

I Conferência IA com Direitos Sociais_2025

Big Techs e a Nova Divisão Internacional do Trabalho

Sérgio Amadeu da Silveira
Prof. UFABC – pesq. CNPq

**A IA REALMENTE EXISTENTE
não é um ser, muito menos um software
ou um aplicativo instalado no seu
celular ou computador.**



É composta por **gigantescas infraestruturas** que reúnem **base de dados imensas**, computadores, roteadores, placas, processadores, cabos, conexões, instalações descomunais **e muito trabalho humano.**



A IA realmente existente
é uma **MEGAMÁQUINA**



**A megamáquina não se limita
a tecnologias físicas, mas abrange a própria
organização do poder e do trabalho.**

Como afirmou Matteo Pasquinelli e Vladan Joler existe uma **linha de montagem** do Aprendizado de Máquina, a abordagem da IA hegemônica hoje:

Dados → Algoritmos → Modelos

Dados → Algoritmos → Modelos

Poder Computacional



O aprendizado de máquina profundo é um subconjunto do aprendizado de máquina que utiliza **redes neurais artificiais com **múltiplas camadas** para **identificar padrões em dados**.**

Os grandes modelos de linguagem, LLMs, como o GPT, Gemini, Claude, DeepSeek são um tipo específico de deep learning voltado ao processamento e geração de linguagem natural.

Um **LLM** que atende **1 bilhão de acessos diários** precisa de aproximadamente **50.000 GPUs H100**.

São dezenas de milhares de
GPUs só para inferência.



Um Chat (LLM) precisa ser dimensionado para o **pico de carga** — quando milhões de usuários **usam o serviço simultaneamente** — o que pode exigir o **dobro ou o triplo da capacidade média**.

O **hardware extra** é necessário para garantir uma **baixa latência (tempo de resposta rápido)**, evitando que os usuários esperem mais de alguns segundos por uma resposta do modelo.

ESCALABILIDADE
ALTA DISPONIBILIDADE



Os **DATA CENTERS** não são construídos apenas para a **inferência** (o uso do modelo), eles são primariamente necessários para o **treinamento dos modelos** de próxima geração.



Os **DADOS** são o insumo
da IA realmente existente.

Treinar um modelo de ponta (como GPT-5 ou Llama 4) é muito **mais intensivo em computação** do que apenas rodá-lo (inferência).

O treinamento de um LLM de grande escala pode exigir **centenas de milhares de GPUs rodando por meses.**



O Data Center da xAI no Texas consome
cerca de **250 MW para alimentar 200 mil GPUs.**

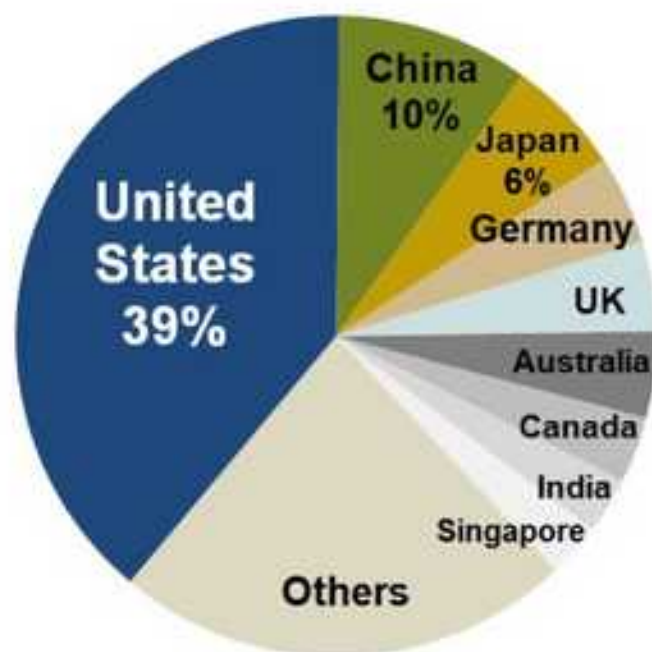
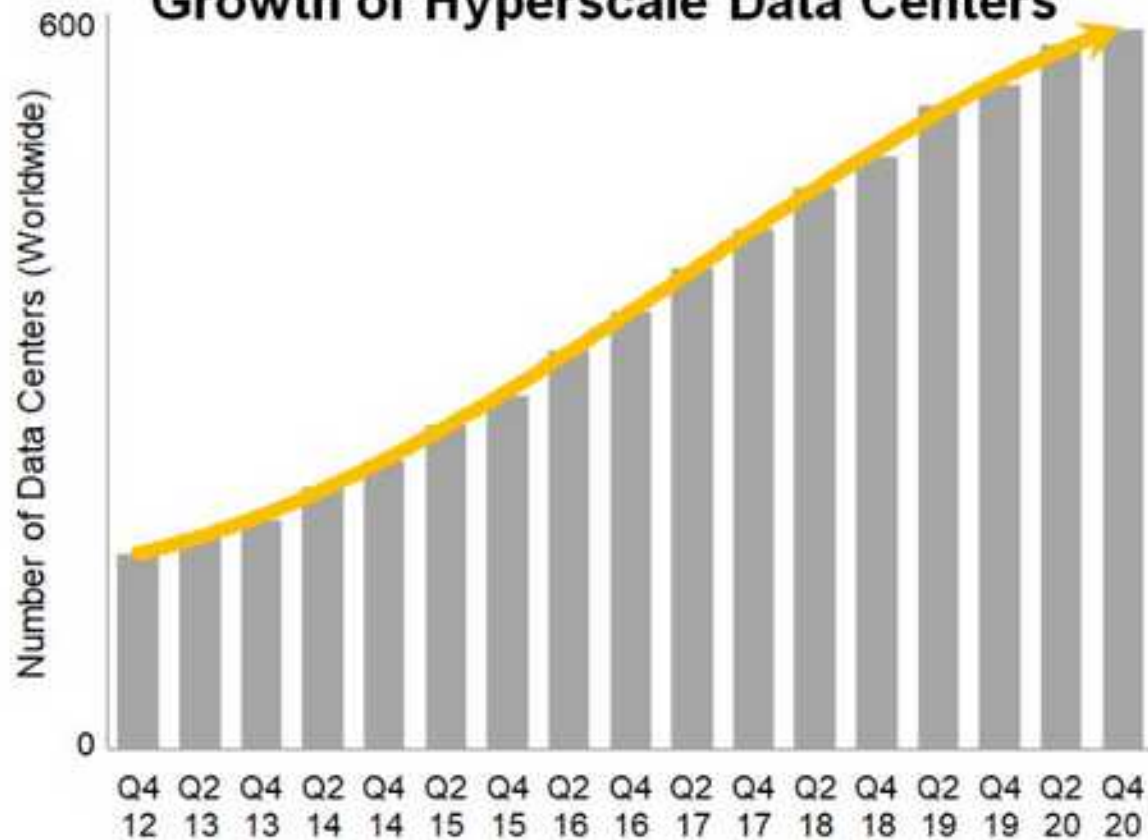
O Data Center da OpenAi no Texas possui **210 subestações de resfriamento a ar** e uma **enorme subestação principal** construída sob medida para sustentar a operação.



1 de 2 | O presidente Donald Trump discursa na Sala Roosevelt, na Casa Branca, terça-feira, 21 de janeiro de 2025, em Washington, enquanto o CEO da Softbank, Masayoshi Son, o diretor de tecnologia da Oracle, Larry Ellison, e o CEO da OpenAI, Sam Altman, ouvem. (Foto AP/Julia Demaree Nikhinson)

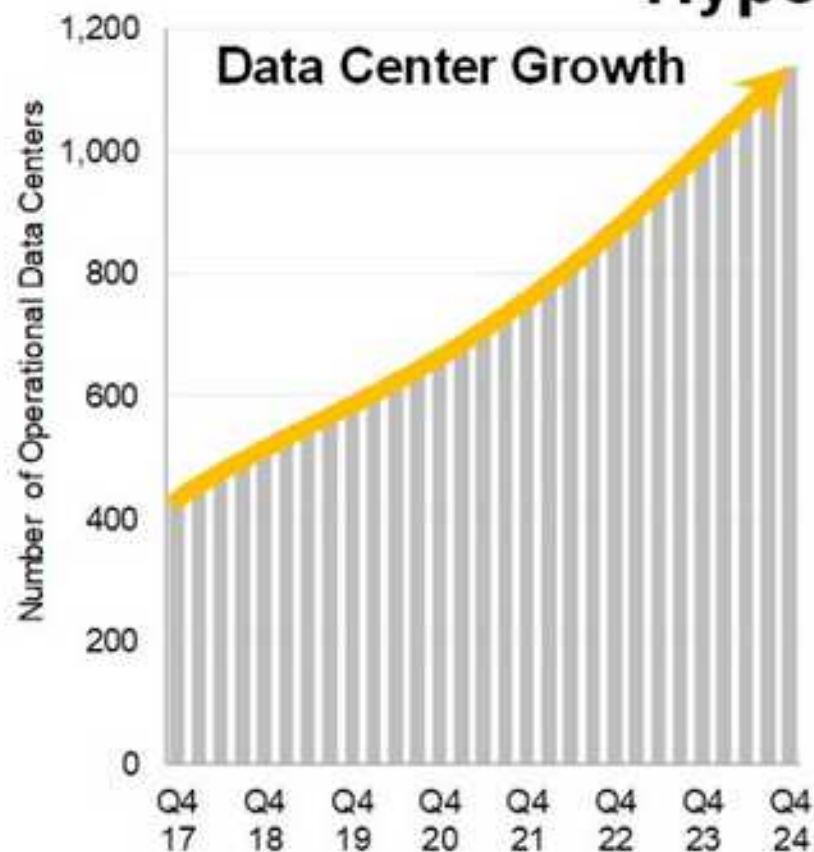
Para manter a liderança em IA, as Big Techs precisam ter **capacidade massiva e ociosa** disponível para o próximo **ciclo de desenvolvimento** do “**próximo grande modelo**”.

Growth of Hyperscale Data Centers



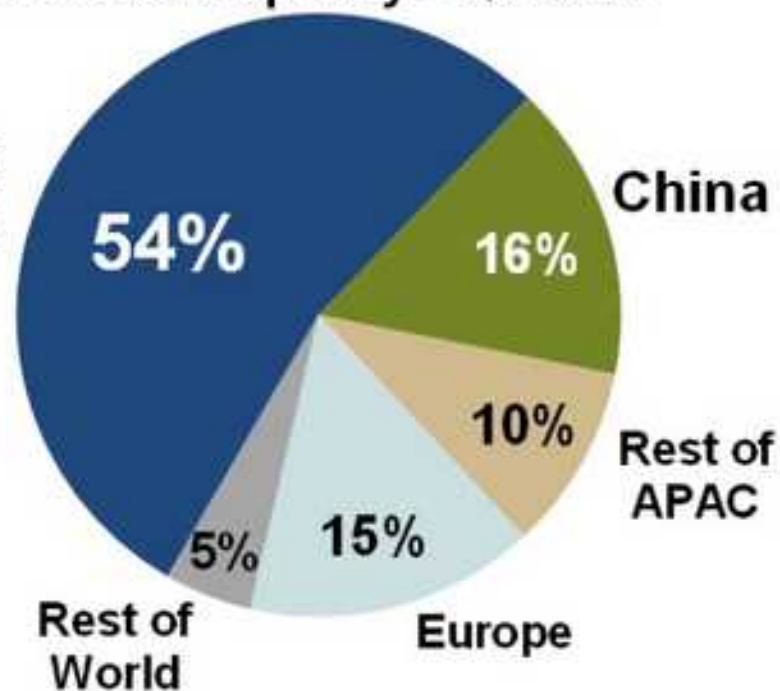
Source: Synergy Research Group

Hyperscale Data Centers



Data Center Capacity - Q4 2024

United States



Source: Synergy Research Group

Na IA realmente existente cada **parâmetro** (ou peso) representa **um valor ajustável que o modelo aprende e usa para armazenar padrões** e informações durante o treinamento.

Um modelo de **1 trilhão de parâmetros** possui maior **capacidade de armazenamento e conhecimento** que um modelo com 70 bilhões de parâmetros.

A **complexidade de um modelo** é determinada principalmente **pela sua arquitetura** e pelo **número de parâmetros**, mas é também delimitada pela **quantidade e qualidade dos dados de treinamento** e pelos objetivos de aprendizado utilizados durante o seu treinamento.

O tempo e os recursos (horas-GPU/TPU)
necessários para treinar um **modelo de 1 trilhão de**
parâmetros exigem **infraestruturas de**
supercomputação massivas
para serem concluídos.

Mesmo na fase de uso (inferência),
o modelo de 1 trilhão de parâmetros exige muito **mais poder de processamento e memória (VRAM)** para gerar uma resposta, tornando-o inerentemente mais complexo de implantar e operar.

Models	Publication Venue	License Type	Model Creators	Purpose	No. of Params	Commercial Use	Steps Trained	Data/ Tokens	Data Cleaning	No. of Processing Units	Processing Unit Type	Training Time
T5 [10]	JMLR'20	Apache-2.0	Google	General	11B	✓	1M	1T	Heur+Dedup	1024	TPU v3	-
GPT-3 [6]	NeurIPS'20	-	OpenAI	General	175B	×	-	300B	Dedup+QF	-	V100	-
mT5 [41]	NAACL'21	Apache-2.0	Google	General	13B	✓	1M	1T	-	-	-	-
PanGu- α [103]	arXiv'21	Apache-2.0	Huawei	General	200B	✓	260k	1.1TB	Heur+Dedup	2048	Ascend 910	-
CPM-2 [12]	AI Open'21	MIT	Tsinghua	General	198B	✓	1M	2.6TB	Dedup	-	-	-
Codex [130]	arXiv'21	-	OpenAI	Coding	12B	×	-	100B	Heur	-	-	-
ERNIE 3.0 [105]	arXiv'21	-	Baidu	General	10B	×	120k*	375B	Heur+Dedup	384	V100	-
Jurassic-1 [107]	White-Paper'21	Apache-2.0	AI21	General	178B	✓	-	300B	-	800	GPU	-
HyperCLOVA [109]	EMNLP'21	-	Naver	General	82B	×	-	300B	Clf+Dedup+PF	1024	A100	321h
Yuan 1.0 [110]	arXiv'21	Apache-2.0	-	General	245B	✓	26k*	180B	Heur+Clf+Dedup	2128	GPU	-
Gopher [113]	arXiv'21	-	Google	General	280B	×	-	300B	QF+Dedup	4096	TPU v3	920h
ERNIE 3.0 Titan [35]	arXiv'21	-	Baidu	General	260B	×	-	300B	Heur+Dedup	-	Ascend 910	-
GPT-NeoX-20B [115]	BigScience'22	Apache-2.0	EleutherAI	General	20B	✓	150k	825GB	None	96	40G A100	-
OPT [14]	arXiv'22	MIT	Meta	General	175B	✓	150k	180B	Dedup	992	80G A100	-
BLOOM [13]	arXiv'22	RAIL-1.0	BigScience	General	176B	✓	-	366B	Dedup+PR	384	80G A100	2520h
Galactica [137]	arXiv'22	Apache-2.0	Meta	Science	120B	×	225k	106B	Dedup	128	80GB A100	-
GLaM [116]	ICML'22	-	Google	General	1.2T	×	600k*	600B	Clf	1024	TPU v4	-
LaMDA [139]	arXiv'22	-	Google	Dialog	137B	×	3M	2.81T	Filtered	1024	TPU v3	1384h
MT-NLG [112]	arXiv'22	Apache-v2.0	MS.+Nvidia	General	530B	×	-	270B	-	4480	80G A100	-
AlphaCode [131]	Science'22	Apache-v2.0	Google	Coding	41B	✓	205k	967B	Heur+Dedup	-	TPU v4	-
Chinchilla [119]	arXiv'22	-	Google	General	70B	×	-	1.4T	QF+Dedup	-	TPUv4	-
PaLM [15]	arXiv'22	-	Google	General	540B	×	255k	780B	Heur	6144	TPU v4	-
AlexaTM [128]	arXiv'22	Apache v2.0	Amazon	General	20B	×	500k	1.1T	Filtered	128	A100	2880h
U-PaLM [122]	arXiv'22	-	Google	General	540B	×	20k	-	-	512	TPU v4	120h
UL2 [123]	ICLR'23	Apache-2.0	Google	General	20B	✓	2M	1T	-	512	TPU v4	-
GLM [33]	ICLR'23	Apache-2.0	Multiple	General	130B	×	-	400B	-	768	40G A100	1440h
CodeGen [129]	ICLR'23	Apache-2.0	Salesforce	Coding	16B	✓	650k	577B	Heur+Dedup	-	TPU v4	-
LLaMA [125]	arXiv'23	-	Meta	General	65B	×	350k	1.4T	Clf+Heur+Dedup	2048	80G A100	504h
PanGu Σ [128]	arXiv'23	-	Huawei	General	1.085T	×	-	329B	-	512	Ascend 910	2400h
BloombergGPT [140]	arXiv'23	-	Bloomberg	Finance	50B	×	139k	569B	Dedup	512	40G A100	1272h
Xuan Yuan 2.0 [143]	arXiv'23	RAIL-1.0	Du Xiaoman	Finance	176B	✓	-	366B	Filtered	80GB	A100	-
CodeT5+ [38]	arXiv'23	BSD-3	Salesforce	Coding	16B	✓	110k	51.5B	Dedup	16	40G A100	-
StarCoder [136]	arXiv'23	OpenRAIL-M	BigCode	Coding	15.5B	✓	250k	1T	Dedup+QF+PF	512	80G A100	624h
LLaMA-2 [21]	arXiv'23	LLaMA-2.0	Meta	General	70B	✓	500k	2T	Minimal Filtering	-	80G A100	1.7Mh
PaLM-2 [120]	arXiv'23	-	Google	General	-	×	-	-	Ddedup+PF+QF	-	-	-

Modelo (ano)	Empresa	Tempo/ processador	Custo treinamento (US\$)
GPT-4 (2023)	OpenAI	NVIDIA A100 sem informação de tempo	\$ 100 milhões (estimativa)
Gemini Ultra (2023)	Google	TPUs com milhares de chips sem informação	\$ 30 milhões (estimativa)
Falcon 180B (2023)	TII (EAU)	4.096 NVIDIA A100 sem tempo estimado	\$ 41 milhões (estimativa)
DBRX (2024)	Databricks	3 meses com 3.000 NVIDIA H100	\$ 10 milhões (reportado)
DeepSeek - V3 (2024)	DeepSeek	2 meses com 2.048 GPUs (relatado)	\$ 5,6 milhões (apenas computação)

Modelos Maiores = Mais Capacidade

Modelos com bilhões ou trilhões de parâmetros, como GPT-4 ou LLaMA 3 70B, têm uma capacidade muito maior de **armazenar conhecimento factual, entender nuances e executar raciocínio complexo** do que modelos menores.

Number of Significant Machine Learning Systems by Sector, 2002–22

Source: Epoch, 2022 | Chart: 2023 AI Index Report



The privatization of AI research

The results underscore another growing problem in AI, too: the sheer intensity of resources now required to produce paper-worthy results has made it increasingly challenging for people working in academia to continue contributing to research.

GCDs	Número de <i>startups</i> de IA adquiridas desde o lançamento do Chat GPT	Número total de <i>startups</i> com pelo menos 5% sob controle	Número <i>startups</i> estrangeiras com pelo menos 5% sob controle e % do total de aquisições da GCD
Google	136	2445	1369 (56%)
Microsoft	50	823	600 (73%)
Amazon	39	268	110 (41%)
Meta		49	30 (61%)
Apple		16	2 (12%)

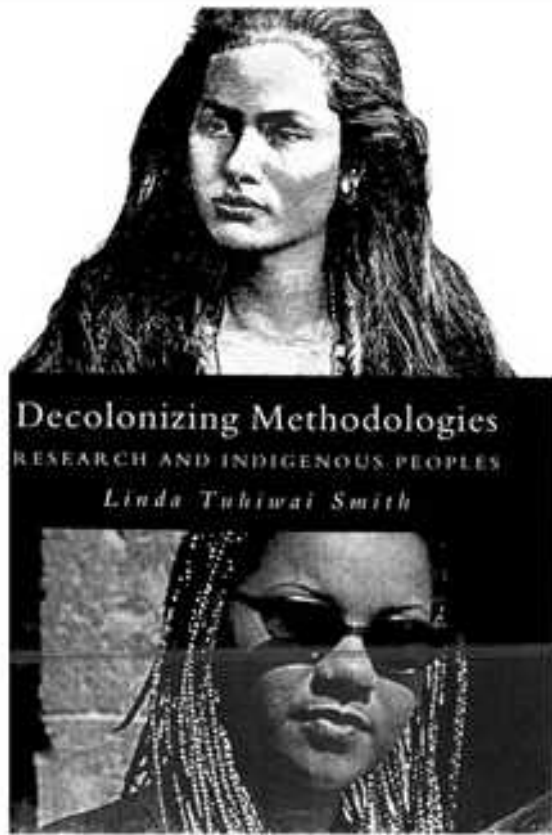
LASTRES, Helena Maria Martins; CASSIOLATO, José Eduardo; DANTAS, Marcos. Economia política da dados e soberania digital. Avaré, SP: Editora Contracorrente; Rio de Janeiro: Centro Internacional Celso Furtado, 2025.

Nosso papel na atual
Divisão Internacional do Trabalho tem sido de:

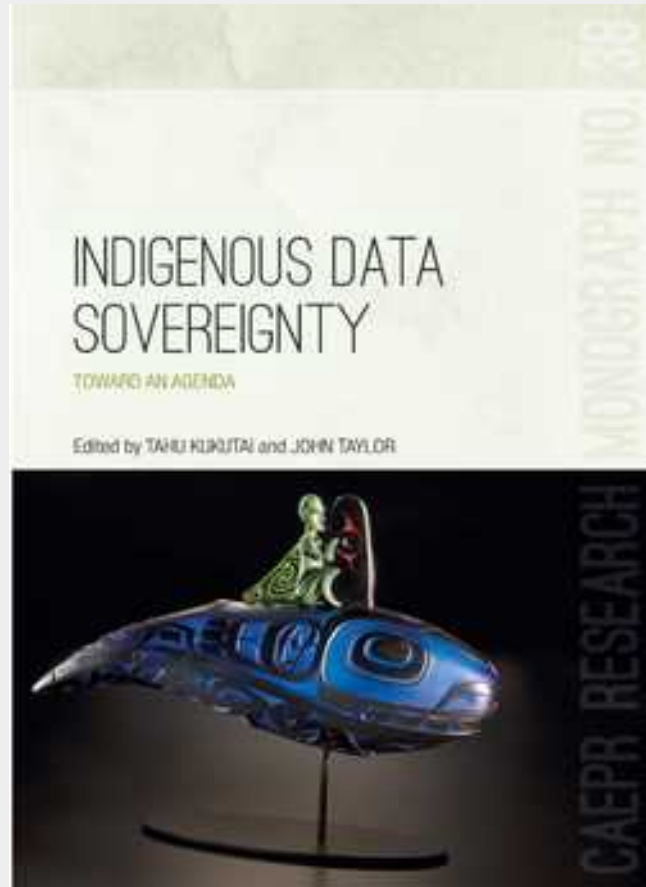
- 1- **Entregar os dados** da nossa população.
- 2- **Adquirir produtos e serviços** construídos, a partir dos nossos dados.
- 3- **Criar aplicativos** nos modelos pré-treinados das Big Techs, integrando seu **ecossistema**.



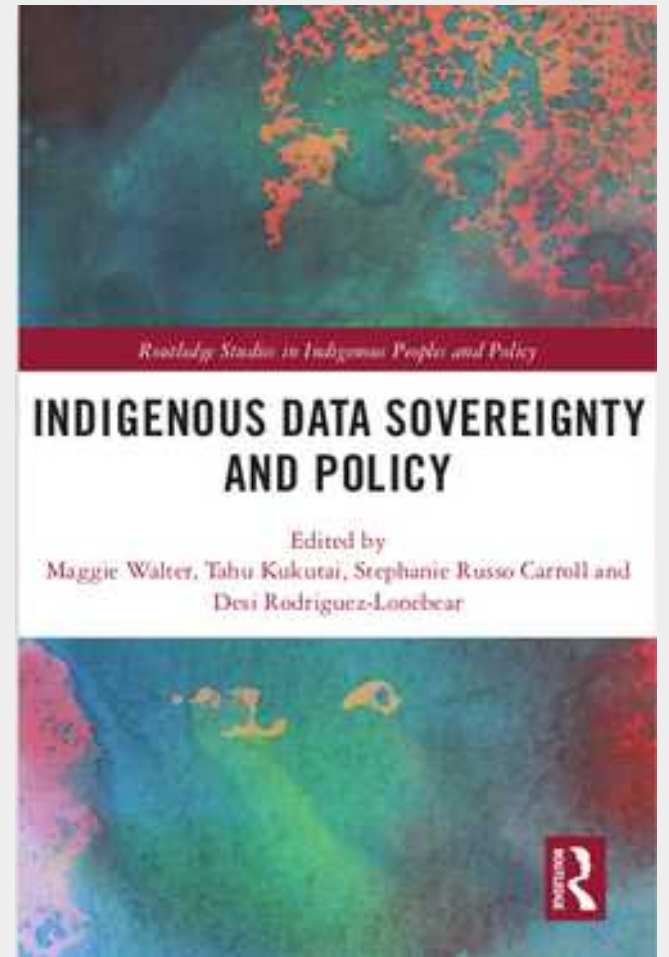
O termo "soberania alimentar" foi cunhado pela primeira vez em **1996** por membros da **Via Campesina**, uma organização internacional de agricultores, e posteriormente adotado por várias organizações internacionais, incluindo o **Banco Mundial** e as **Nações Unidas**.



1999



2016



2021

Soberania Digital Soberania de dados

71,53%

das universidades brasileiras
utilizam os e-mails do Google

7,64%

utilizam e-mails da Microsoft

(Pesquisa EDUCAÇÃO VIGIADA)

[Tempo de Aprender](#)[Educação Conectada](#)[Diploma Digital](#)[Conta pra Mim](#)[ID Estudantil](#)[Novos Caminhos](#)[Escola Cívico-Militar](#)[Caderno PNA](#)[Future-se](#)[Prouni](#)[Fies](#)[Sisu](#)[Educação em Prática](#)[Gabinete do Ministro](#)**ACESSO À
INFORMAÇÃO****TECNOLOGIA**

Microsoft destaca Sisu em nuvem como case de sucesso

Segunda-feira, 23 de março de 2020, 14h38

[Twitter](#)[Compartilhar](#)**Multinacional de tecnologia destina espaço em seu site para programa do MEC***Dyelle Menezes e Guilherme Pera, do Portal MEC*

A Microsoft destacou o Sistema de Seleção Unificada (Sisu) do primeiro semestre de 2020 como um case de sucesso. O Ministério da Educação (MEC) migrou o programa para a nuvem da multinacional de tecnologia para aumentar a capacidade de acessos.

Segundo a empresa, a mudança "garantiu a estabilidade e escalabilidade na capacidade de processamento para atender à demanda crescente de estudantes". Na primeira edição da seleção, em 2010, 83.125 vagas foram ofertadas pelo sistema, ante 237.128 na 1ª/2020.

A multinacional pontou ainda que "o MEC levou em consideração a necessidade de investimento em infraestrutura para comportar um sistema que recebe um enorme volume de acessos em apenas 12 dias por ano". No primeiro semestre do ano, foram 3.458.358 inscrições, feitas por 1.795.211 pessoas, já que cada candidato pode optar por

Lutar pelos nossos dados!

Construir nossas **infraestruturas** de
processamento **públicas e soberanas!**

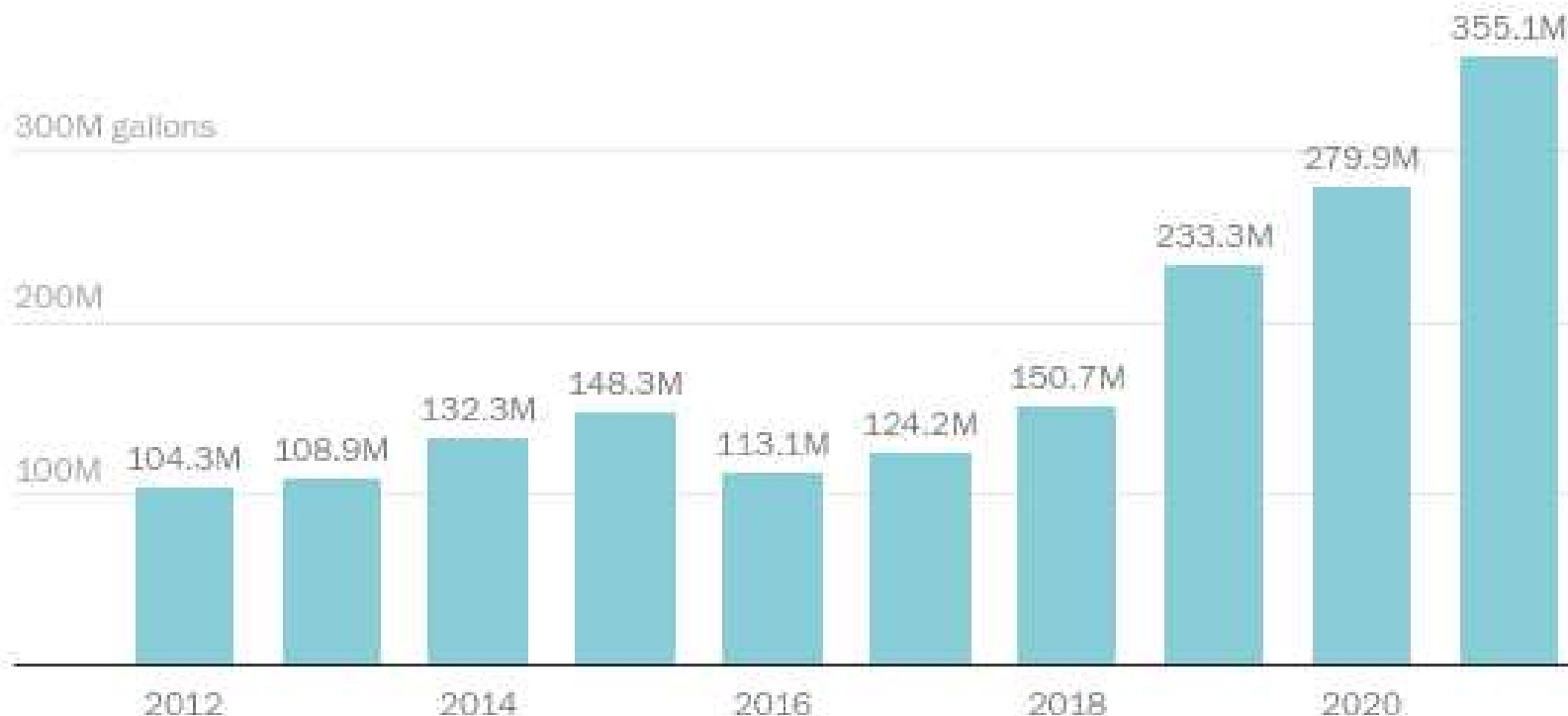
Criar políticas científicas para desenvolver
modelos ambientalmente sustentáveis.





Google's annual water use in The Dalles, in gallons

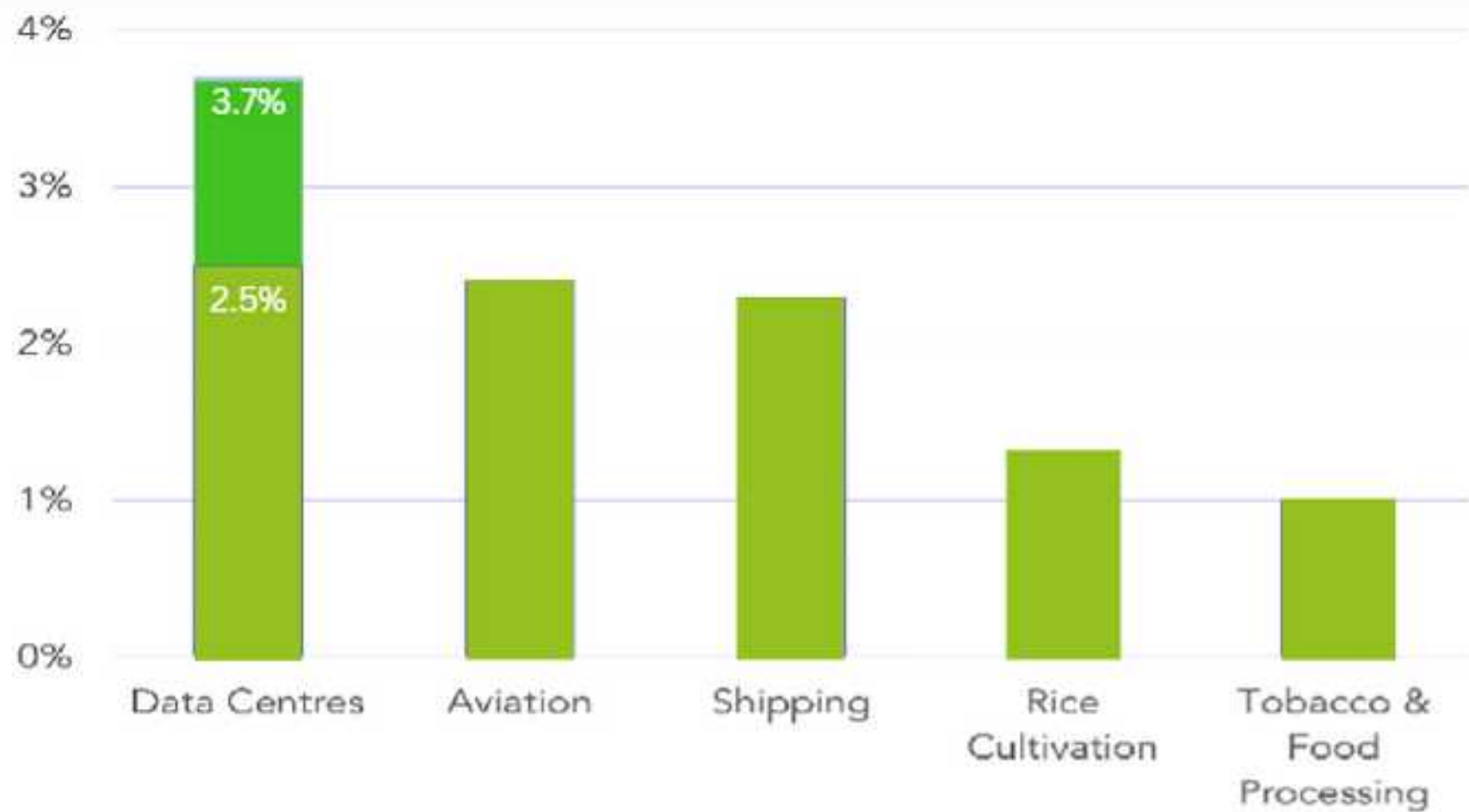
Google's data centers in The Dalles use nearly three times more water than they did five years ago and now account for more than a quarter of all the city's water use.



Source: City of The Dalles • [Get the data](#)

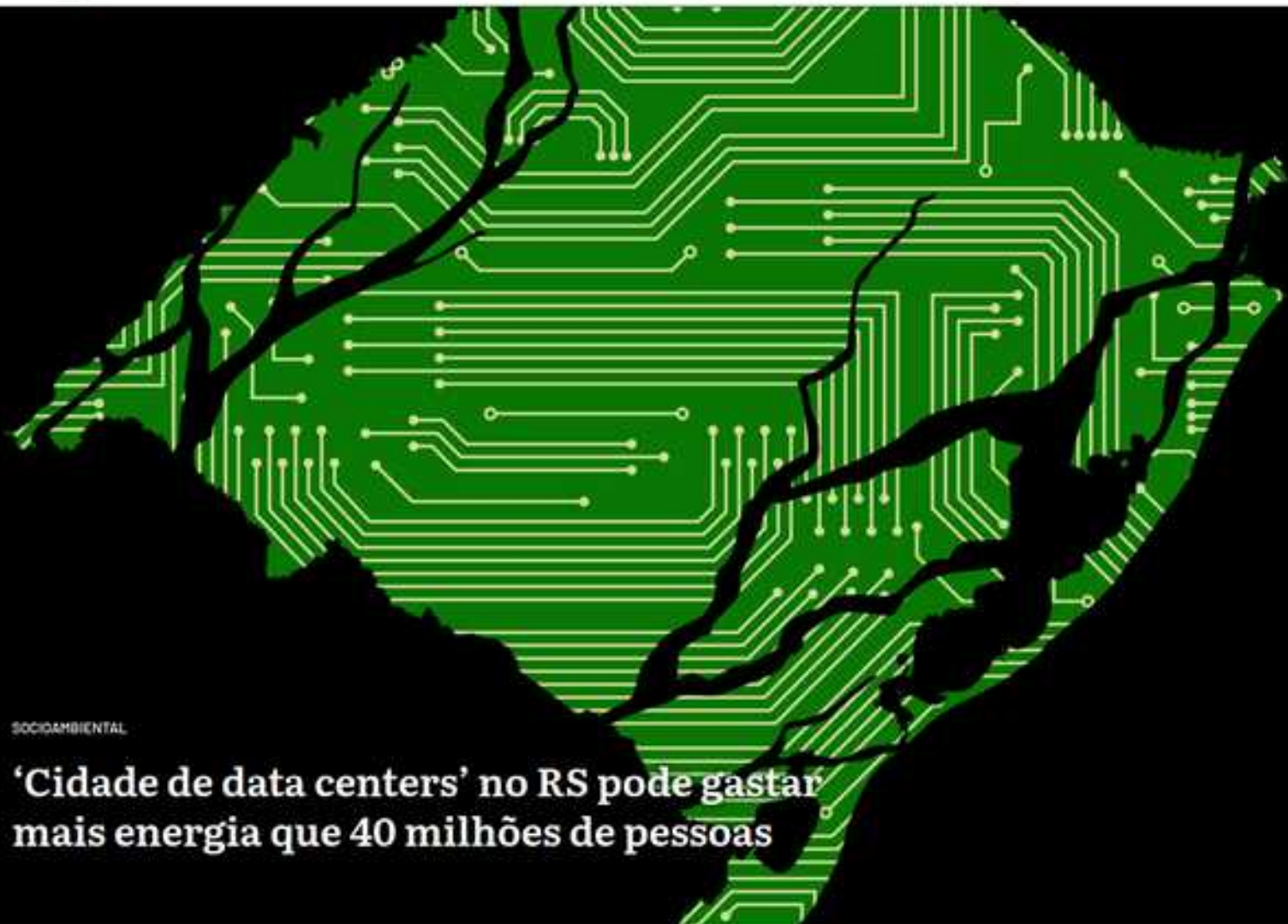


OREGONLIVE
The Oregonian



Os data centers são responsáveis por 2,5% a 3,7% das emissões globais de GEE.





SOCIOAMBIENTAL

**‘Cidade de data centers’ no RS pode gastar
mais energia que 40 milhões de pessoas**



Evento com presença do governador gaúcho Eduardo Leite (PSDB) marca assinatura de protocolo de intenções para as obras do complexo da Scala Data Centers (Foto: Maurício Tonetto/Secom)

O Scala AI City, como está sendo chamado o empreendimento em Eldorado do Sul, é superlativo: o investimento inicial é de R\$ 3 bilhões, podendo chegar a um total de R\$ 500 bilhões, segundo dados divulgados pelo governo do estado. A título de comparação, o maior investimento privado da história do Rio Grande do Sul alcançou R\$ 24 bilhões: a ampliação de uma fábrica de celulose da chilena CMPC, em 2024.



DISCOVER WHAT IEEE MEMBER
DISCOUNTS CAN OFFER YOU

EXCLUSIVE
DISCOUNT

NEWS | **ENERGY**

Big Tech Backs Small Nuclear › Google and Amazon invest in small modular reactors to power data centers

BY [EMILY WALTON](#) | 18 OCT 2024 | 4 MIN READ | [\[i\]](#)

Emily Walton is the power and energy editor at IEEE Spectrum.



A Irlanda queria construir data centers para o boom da IA. Agora, eles temem apagões

A Irlanda atrai empresas globais de tecnologia desde o "Tigre Celta"

Matt O'Brien • Quinta-feira, 19 de dezembro de 2024, 13h25 GMT



Um faxineiro esfrega o chão do saguão de recepção de um prédio vazio de data centers, em Dublin, Irlanda. (Copyright 2024 The Associated Press. Todos os direitos reservados)

EirGrid warning over 'mass exodus' of data centres from Ireland

Electricity network operator fears that without firm plans for new investments, some large data centres will exit the

POLITICO

[EU-US relations](#) [War in Ukraine](#) [Newsletters](#) [Podcasts](#) [Poll of Polls](#) [Policy news](#) [Events](#)

NEWS > TECHNOLOGY

Big Tech's AI dream turns into Ireland's energy nightmare

Energy-hungry artificial intelligence is stressing Ireland's power grids, jeopardizing the country's position as Europe's tech hub.



Warning of pressures on power supply since as early as 2016. Picture: File

Lutar por direitos diante dos sistemas automatizados.

Lutar pelos dados, pela **participação sindical** nos projetos de IA, reivindicar **ganhos de produtividade**, exigir **relatórios de impactos da IA nas categorias**, exigir **relatórios de impactos socioambientais** para as infraestruturas de IA.

Obrigado.

sergio.amadeu@ufabc.edu.br

