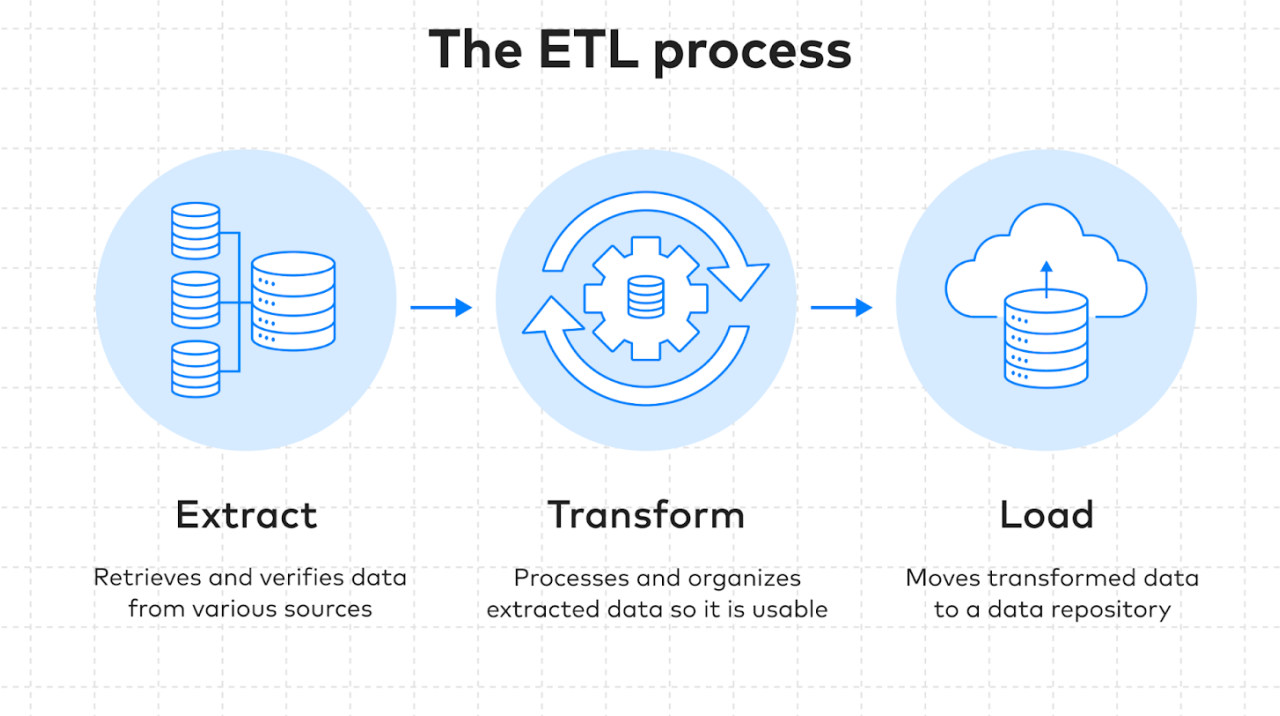
> Na foto: Clive Humby. Fonte: Retailer Insider.

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

"Data is the new Oil"

Dados **geram valor** para organizações que conseguem:

- Explorá-los
- Refiná-los.



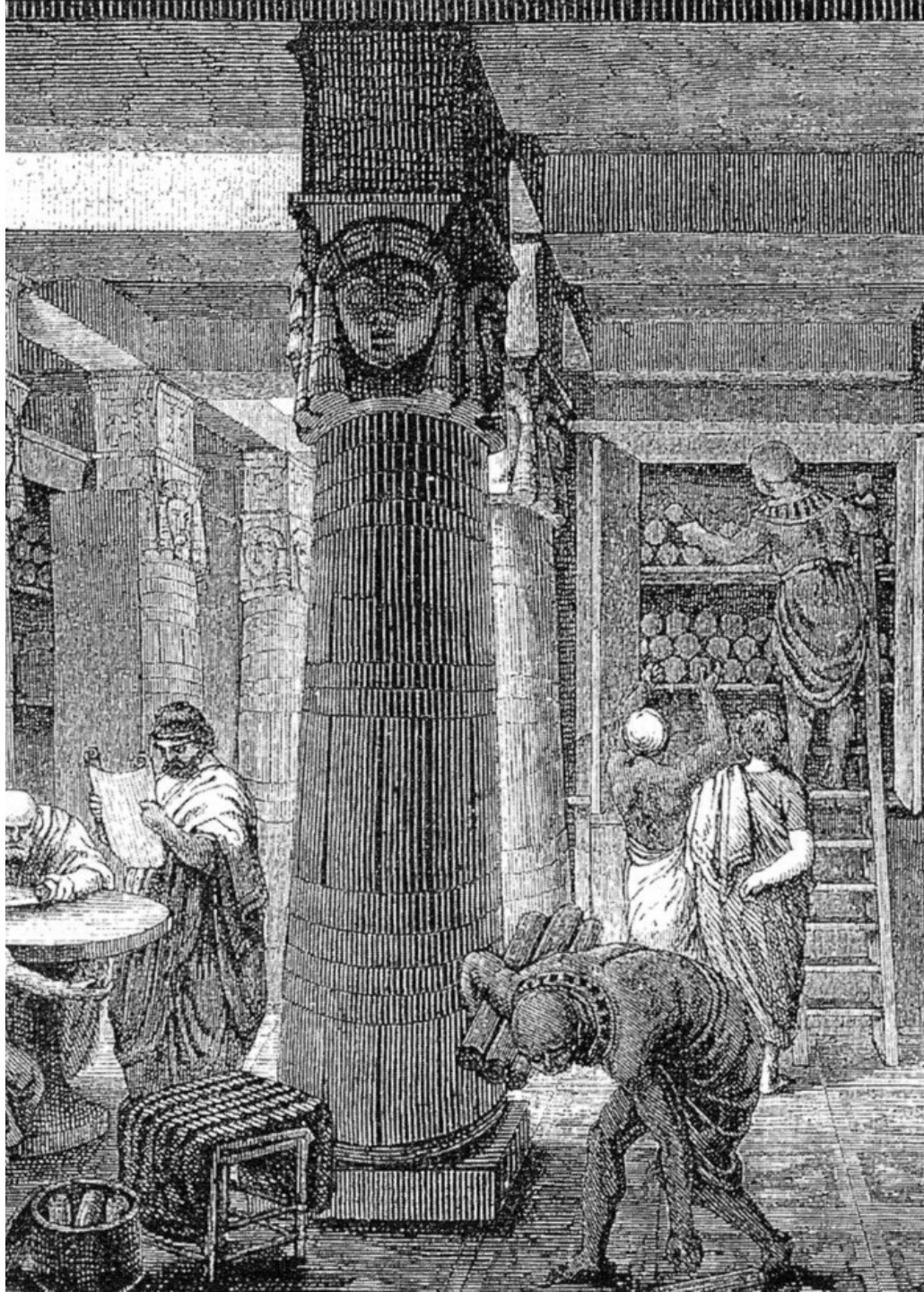
> Na imagem: Processo de Extração, Transformação e Carregamento de Dados.
> Fonte: Fivetran

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

"Data is the new oil"

- Biblioteca de Alexandria (século 3 AEC)
- 40.000 a 400.000 textos/pergaminhos (~100.000 livros).
- **ETL** (extract, transform, load) por humanos.
- **Single point of failure**.
- Infelizmente, **sem backup**.

> Na imagem: Biblioteca de Alexandria pelo artista Von Corven.
> Fonte: Wikipedia

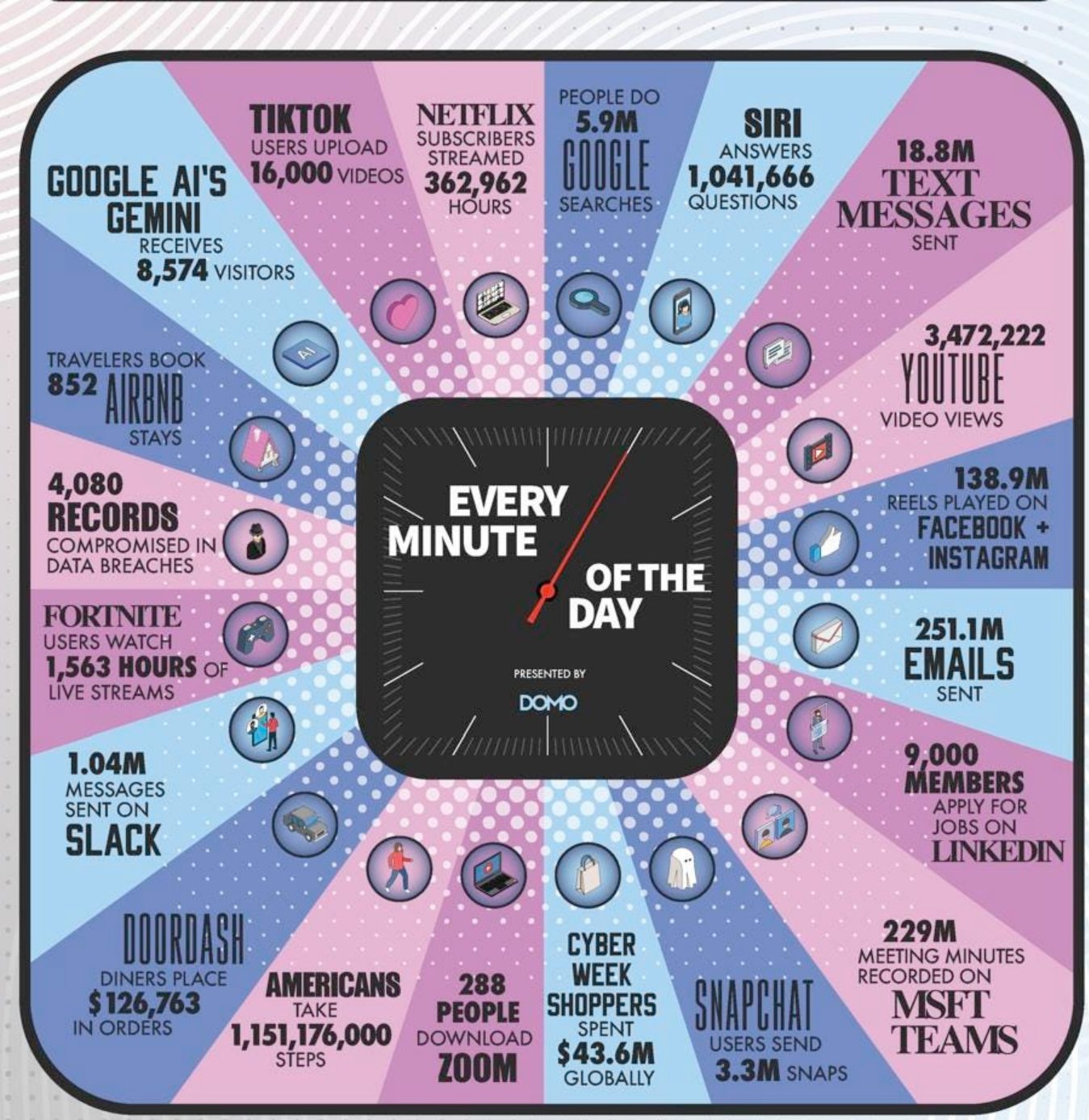


Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

"Data is the new oil"

- **Geração massiva** de dados por máquinas e humanos.
- **GenAI** pronta para destronar os pilares competitivos da Internet.
- **ETL automático** (pipelines manuais e baseados em agentes)

> Na imagem: Report da DOMO: Data Never Sleeps, 12a edição.



Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

"Data is the new oil" (Será?)

Vários **problemas** com essa frase. Dados são:

- (virtualmente) infinitos
- fáceis de replicar
- fáceis de transportar
- podem perder valor facilmente

Mas, assim como petróleo:

- valor é dependente do contexto
- demanda energia e recursos finitos
- motivos de disputa

> Na imagem: Bernard Marr. Fonte: bernardmarr.com

Bernard Marr & Co.
Future • Business • Success

English

Here's Why Data Is Not The New Oil

Written by
Bernard Marr

Bernard Marr is a world-renowned futurist, influencer and thought leader in the fields of business and technology, with a passion for using technology for the good of humanity. He is a best-selling and award-winning author of over 20 books, writes a regular column for Forbes and advises and coaches many of the world's best-known organisations. He has a combined following of 5 million people across his social media channels and newsletters and was ranked by LinkedIn as one of the top 5 business influencers in the

bernardmarr.com

It's a claim you've probably heard multiple times – "Data is the new oil!"

Now it's true, that in some ways, the analogy fits – it's easy to draw parallels due to the way information (data) is used to power much of the transformative technology we see today – artificial intelligence, automation and advanced, predictive analytics.

However, in many ways, it's also lazy and inaccurate – and while it's handy

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

IA moderna acontece em um *cenário heterogêneo* de sistemas de informação empresariais (ERP), data lakes analíticos, gateways de IoT...

(*Tudo isso geralmente desconectado*)

Dados Insulares

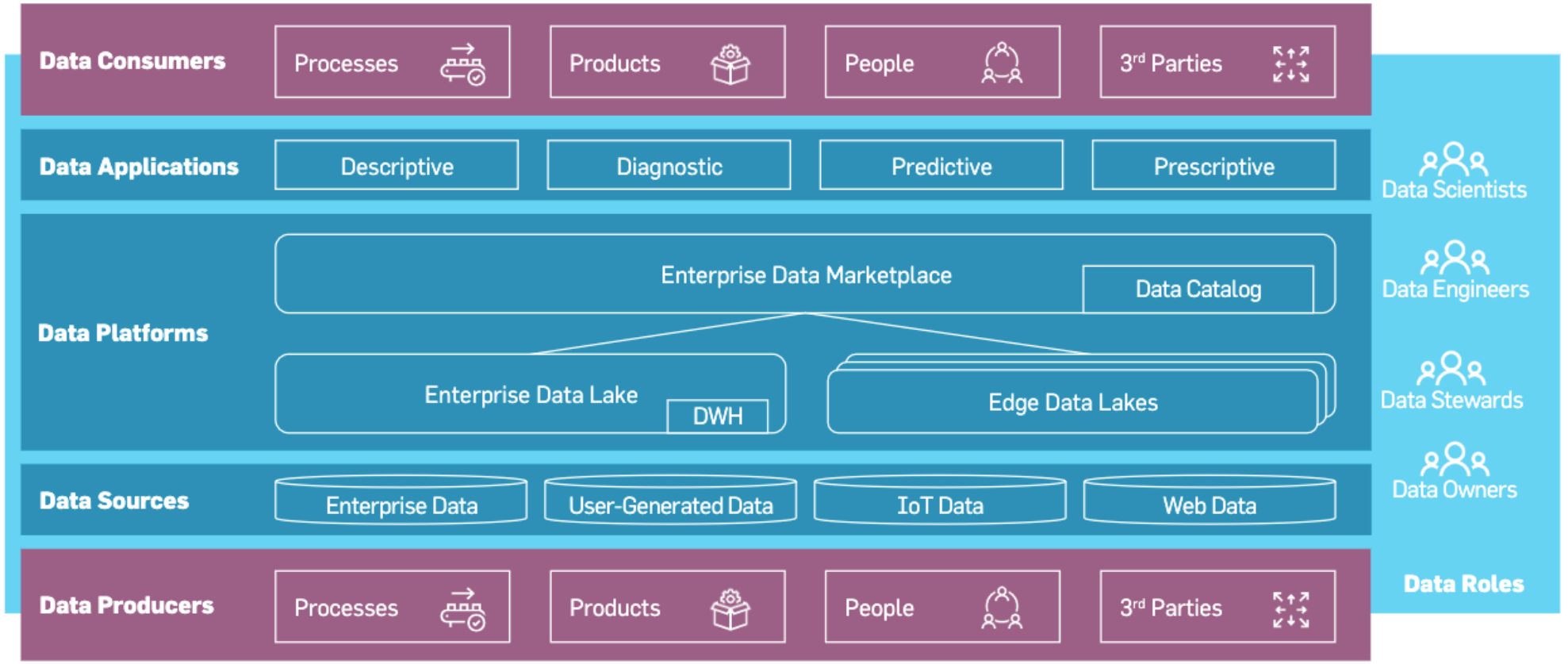
- **Soluções Isoladas:** A IA é frequentemente implementada em "ilhas".
  - Extrações de dados **específicas** para casos de uso e modelos customizados.
  - Sem **diretrizes** de metadados e bancos de dados isolados
  - Dificuldade de **reuso** e de escalabilidade.
- **O Custo da Ineficiência:** Casos de uso idênticos são muitas vezes recriados *do zero*, gerando cargas altas e redundantes nos sistemas de origem críticos.

> Na imagem: GRÖGER, Christoph. There is no AI without data. Communications of the ACM, v. 64, n. 11, p. 98-108, 2021.

Rumo à IA "Industrializada"

- Estabelecer um ecossistema de dados abrangente.
- Problema **socio-técnico**: Atores, sistemas e culturas heterogêneos.
- **Pilar de Gestão de Dados:** Resolve a modelagem unificada e silos de metadados.
- **Pilar de Democratização:** Foca em transformar a descoberta e engenharia de dados em capacidades de autoatendimento para todos os usuários.
- **Pilar de Governança:** Estabelece papéis formais como Data Owners para garantir conformidade e provimento ágil.

Figure 3. Core elements of a data ecosystem for industrial enterprises.



GRÖGER, Christoph. There is no AI without data. Communications of the ACM, v. 64, n. 11, p. 98-108, 2021.

Escalar a IA em larga escala exige um *design colaborativo* de sistemas de dados e modelos inteligentes.

Pilar de Co-Design

Explorar a **sinergia** entre modelos de IA e sistemas de dados (DBMS).

Análise é realizada **fora do banco de dados**.

- Acesso aos dados é geralmente um gargalo.

Mova a análise, não os dados

> Fonte da imagem: This is Engineering via Unsplash

Exemplo

Em várias aplicações, objetos de interesse são raros em dados massivos.

- Como encontrar uma agulha num palheiro.

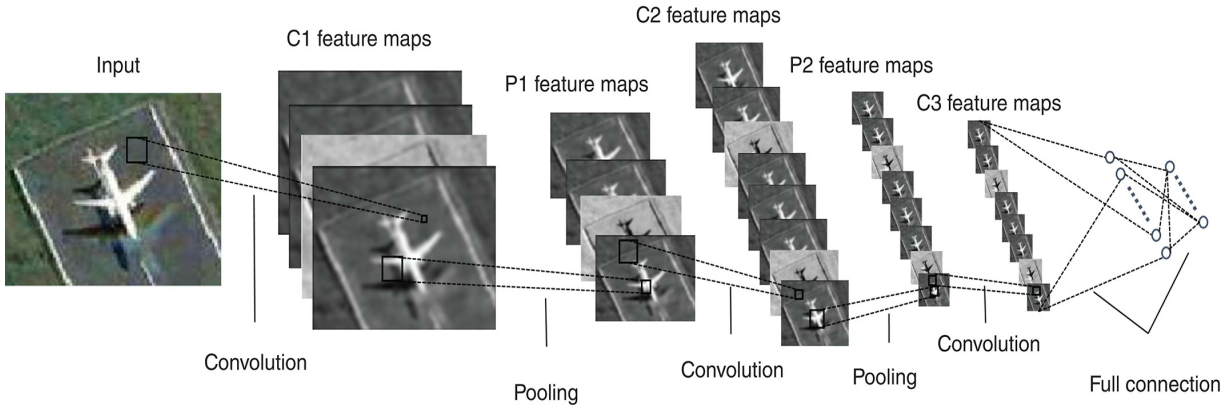
Exemplo: Encontrar turbinas eólicas em imagens de satélite.

- Objeto difícil de descrever em SQL.
- Fácil de conseguir exemplos (imagens).

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

Exemplo

Podemos *anotar* alguns dados e treinar um classificador (*e.g.*, CNN) com IA.



Com o classificador treinado, podemos aplicá-lo ao **banco de dados completo** para encontrar os objetos de interesse.

Considere executar uma CNN *1 bilhão* de patches de imagens...

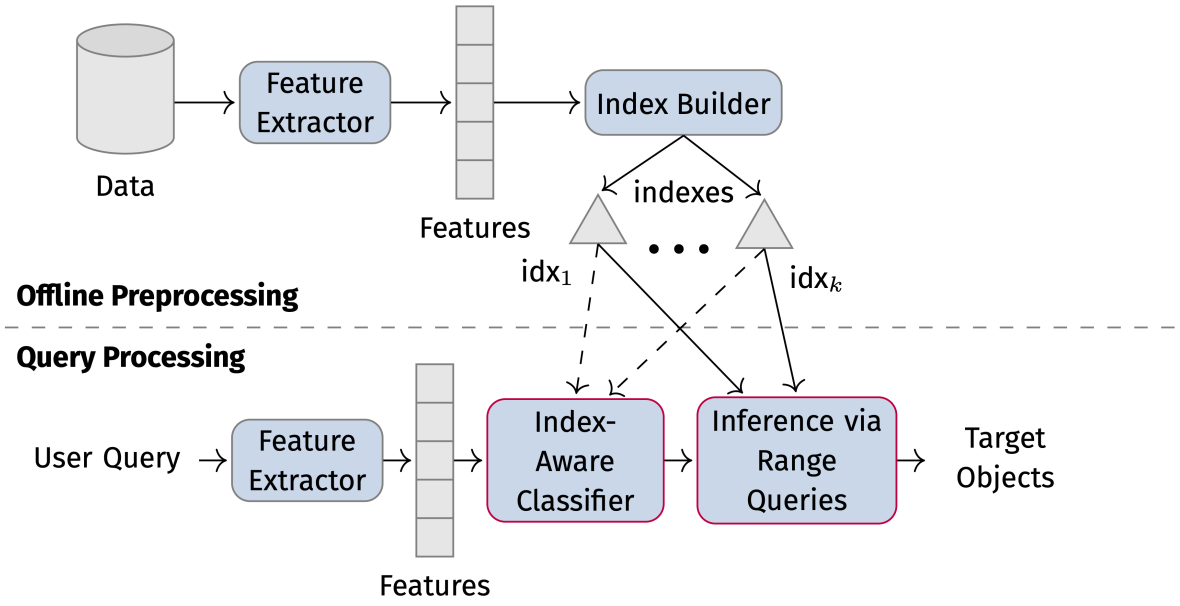
(*Vai por mim. Vai demorar*)

Co-Design: Usando Índices

A busca tradicional exige escanear todo o banco de dados.

- Ineficiente
- Ignora o potencial do DBMS.

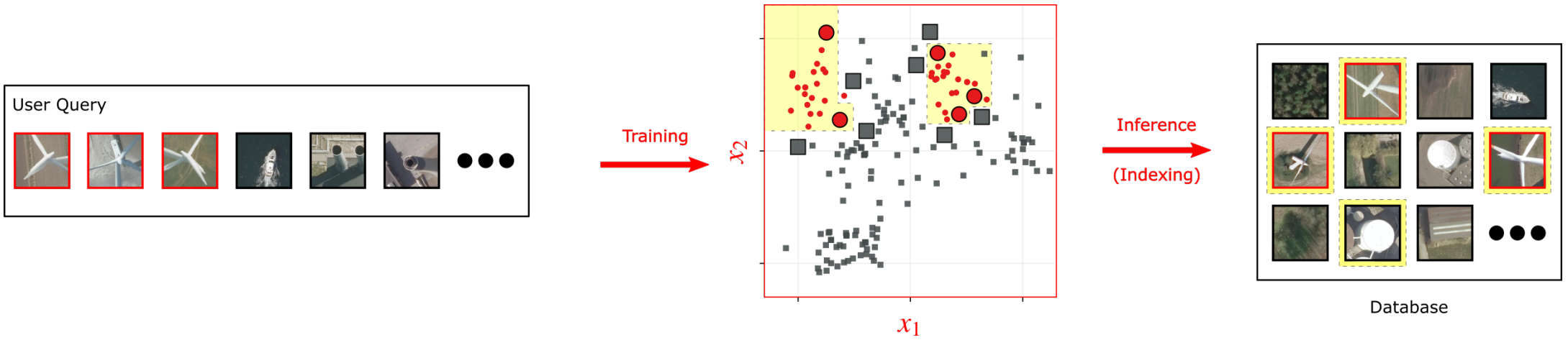
Co-design de índices multidimensionais e modelos de IA.



> Fonte: LÜLF, Christian et al. Fast search-by-classification for large-scale databases using index-aware decision trees and random forests. arXiv preprint arXiv:2306.02670, 2023.

Co-Design: Usando Índices

Treinar modelos baseados em árvores de decisão.

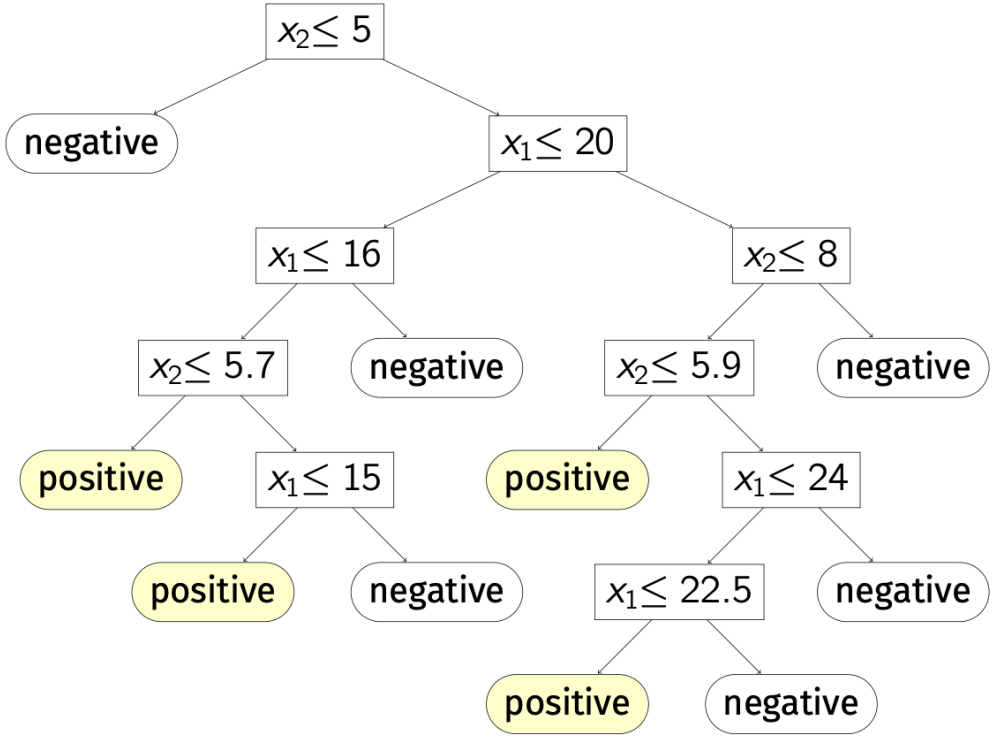


> Fonte: LÜLF, Christian et al. Fast search-by-classification for large-scale databases using index-aware decision trees and random forests. arXiv preprint arXiv:2306.02670, 2023.

Co-Design: Usando Índices

Index-aware: Modelos $\longrightarrow$ SQL sobre índices pré-existentes, transformando a inferência em consultas de intervalo.

Pense que a árvore de decisão vira a cláusula **WHERE** de uma consulta SQL.

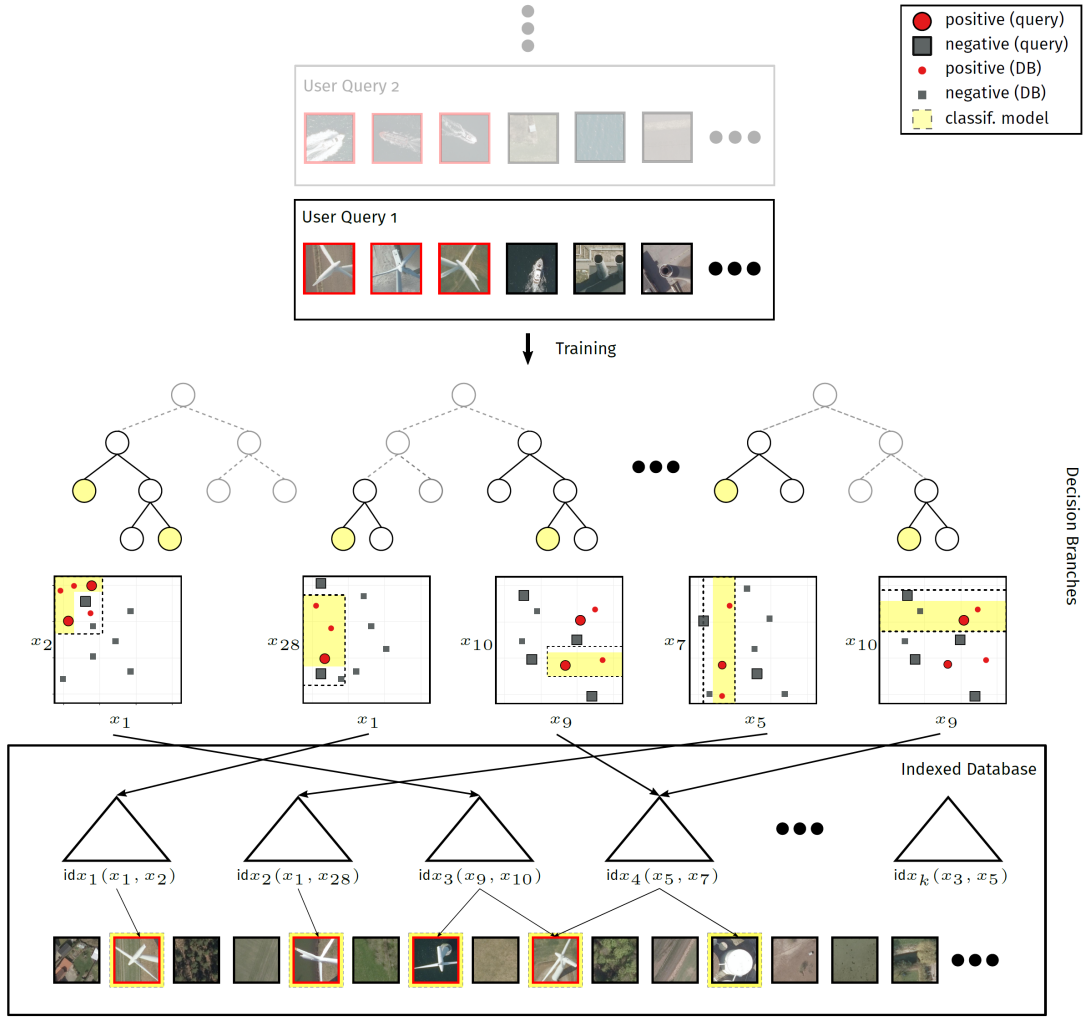


> Fonte: LÜLF, Christian et al. Fast search-by-classification for large-scale databases using index-aware decision trees and random forests. arXiv preprint arXiv:2306.02670, 2023.

Co-Design: Usando Índices

1. Índices são pré-definidos sobre pequenos subconjuntos de *features*.
2. *Decision Branches* são treinadas nestes subconjuntos.
3. Durante a inferência, os índices são utilizados, evitando *full scan*.

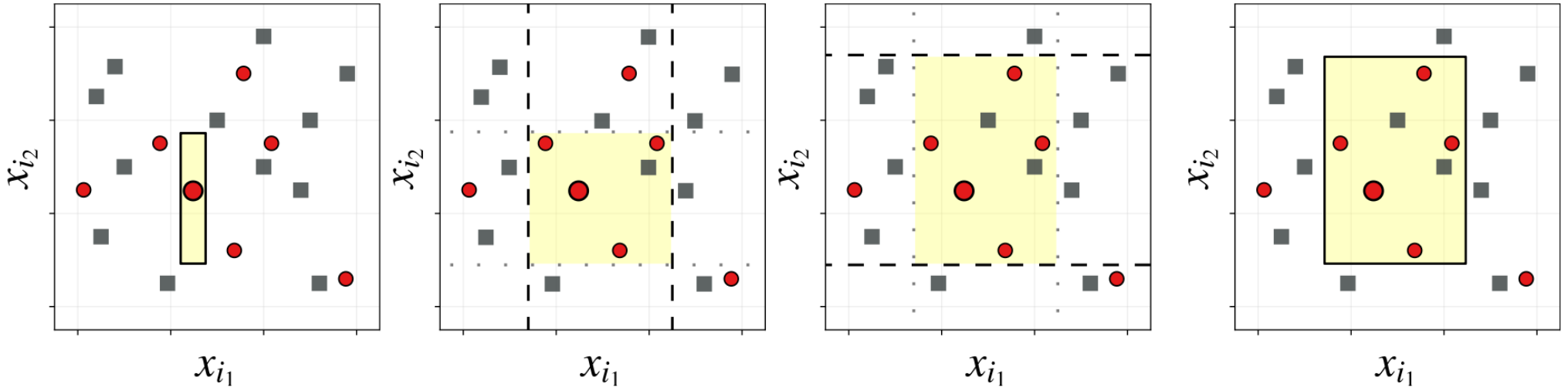
> Fonte: LÜLF, Christian et al. Fast search-by-classification for large-scale databases using index-aware decision trees and random forests. arXiv preprint arXiv:2306.02670, 2023.



Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

Co-Design: Usando Índices

Construir modelos bottom-up para facilitar a utilização dos índices. (Por que?)



> Fonte: LÜLF, Christian et al. Fast search-by-classification for large-scale databases using index-aware decision trees and random forests. arXiv preprint arXiv:2306.02670, 2023.

Co-Design: Usando Índices

Melhor desempenho de classificação em comparação com **Random Forest**.

Muito mais **~200 vezes mais rápido**.

> Fonte: LÜLF, Christian et al. Fast search-by-classification for large-scale databases using index-aware decision trees and random forests. arXiv preprint arXiv:2306.02670, 2023.

Model	T_{train}	T_{query}	T_{total}	F_1-score
DBranch	0.398	1.047	1.445	0.833
DTree	0.855	1,043.433	1,044.288	0.829
DBEns	0.993	5.666	6.658	0.914
RForest	0.274	1,319.688	1,319.961	0.904
ExTrees	0.122	1,332.026	1,332.148	0.950

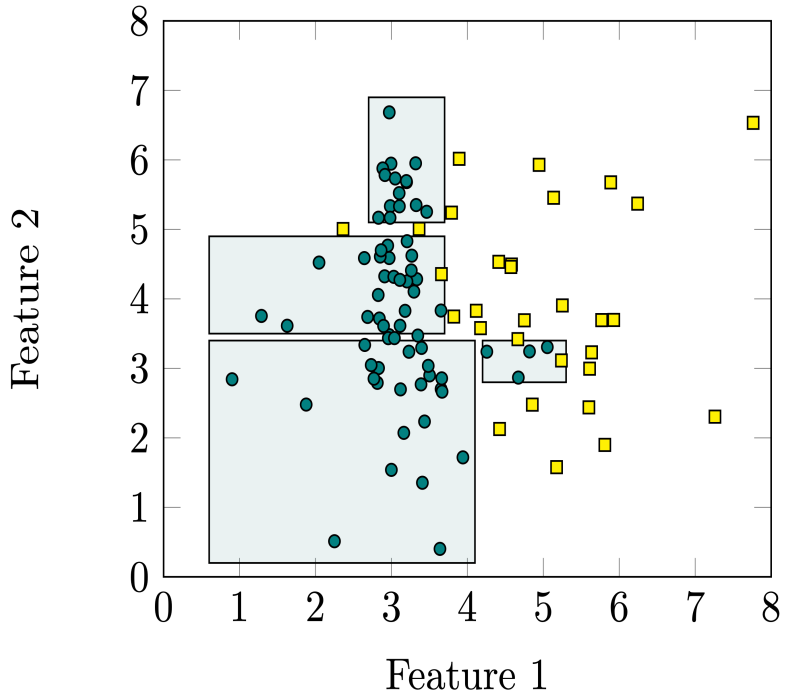
Mas esse processo não é end-to-end...

(É preciso confiar nas features extraídas...)

Co-Design: Generalizando Modelos $\longrightarrow$ SQL

Objetivo: Encontrar um conjunto de *hyperboxes* em um espaço multidimensional de *features*. As *hyperboxes* devem "cobrir" objetos de interesse.

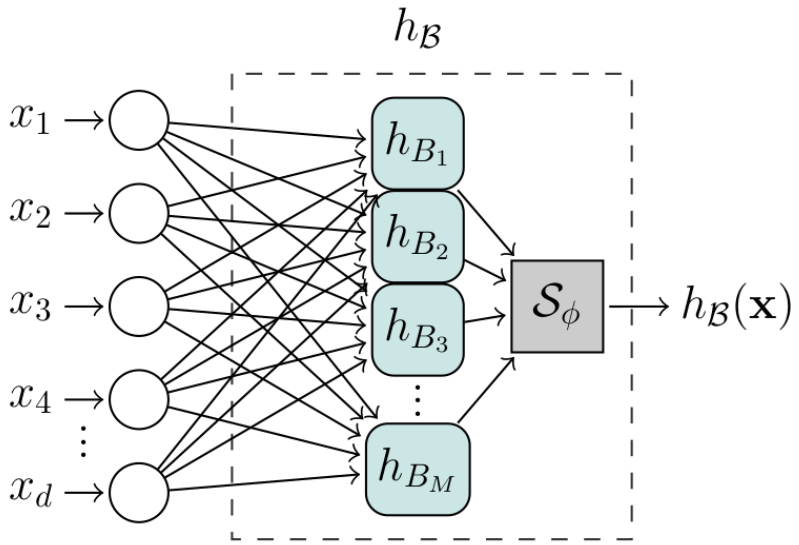
- Minimizar uma função de perda $\mathcal{L}$.
- Fácil de treinar e executar (inferência).
- Poder transferir aprendizado entre domínios.



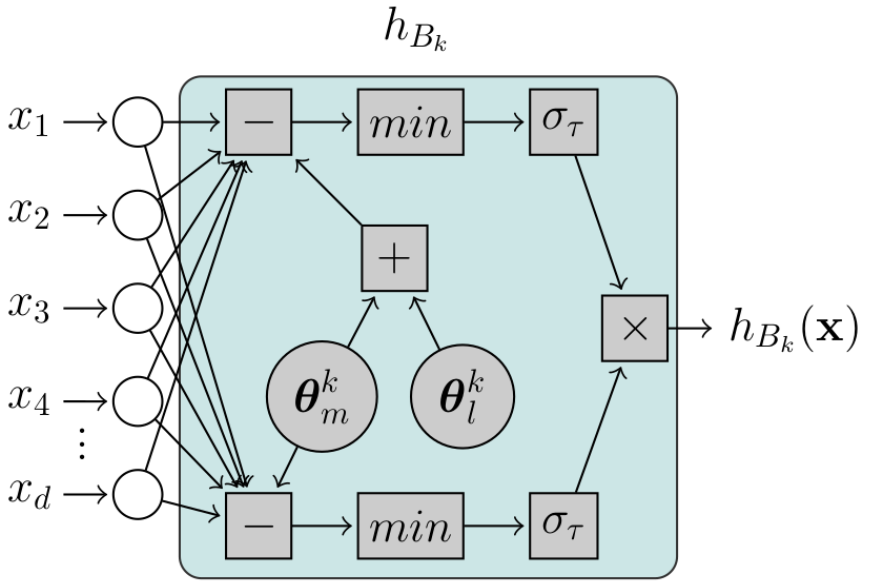
Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

Co-Design: Generalizando Modelos $\longrightarrow$ SQL

HyperNN: Deep Learning $\longrightarrow$ Hyperboxes $\longrightarrow$ SQL



HyperNN Network Structure

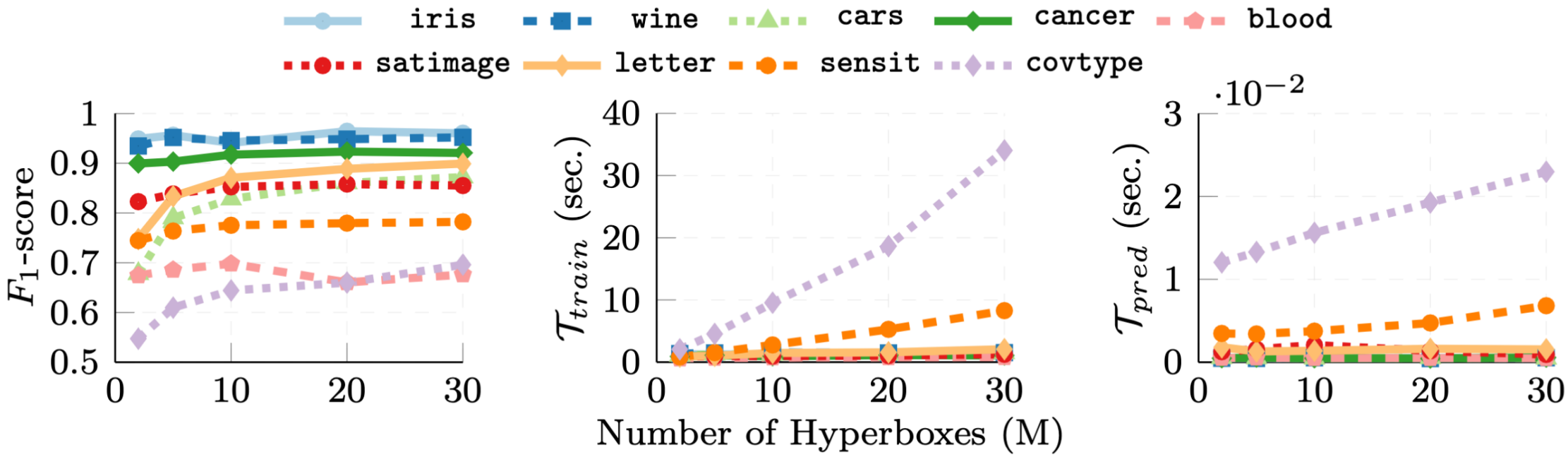


Hyperbox Neuron

Fonte: MARTINS, Denis Mayr Lima; LÜLF, Christian; GIESEKE, Fabian. End-to-End Neural Network Training for Hyperbox-Based Classification. arXiv preprint arXiv:2307.09269, 2023.

Co-Design: Generalizando Modelos $\longrightarrow$ SQL

HyperNN: Deep Learning $\longrightarrow$ Hyperboxes $\longrightarrow$ SQL
Decisão da Rede Neural pode ser traduzida facilmente em SQL.



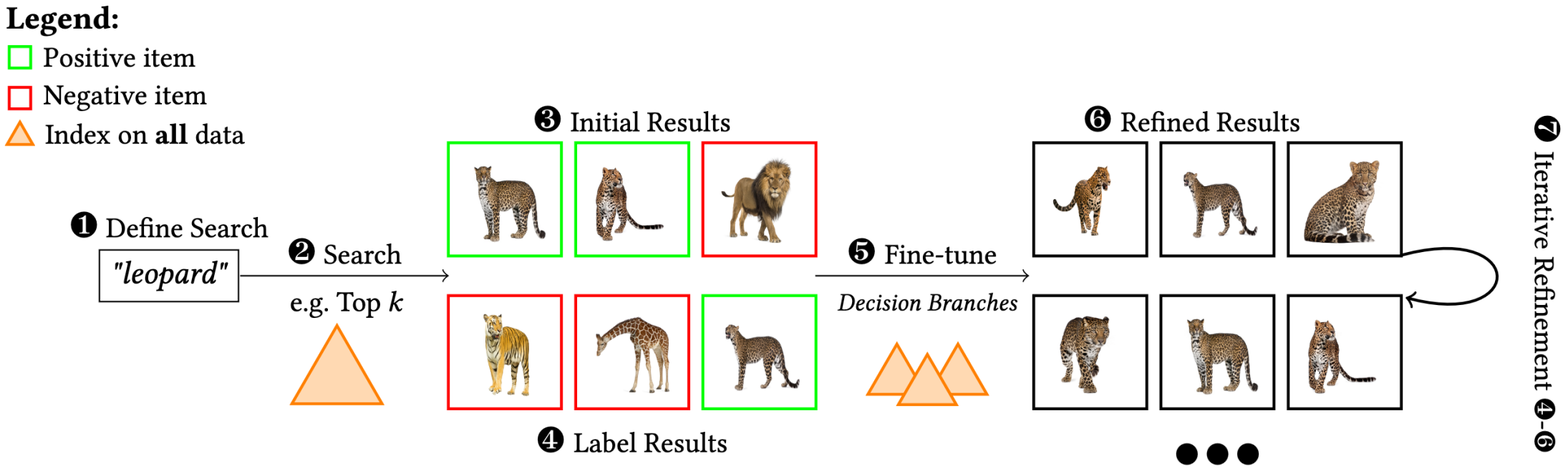
> Fonte: MARTINS, Denis Mayr Lima; LÜLF, Christian; GIESEKE, Fabian. End-to-End Neural Network Training for Hyperbox-Based Classification. arXiv preprint arXiv:2307.09269, 2023.

Ok. Mas e os LLMs...

(*Como fazer esse co-design?*)

Co-Design: Embeddings + Index

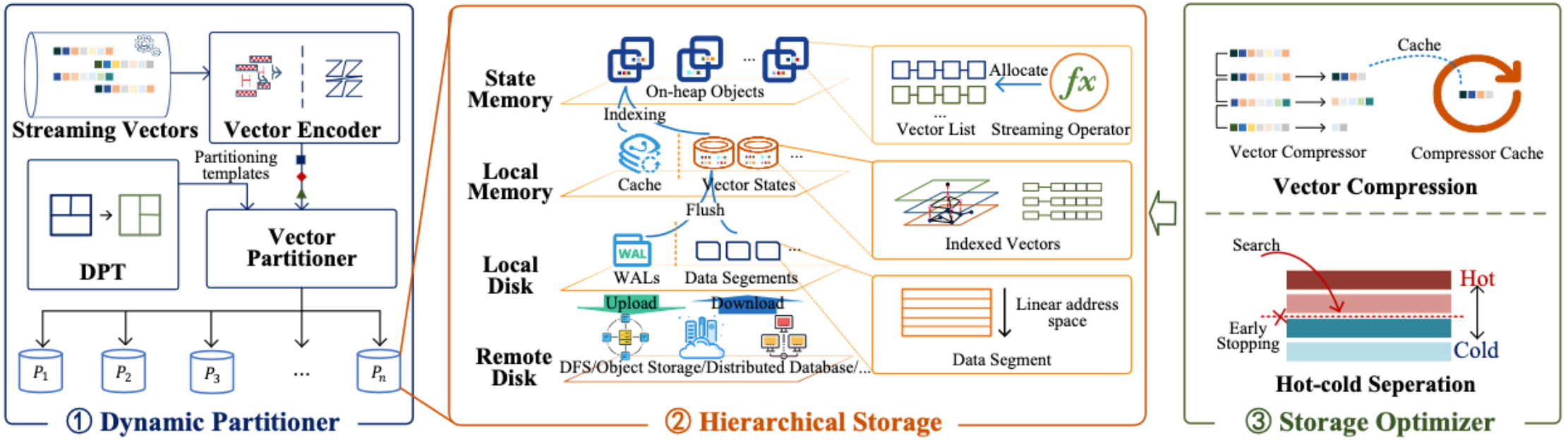
Arquitetura CLIP para *text-image embeddings*.
Busca tradicional + Refinamento via *Decision Branches*.



> LÜLF, Christian et al. Clip-branches: Interactive fine-tuning for text-image retrieval. In: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval. 2024. p. 2719-2723.

Co-Design: Eficiência de RAG

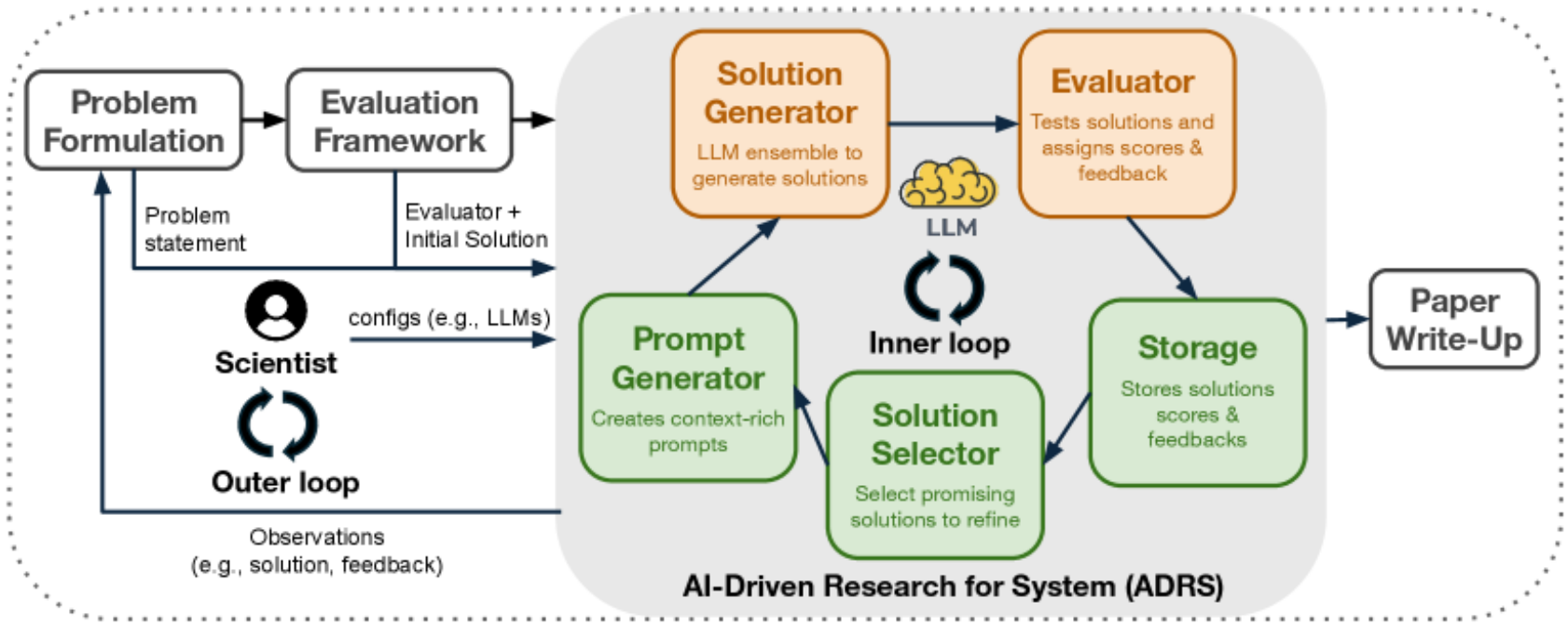
VStream: 251–373× melhor em eficiência de consultas e 1.5–2.0× menor em termos de sobrecarga de memória.



Fonte: GONG, Shenghao et al. VStream: A distributed streaming vector search system. Proceedings of the VLDB Endowment, v. 18, n. 6, p. 1593-1606, 2025.

Perspectivas Futuras

AI-Driven Research for Systems (ADRS): *"Much like an academic advisor guides a student, the researcher of the future will act as a guide for these AI systems."*



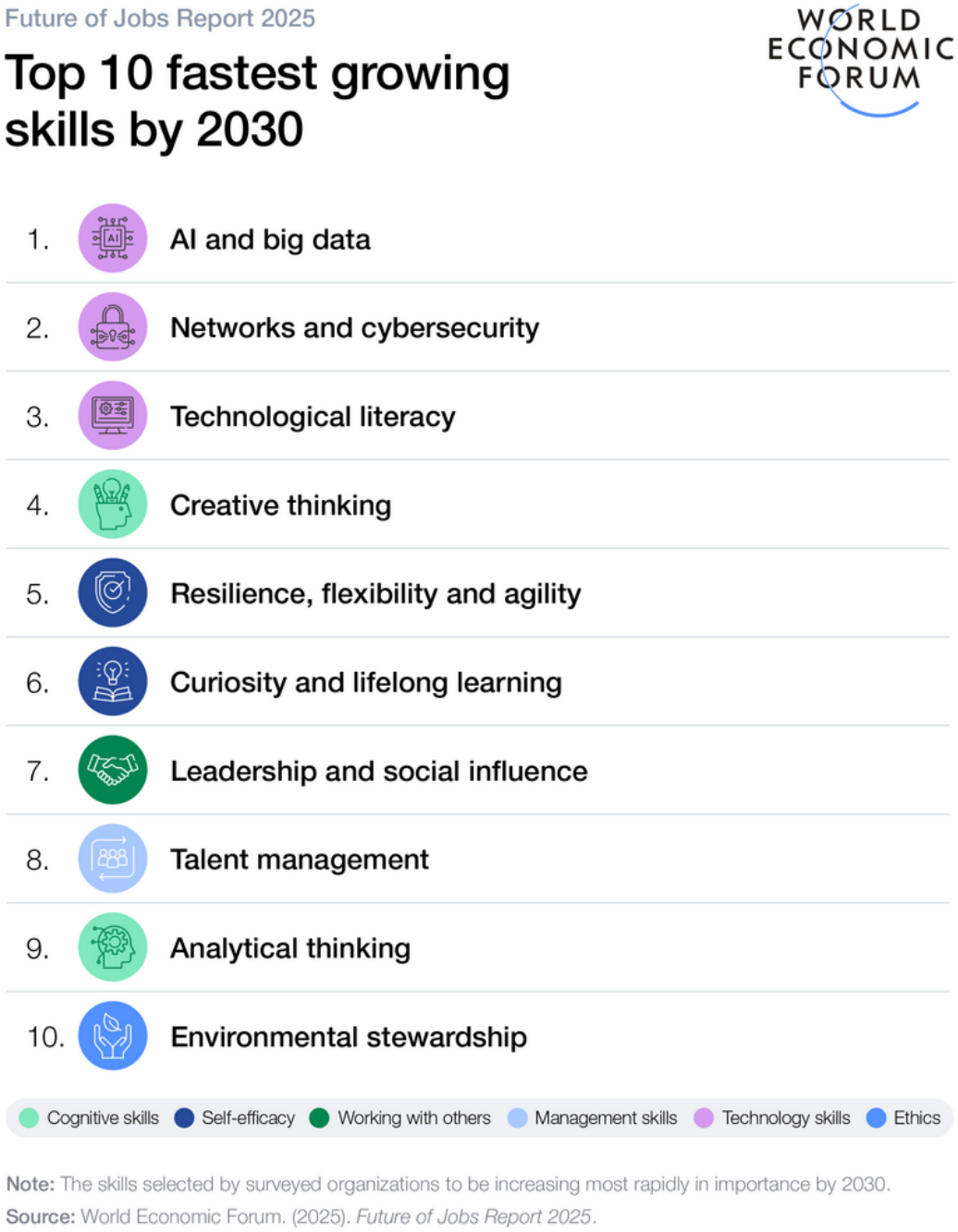
CHENG, Audrey et al. Barbarians at the gate: How ai is upending systems research. arXiv preprint arXiv:2510.06189, 2025.

Perspectivas Futuras

"In the next five years, 170 million jobs are projected to be created and 92 million jobs to be displaced (...)."

> Fonte: The Future of Jobs - Report 2025

Como se preparar para **posições de trabalho** e **problemas de pesquisa** que *ainda não existem*?



Prof. Dr. Denis M. L. Martins | martins.denis@usp.br

Palavras Finais

- **Industrialização:** A IA em produção exige sair da "IA Insular" para um ecossistema robusto e integrado.
- **Co-design**:
  - Implementação de treinamento e/ou inferência de modelos **in-database**.
  - Gerenciamento e otimização de prompts próximo aos dados que eles utilizam.
  - Utilizar o potencial do DBMS ao favor da IA.

Fonte da imagem: Gery Wibowo via Unsplash.

Prof. Dr. Denis M. L. Martins | martins.denis@usp.br