REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

# APRENDIZADO AUTO-SUPERVISIONADO PARA EXPLORAÇÃO DE DADOS NÃO ESTRUTURADOS: NOVAS FRONTEIRAS DA INTELIGÊNCIA ARTIFICIAL ESCALÁVEL

## SELF-SUPERVISED LEARNING FOR UNSTRUCTURED DATA EXPLORATION: NEW FRONTIERS OF SCALABLE ARTIFICIAL INTELLIGENCE

Ciências Exatas e da Terra, Engenharias • 14/07/2026

REGISTRO DOI: 10.70773/revistatopicos/783480421

Rômulo Ferreira dos Santos[1]

**RESUMO**

O crescimento exponencial da produção de dados digitais tem ampliado a necessidade de métodos capazes de extrair conhecimento relevante sem depender de extensos processos de rotulação manual. Textos, imagens, vídeos, áudios, sinais digitais, registros de sensores, fluxos de navegação, documentos corporativos, dados biomédicos, registros geoespaciais e interações multimodais constituem grandes volumes de dados não estruturados ou semi-estruturados, cuja exploração por modelos supervisionados tradicionais exige tempo, custo, especialistas e infraestrutura de anotação. Nesse contexto, o aprendizado auto-supervisionado emerge como paradigma central da inteligência artificial escalável, ao permitir que modelos aprendam representações úteis a partir dos próprios dados, por meio de tarefas pretextuais, mascaramento, reconstrução, previsão temporal, contraste entre amostras, alinhamento multimodal e destilação sem rótulos humanos explícitos. Este artigo explora o potencial do aprendizado auto-supervisionado para análise de grandes volumes de dados não estruturados em diferentes domínios computacionais, com ênfase em linguagem natural, visão computacional, fala, sinais digitais, séries temporais, dados multimodais e sistemas fundacionais. A pesquisa adota abordagem qualitativa, exploratória, bibliográfica e propositiva, articulando literatura sobre aprendizado profundo, representação de dados, modelos Transformer, contrastive learning, masked modeling, foundation models, data-centric AI, eficiência computacional e ética algorítmica. A análise evidencia que técnicas auto-supervisionadas podem reduzir dependência de dados rotulados, melhorar generalização, favorecer transferência para tarefas downstream, ampliar desempenho em cenários de baixa rotulação e sustentar a construção de modelos fundacionais. Propõe-se o Modelo Integrado de Exploração Auto-Supervisionada

de Dados Não Estruturados, denominado MIEAS-DNE, estruturado em camadas de ingestão, pré-processamento, geração de tarefas pretextuais, aprendizagem de representações, avaliação, adaptação e governança. Conclui-se que o aprendizado auto-supervisionado representa uma das principais fronteiras da IA escalável, mas seu avanço exige atenção a custo computacional, viés, explicabilidade, privacidade, qualidade dos dados e uso responsável.
**Palavras-chave:** aprendizado auto-supervisionado; dados não estruturados; inteligência artificial; representação de dados; machine learning; modelos fundacionais; aprendizagem contrastiva; modelagem mascarada.

**ABSTRACT**

The exponential growth of digital data production has increased the need for methods capable of extracting relevant knowledge without relying on extensive manual labeling processes. Texts, images, videos, audio, digital signals, sensor records, browsing flows, corporate documents, biomedical data, geospatial records, and multimodal interactions constitute large volumes of unstructured or semi-structured data whose exploration by traditional supervised models requires time, cost, experts, and annotation infrastructure. In this context, self-supervised learning emerges as a central paradigm of scalable artificial intelligence by allowing models to learn useful representations from the data itself through pretext tasks, masking, reconstruction, temporal prediction, contrast between samples, multimodal alignment, and distillation without explicit human labels. This article explores the potential of self-supervised learning for analyzing large volumes of unstructured data across different computational domains, with emphasis on natural language, computer vision, speech, digital signals, time series, multimodal data, and foundation systems. The research adopts a qualitative,

exploratory, bibliographic, and propositional approach, articulating literature on deep learning, representation learning, Transformer models, contrastive learning, masked modeling, foundation models, data-centric AI, computational efficiency, and algorithmic ethics. The analysis shows that self-supervised techniques can reduce dependence on labeled data, improve generalization, support transfer to downstream tasks, enhance performance in low-label scenarios, and sustain the construction of foundation models. The article proposes the Integrated Model for Self-Supervised Exploration of Unstructured Data, named MIEAS-DNE, structured in layers of ingestion, preprocessing, pretext task generation, representation learning, evaluation, adaptation, and governance. It concludes that self-supervised learning represents one of the main frontiers of scalable AI, but its development requires attention to computational cost, bias, explainability, privacy, data quality, and responsible use.
**Keywords:** self-supervised learning; unstructured data; artificial intelligence; data representation; machine learning; foundation models; contrastive learning; masked modeling.

## 1. INTRODUÇÃO

A expansão da produção de dados digitais tornou-se uma das marcas estruturantes da sociedade informacional contemporânea. Organizações públicas, empresas privadas, instituições científicas, plataformas digitais, sistemas de saúde, redes sociais, sensores urbanos, dispositivos móveis, aplicações industriais e ambientes educacionais geram continuamente volumes massivos de dados, frequentemente em formatos não estruturados, como textos, imagens, vídeos, áudios, sinais, logs, documentos, mensagens,

registros de navegação e fluxos multimodais (GOODFELLOW; BENGIO; COURVILLE, 2016; KITCHIN, 2014).

A inteligência artificial tradicional, especialmente em sua vertente supervisionada, obteve avanços expressivos a partir de grandes bases rotuladas. Entretanto, a rotulação manual de dados é cara, lenta, sujeita a inconsistências, dependente de especialistas e, em muitos domínios, praticamente inviável em larga escala. Em áreas como medicina, direito, sensoriamento remoto, fala, sinais industriais e dados científicos, a anotação exige conhecimento especializado, validação rigorosa e critérios éticos, o que limita a expansão de modelos supervisionados puros (LECUN; BENGIO; HINTON, 2015; GOODFELLOW; BENGIO; COURVILLE, 2016).

O aprendizado auto-supervisionado surge como resposta a esse desafio. Em vez de depender exclusivamente de rótulos humanos, os modelos aprendem a partir da estrutura interna dos próprios dados. O sistema cria tarefas auxiliares, chamadas tarefas pretextuais, nas quais parte da informação é ocultada, corrompida, reorganizada ou contrastada, e o modelo precisa reconstruir, prever, alinhar ou distinguir padrões. Dessa forma, os dados tornam-se fonte de supervisão indireta (DOERSCH; GUPTA; EFROS, 2015; DEVLIN et al., 2019; CHEN et al., 2020).

A ideia central é que dados não estruturados possuem regularidades latentes. Textos apresentam gramática, semântica, contexto e coocorrência; imagens apresentam bordas, formas, texturas, objetos e composição espacial; áudios apresentam ritmo, fonemas, timbre e padrões temporais; sinais digitais apresentam dependência sequencial, frequência, anomalias e regimes de operação; vídeos apresentam movimento, continuidade temporal e relação entre

objetos. O aprendizado auto-supervisionado busca capturar essas regularidades em representações vetoriais úteis para tarefas posteriores (BENGIO; COURVILLE; VINCENT, 2013; LECUN; MISRA, 2021).

A importância desse paradigma tornou-se evidente com o avanço dos modelos de linguagem. BERT demonstrou que modelos Transformer poderiam ser pré-treinados em grandes volumes de texto não rotulado por meio de tarefas como previsão de tokens mascarados, produzindo representações transferíveis para tarefas como classificação, resposta a perguntas, inferência textual e extração de entidades (DEVLIN et al., 2019). Posteriormente, modelos baseados em pré-treinamento autoregressivo, mascarado ou híbrido consolidaram a noção de que o conhecimento linguístico pode ser extraído em escala a partir de corpora massivos (RADFORD et al., 2019; BROWN et al., 2020).

Na visão computacional, métodos auto-supervisionados também avançaram rapidamente. SimCLR demonstrou que representações visuais de alta qualidade poderiam ser aprendidas por contraste entre diferentes aumentos da mesma imagem, sem rótulos humanos (CHEN et al., 2020). Métodos como MoCo, BYOL, DINO e MAE ampliaram essa fronteira, explorando bancos de memória, destilação sem rótulos, Transformers visuais e reconstrução de partes mascaradas da imagem (HE et al., 2020; GRILL et al., 2020; CARON et al., 2021; HE et al., 2022).

Na área de fala, wav2vec 2.0 demonstrou que representações robustas poderiam ser aprendidas a partir de áudio bruto não transcrito, com posterior adaptação a tarefas de reconhecimento automático de fala usando quantidade muito menor de dados

rotulados (BAEVSKI et al., 2020). Esse avanço é particularmente relevante para línguas de poucos recursos, dialetos, sotaques regionais e contextos nos quais transcrição manual é custosa (BAEVSKI et al., 2020; HSU et al., 2021).

O aprendizado auto-supervisionado também sustenta parte relevante da discussão sobre modelos fundacionais. Esses modelos são pré-treinados em larga escala e posteriormente adaptados a múltiplas tarefas por fine-tuning, prompting, few-shot learning ou adaptação eficiente de parâmetros. A capacidade de aprender representações gerais a partir de dados massivos não rotulados ou fracamente supervisionados tornou-se base para sistemas contemporâneos de linguagem, visão, fala e multimodalidade (BOMMASANI et al., 2021; RADFORD et al., 2021).

Diante desse cenário, este artigo parte do seguinte problema de pesquisa: **como o aprendizado auto-supervisionado pode ampliar a exploração de dados não estruturados e sustentar modelos de inteligência artificial escaláveis em cenários marcados por escassez de dados rotulados, heterogeneidade informacional e alta complexidade computacional?**

O objetivo geral é analisar o potencial do aprendizado auto-supervisionado para a exploração de dados não estruturados, discutindo seus fundamentos, métodos, aplicações, limitações e implicações para a inteligência artificial escalável. Como objetivos específicos, busca-se: diferenciar aprendizado auto-supervisionado de aprendizado supervisionado, não supervisionado e semi-supervisionado; examinar métodos contrastivos, generativos, mascarados, preditivos e multimodais; analisar aplicações em texto, imagem, fala, sinais e dados complexos; propor um modelo

integrado de exploração auto-supervisionada; e discutir desafios éticos, computacionais e científicos (BENGIO; COURVILLE; VINCENT, 2013; DEVLIN et al., 2019; CHEN et al., 2020).

Defende-se como tese central que o aprendizado auto-supervisionado constitui uma das principais fronteiras da inteligência artificial escalável, pois permite transformar dados brutos em representações generalizáveis, reduzindo dependência de rotulação manual e ampliando a capacidade de adaptação dos modelos a múltiplos domínios. Contudo, seu uso responsável exige governança de dados, avaliação robusta, controle de viés, eficiência computacional e atenção aos impactos sociais da automação (BOMMASANI et al., 2021; MITCHELL et al., 2019; SCHWARTZ et al., 2020).

## 2. METODOLOGIA

Este estudo adota abordagem qualitativa, exploratória, bibliográfica e propositiva. A escolha por esse delineamento justifica-se pela natureza emergente, interdisciplinar e altamente dinâmica do aprendizado auto-supervisionado, campo que articula aprendizado profundo, estatística, ciência de dados, processamento de linguagem natural, visão computacional, fala, sistemas multimodais, representação vetorial, modelos fundacionais e ética em inteligência artificial (CRESWELL, 2014; TORRACO, 2005).

A pesquisa foi estruturada como revisão narrativa integrativa com proposição de modelo conceitual. A revisão narrativa integrativa permite reunir contribuições de diferentes áreas técnicas, especialmente quando o tema se desenvolve rapidamente e combina literatura acadêmica, relatórios técnicos, benchmarks,

modelos abertos, pré-publicações e conferências de referência em inteligência artificial (WHITTEMORE; KNAFL, 2005; TORRACO, 2005).

A pergunta norteadora foi: **quais métodos, arquiteturas e estratégias de avaliação permitem utilizar aprendizado auto-supervisionado para extrair conhecimento de dados não estruturados em ambientes computacionais escaláveis?**

Foram considerados como eixos de análise: aprendizado auto-supervisionado; representação de dados; tarefas pretextuais; aprendizado contrastivo; masked language modeling; masked image modeling; autoencoders; Transformers; modelos fundacionais; aprendizagem multimodal; dados não estruturados; transferência de aprendizagem; fine-tuning; avaliação de representações; eficiência computacional; e governança ética da IA (BENGIO; COURVILLE; VINCENT, 2013; VASWANI et al., 2017; BOMMASANI et al., 2021).

Entre os autores e trabalhos utilizados, destacam-se Bengio, Courville e Vincent (2013), LeCun, Bengio e Hinton (2015), Vaswani et al. (2017), Devlin et al. (2019), Radford et al. (2019, 2021), Chen et al. (2020), He et al. (2020, 2022), Grill et al. (2020), Caron et al. (2021), Baevski et al. (2020, 2022), Brown et al. (2020), Bommasani et al. (2021), Dosovitskiy et al. (2021), Hinton e Salakhutdinov (2006), Kingma e Welling (2014), Vincent et al. (2008), Mikolov et al. (2013), Peters et al. (2018), Howard e Ruder (2018), além de estudos sobre avaliação, viés, sustentabilidade computacional e modelos responsáveis (MITCHELL et al., 2019; BENDER et al., 2021; SCHWARTZ et al., 2020).

Foram incluídos trabalhos que abordassem aprendizado auto-supervisionado, pré-treinamento, representação de dados, modelos de linguagem, visão computacional, fala, dados multimodais, autoencoders, contrastive learning, masked modeling, modelos fundacionais e aplicações em dados não estruturados. Foram excluídos materiais estritamente promocionais, textos sem autoria identificável, publicações sem relação com aprendizagem de representações e estudos centrados apenas em aprendizado supervisionado sem discussão de auto-supervisão.

A análise foi organizada em cinco etapas. A primeira etapa delimitou os conceitos de aprendizado supervisionado, não supervisionado, semi-supervisionado, auto-supervisionado e fracamente supervisionado. A segunda etapa examinou os principais mecanismos auto-supervisionados, como reconstrução, previsão, contraste, mascaramento, ordenação temporal, distilação e alinhamento multimodal. A terceira etapa analisou aplicações por modalidade de dados. A quarta etapa identificou desafios técnicos e éticos. A quinta etapa resultou na proposição do **MIEAS-DNE — Modelo Integrado de Exploração Auto-Supervisionada de Dados Não Estruturados.**

Por se tratar de estudo bibliográfico e propositivo, não houve coleta empírica direta. A proposta apresentada tem finalidade conceitual e metodológica, podendo orientar pesquisas aplicadas, desenvolvimento de sistemas inteligentes, projetos corporativos de análise de dados, aplicações científicas e estratégias de IA em contextos de baixa rotulação.

## 3. DADOS NÃO ESTRUTURADOS E O DESAFIO DA ROTULAÇÃO EM ESCALA

Dados não estruturados são aqueles que não se organizam naturalmente em tabelas rígidas, esquemas relacionais simples ou campos previamente definidos. Textos livres, áudios, imagens, vídeos, documentos digitalizados, logs, mensagens, sinais biomédicos, registros de sensores, arquivos jurídicos, prontuários, publicações em redes sociais e dados multimodais constituem exemplos típicos desse universo informacional (KITCHIN, 2014; GOODFELLOW; BENGIO; COURVILLE, 2016).

A dificuldade de exploração desses dados decorre de sua variedade, volume, ruído, ambiguidade e dependência contextual. Um texto pode conter ironia, polissemia, abreviações, erros, contexto cultural e relações implícitas. Uma imagem pode envolver oclusões, iluminação variável, múltiplos objetos e semântica visual complexa. Um áudio pode conter ruído, sotaques, sobreposição de falas e variação temporal. Um sinal industrial pode apresentar regimes normais e anômalos difíceis de distinguir (BENGIO; COURVILLE; VINCENT, 2013; LECUN; MISRA, 2021).

Modelos supervisionados tradicionais exigem rótulos explícitos. Para classificar imagens médicas, especialistas precisam marcar achados. Para treinar sistemas jurídicos, juristas precisam classificar documentos. Para reconhecimento de fala, transcritores precisam alinhar áudio e texto. Para detecção de anomalias industriais, engenheiros precisam identificar falhas. Em muitos casos, o custo de rotulação supera a capacidade operacional da organização (LECUN; BENGIO; HINTON, 2015; GOODFELLOW; BENGIO; COURVILLE, 2016).

A rotulação também pode ser inconsistente. Diferentes especialistas podem discordar sobre diagnóstico, relevância jurídica, emoção expressa, categoria semântica ou fronteira temporal de um evento. A

qualidade do modelo supervisionado depende da qualidade dos rótulos; quando estes são ruidosos, enviesados ou incompletos, o aprendizado sofre limitações (NORTHCUTT; JIANG; CHUANG, 2021; MITCHELL et al., 2019).

Além disso, muitos domínios apresentam escassez de dados rotulados por razões estruturais. Línguas indígenas, dialetos regionais, doenças raras, falhas industriais incomuns, eventos extremos, dados ambientais específicos e contextos locais possuem poucos exemplos anotados. O aprendizado auto-supervisionado torna-se particularmente relevante nesses cenários, pois permite aproveitar grandes volumes de dados brutos disponíveis (BAEVSKI et al., 2020; BOMMASANI et al., 2021).

O desafio central não é apenas classificar dados, mas aprender representações. Uma representação é uma forma computacional de organizar informação de modo que relações relevantes fiquem mais acessíveis ao modelo. Boas representações aproximam elementos semanticamente semelhantes, separam padrões distintos e capturam regularidades úteis para várias tarefas downstream (BENGIO; COURVILLE; VINCENT, 2013; MIKOLOV et al., 2013).

O aprendizado auto-supervisionado desloca o foco da tarefa específica para a representação geral. Em vez de treinar um modelo apenas para uma classificação estreita, busca-se pré-treinar um modelo em dados amplos e depois adaptá-lo a diferentes tarefas. Essa lógica explica parte do sucesso dos modelos fundacionais contemporâneos (DEVLIN et al., 2019; BROWN et al., 2020; BOMMASANI et al., 2021).

Portanto, a exploração de dados não estruturados exige métodos que consigam aprender a partir da abundância de dados brutos e da escassez de rótulos. O aprendizado auto-supervisionado oferece esse caminho ao transformar o próprio dado em fonte de supervisão (LECUN; MISRA, 2021; ZHANG et al., 2025).

## 4. FUNDAMENTOS DO APRENDIZADO AUTO-SUPERVISIONADO

O aprendizado auto-supervisionado é um paradigma no qual o modelo aprende a partir de sinais de supervisão gerados automaticamente dos próprios dados. Diferentemente do aprendizado supervisionado, que depende de rótulos humanos, o aprendizado auto-supervisionado cria tarefas internas, como prever uma parte ausente do dado, reconstruir uma entrada corrompida, identificar relações temporais, aproximar versões aumentadas da mesma amostra ou alinhar modalidades diferentes (DOERSCH; GUPTA; EFROS, 2015; CHEN et al., 2020; LECUN; MISRA, 2021).

A ideia de aprender representações sem rótulos explícitos não é inteiramente nova. Autoencoders, modelos generativos, embeddings distribuídos e métodos de redução de dimensionalidade já buscavam capturar estrutura latente dos dados. Hinton e Salakhutdinov (2006) mostraram a relevância de representações compactas, enquanto Vincent et al. (2008) desenvolveram denoising autoencoders capazes de reconstruir entradas corrompidas, antecipando princípios usados em métodos modernos de mascaramento e reconstrução (HINTON; SALAKHUTDINOV, 2006; VINCENT et al., 2008).

O que tornou o aprendizado auto-supervisionado contemporâneo especialmente poderoso foi a combinação entre grandes bases de

dados, arquiteturas profundas, Transformers, GPUs, otimização em larga escala e tarefas pretextuais eficazes. O Transformer, proposto por Vaswani et al. (2017), tornou-se arquitetura central para modelos de linguagem, visão e multimodalidade, devido à capacidade de modelar dependências longas por mecanismos de atenção (VASWANI et al., 2017; DOSOVITSKIY et al., 2021).

No processamento de linguagem natural, o aprendizado auto-supervisionado aparece em tarefas como previsão da próxima palavra, modelagem de linguagem mascarada e ordenação de sentenças. Word2Vec já havia demonstrado que relações semânticas poderiam emergir da previsão de contexto linguístico (MIKOLOV et al., 2013). BERT ampliou essa lógica ao mascarar tokens e treinar Transformers bidirecionais em grandes corpora, gerando representações transferíveis (DEVLIN et al., 2019).

Na visão computacional, tarefas pretextuais iniciais incluíam prever rotação, resolver quebra-cabeças de patches, colorir imagens e reconstruir entradas. Posteriormente, métodos contrastivos ganharam destaque ao aproximar representações de duas versões aumentadas da mesma imagem e afastar representações de imagens diferentes (DOERSCH; GUPTA; EFROS, 2015; GIDARIS; SINGH; KOMODAKIS, 2018; CHEN et al., 2020).

O aprendizado auto-supervisionado também se diferencia do aprendizado não supervisionado tradicional. Embora ambos usem dados sem rótulos humanos, o auto-supervisionado define objetivos de predição explícitos a partir dos próprios dados, enquanto o não supervisionado clássico frequentemente busca agrupamentos, redução de dimensionalidade ou estimação de densidade sem

tarefa pretextual claramente estruturada (BENGIO; COURVILLE; VINCENT, 2013; LECUN; MISRA, 2021).

A vantagem do aprendizado auto-supervisionado está em sua capacidade de produzir representações generalizáveis. Após pré-treinamento, o modelo pode ser adaptado a tarefas específicas com poucos rótulos. Essa lógica reduz custo de anotação, melhora desempenho em domínios de baixa rotulação e permite reutilização de modelos em múltiplos contextos (DEVLIN et al., 2019; BAEVSKI et al., 2020; HE et al., 2022).

Entretanto, a qualidade da representação depende da tarefa pretextual. Uma tarefa inadequada pode levar o modelo a aprender atalhos irrelevantes. Por exemplo, em imagens, aumentos mal escolhidos podem destruir informações importantes; em texto, mascaramento mal calibrado pode limitar contexto; em sinais, janelas temporais inadequadas podem ocultar padrões críticos. A construção da tarefa é, portanto, elemento epistemológico central do método (CHEN et al., 2020; BENGIO; COURVILLE; VINCENT, 2013).

## 5. APRENDIZADO CONTRASTIVO E REPRESENTAÇÕES DISCRIMINATIVAS

O aprendizado contrastivo é uma das abordagens mais influentes do aprendizado auto-supervisionado. Seu princípio é aproximar representações de amostras semanticamente relacionadas e afastar representações de amostras diferentes. Em visão computacional, isso geralmente ocorre por meio de diferentes aumentos da mesma imagem, como cortes, alterações de cor, desfoque e transformações geométricas (CHEN et al., 2020; HE et al., 2020).

SimCLR, proposto por Chen et al. (2020), mostrou que um framework relativamente simples, baseado em aumentos fortes, codificador neural, projeção não linear e perda contrastiva, poderia aprender representações visuais competitivas com métodos supervisionados. O estudo demonstrou que a composição dos aumentos, o tamanho do batch e o tempo de treinamento são fatores críticos para o desempenho (CHEN et al., 2020).

MoCo, proposto por He et al. (2020), desenvolveu uma estratégia baseada em fila dinâmica e encoder momentum para manter grande conjunto de exemplos negativos sem depender exclusivamente de batches enormes. Essa abordagem foi importante para tornar o aprendizado contrastivo mais eficiente em certos cenários computacionais (HE et al., 2020).

O aprendizado contrastivo também foi aplicado a texto, fala, séries temporais e dados multimodais. Em CLIP, Radford et al. (2021) treinaram um codificador de imagem e um codificador de texto para alinhar pares imagem-legenda em larga escala. Embora utilize supervisão natural de pares coletados da web, o método ilustra o poder do contraste multimodal para criar representações transferíveis e uso zero-shot em tarefas visuais (RADFORD et al., 2021).

A força do aprendizado contrastivo está em aprender invariâncias. O modelo deve reconhecer que duas versões transformadas da mesma amostra mantêm identidade semântica, apesar de mudanças superficiais. Em imagens, isso ajuda a ignorar variações de cor ou enquadramento. Em áudio, pode ajudar a ignorar ruído. Em texto, pode preservar significado diante de paráfrases. Em séries

temporais, pode identificar padrões persistentes apesar de ruídos e deslocamentos (CHEN et al., 2020; GRILL et al., 2020).

Entretanto, o aprendizado contrastivo possui desafios. A escolha de exemplos negativos pode introduzir falsos negativos, isto é, amostras semanticamente semelhantes tratadas como diferentes. Em bases desbalanceadas, esse problema pode prejudicar representações. Além disso, alguns métodos exigem grandes batches, memória e tempo computacional significativo (CHEN et al., 2020; HE et al., 2020).

Métodos como BYOL e SimSiam buscaram reduzir dependência de exemplos negativos. BYOL utiliza duas redes, uma online e uma target, com atualização por média móvel, aprendendo representações por previsão entre diferentes visualizações da mesma imagem sem contraste explícito com negativos (GRILL et al., 2020). Essa linha mostrou que a dinâmica de treinamento e a assimetria arquitetural podem evitar colapso representacional sem necessidade de negativos (GRILL et al., 2020; CHEN; HE, 2021).

O aprendizado contrastivo contribuiu decisivamente para demonstrar que representações visuais, textuais e multimodais poderiam ser aprendidas sem rótulos tradicionais. Sua influência permanece central em modelos fundacionais, sistemas multimodais e aplicações de busca semântica (CHEN et al., 2020; RADFORD et al., 2021).

## 6. MODELAGEM MASCARADA, RECONSTRUÇÃO E APRENDIZADO GENERATIVO

A modelagem mascarada constitui uma das estratégias mais bem-sucedidas de aprendizado auto-supervisionado. A ideia é ocultar

parte da entrada e treinar o modelo para reconstruir ou prever a informação ausente. Essa lógica tornou-se fundamental no processamento de linguagem natural com BERT e, posteriormente, na visão computacional com Masked Autoencoders (DEVLIN et al., 2019; HE et al., 2022).

Em BERT, uma porcentagem dos tokens de entrada é mascarada, e o modelo deve prever quais tokens foram ocultados a partir do contexto bidirecional. Esse objetivo força a rede a aprender relações sintáticas, semânticas e contextuais, resultando em representações úteis para múltiplas tarefas linguísticas (DEVLIN et al., 2019). A modelagem mascarada tornou-se uma das bases do pré-treinamento moderno em linguagem natural (DEVLIN et al., 2019; LIU et al., 2019).

Na visão computacional, Masked Autoencoders, propostos por He et al. (2022), aplicaram a ideia de mascarar grande parte dos patches de uma imagem e treinar o modelo para reconstruir os pixels ausentes. O método mostrou que uma tarefa aparentemente simples poderia treinar Vision Transformers de forma escalável, com bons resultados em transferência para tarefas downstream (HE et al., 2022).

A modelagem mascarada tem uma vantagem importante: ela explora a estrutura interna dos dados sem exigir exemplos negativos. Em vez de comparar amostras, o modelo aprende a inferir partes ausentes a partir de partes observadas. Isso é especialmente relevante em domínios nos quais a definição de amostras positivas e negativas é difícil, como sinais biomédicos, documentos longos, imagens científicas ou dados multimodais complexos (HE et al., 2022; BAEVSKI et al., 2022).

Autoencoders e variational autoencoders também pertencem à tradição de reconstrução. Autoencoders aprendem a codificar uma entrada em representação comprimida e reconstruí-la. Variational Autoencoders introduzem modelagem probabilística do espaço latente, permitindo geração e interpolação. Embora nem todos sejam chamados de auto-supervisionados em sentido estrito contemporâneo, compartilham a ideia de aprender estrutura dos dados sem rótulos manuais (KINGMA; WELLING, 2014; VINCENT et al., 2008).

Em fala, wav2vec 2.0 combina mascaramento de representações latentes com tarefa contrastiva sobre unidades quantizadas, aprendendo representações de áudio a partir de dados não transcritos. Essa abordagem demonstrou que a fala possui estrutura suficiente para gerar supervisão interna, reduzindo dependência de transcrições extensas (BAEVSKI et al., 2020).

O data2vec, proposto por Baevski et al. (2022), avançou ao propor uma estrutura unificada para texto, visão e fala, na qual o modelo prevê representações latentes contextualizadas de uma rede professora a partir de entradas mascaradas. Esse movimento aponta para uma fronteira importante: objetivos auto-supervisionados independentes de modalidade, capazes de aproximar diferentes domínios de dados (BAEVSKI et al., 2022).

A modelagem mascarada, portanto, representa uma das bases mais promissoras da IA escalável. Ela é conceitualmente simples, altamente adaptável e eficaz em diferentes modalidades. Seu desafio está no custo computacional, na escolha da taxa de mascaramento, na arquitetura adequada e na avaliação da

qualidade das representações aprendidas (HE et al., 2022; BAEVSKI et al., 2022).

## 7. TRANSFORMERS, MODELOS FUNDACIONAIS E ESCALABILIDADE

A arquitetura Transformer transformou profundamente o aprendizado auto-supervisionado. Proposta por Vaswani et al. (2017), ela substituiu recorrência por mecanismos de atenção, permitindo modelar relações entre elementos de uma sequência de forma paralelizável e eficiente. Sua aplicação em linguagem natural, visão computacional, fala e multimodalidade tornou-a uma das arquiteturas centrais da IA contemporânea (VASWANI et al., 2017; DEVLIN et al., 2019; DOSOVITSKIY et al., 2021).

Os modelos fundacionais são geralmente pré-treinados em grandes volumes de dados e adaptados a múltiplas tarefas. Bommasani et al. (2021) definem esses modelos como sistemas amplos que servem de base para diversas aplicações downstream. O aprendizado auto-supervisionado é um dos mecanismos centrais dessa base, pois permite pré-treinamento em escala sem necessidade de rótulos manuais específicos para cada tarefa (BOMMASANI et al., 2021).

Em linguagem natural, modelos como BERT, GPT, RoBERTa, T5 e variantes posteriores demonstraram que pré-treinamento em larga escala seguido de adaptação pode gerar desempenho robusto em ampla variedade de tarefas. Embora existam diferenças entre objetivos autoregressivos, mascarados e sequência-a-sequência, todos exploram estrutura dos próprios textos para aprender representações ou capacidades gerativas (DEVLIN et al., 2019; RADFORD et al., 2019; BROWN et al., 2020; RAFFEL et al., 2020).

Em visão computacional, Vision Transformers e modelos auto-supervisionados como DINO e MAE mostraram que a combinação entre Transformers e pré-treinamento sem rótulos pode gerar representações visuais com propriedades emergentes, inclusive atenção a regiões semanticamente relevantes (CARON et al., 2021; HE et al., 2022; DOSOVITSKIY et al., 2021).

Em multimodalidade, CLIP demonstrou que o alinhamento entre imagens e textos em larga escala permite aprender representações transferíveis para tarefas visuais com uso de linguagem natural como interface de classificação. Esse resultado ampliou o campo de modelos multimodais e abriu caminho para sistemas capazes de relacionar diferentes formas de dados não estruturados (RADFORD et al., 2021).

A escalabilidade desses modelos decorre de três fatores principais. O primeiro é a capacidade de aproveitar grandes bases não rotuladas. O segundo é a reutilização de representações em múltiplas tarefas. O terceiro é a melhoria de desempenho com aumento de dados, parâmetros e computação, embora esse crescimento traga custos ambientais e concentração de infraestrutura (BROWN et al., 2020; KAPLAN et al., 2020; SCHWARTZ et al., 2020).

A IA escalável, entretanto, não deve ser confundida com crescimento ilimitado. Modelos maiores podem melhorar desempenho, mas também aumentam custo computacional, consumo energético, emissões, barreiras de acesso, opacidade e risco de reprodução de vieses em escala. A literatura sobre Green AI defende que eficiência, custo e sustentabilidade devem ser critérios de avaliação ao lado de acurácia (SCHWARTZ et al., 2020; STRUBELL; GANESH; MCCALLUM, 2019).

A fronteira atual está em modelos mais eficientes, adaptáveis e especializados. Técnicas como fine-tuning eficiente em parâmetros, distilação, quantização, pruning, retrieval-augmented generation, mixture of experts, aprendizado contínuo e modelos menores especializados buscam equilibrar desempenho e custo (HAN; MAO; DALLY, 2016; HOUlsBY et al., 2019; LEWIS et al., 2020).

Portanto, o aprendizado auto-supervisionado é infraestrutura epistêmica dos modelos fundacionais, mas sua consolidação exige maturidade técnica e governança. Escalar modelos sem escalar responsabilidade pode ampliar riscos tanto computacionais quanto sociais (BOMMASANI et al., 2021; BENDER et al., 2021).

## 8. APLICAÇÕES EM LINGUAGEM NATURAL, DOCUMENTOS E CONHECIMENTO TEXTUAL

O processamento de linguagem natural foi uma das áreas mais transformadas pelo aprendizado auto-supervisionado. Textos são abundantes, mas rótulos específicos para tarefas são escassos. A possibilidade de pré-treinar modelos em corpora massivos e adaptá-los para classificação, sumarização, tradução, busca semântica, extração de entidades, resposta a perguntas e geração textual tornou-se uma das principais conquistas da IA moderna (DEVLIN et al., 2019; BROWN et al., 2020).

Em documentos corporativos, jurídicos, acadêmicos e administrativos, o aprendizado auto-supervisionado permite construir representações semânticas capazes de organizar, recuperar, agrupar e classificar informações. Modelos de linguagem pré-treinados podem identificar similaridade entre documentos, extrair temas, apoiar análise de contratos, detectar inconsistências e

estruturar grandes acervos textuais (PETERS et al., 2018; DEVLIN et al., 2019).

Em contextos de baixa rotulação, modelos pré-treinados podem ser ajustados com poucos exemplos. Isso é particularmente relevante para domínios especializados, como direito, medicina, contabilidade, engenharia, meio ambiente e administração pública. A transferência de aprendizagem reduz a necessidade de grandes bases anotadas, embora ainda exija validação rigorosa por especialistas (HOWARD; RUDER, 2018; DEVLIN et al., 2019).

A busca semântica é uma aplicação direta. Em vez de procurar apenas palavras-chave, modelos auto-supervisionados podem representar consultas e documentos em espaços vetoriais, permitindo recuperar conteúdos semanticamente relacionados. Essa abordagem é central em sistemas modernos de recuperação de informação, assistentes inteligentes e aplicações de geração aumentada por recuperação (MIKOLOV et al., 2013; LEWIS et al., 2020).

A análise de sentimentos, opiniões e discurso também se beneficia do pré-treinamento. Modelos linguísticos capturam padrões pragmáticos e semânticos que podem ser adaptados a tarefas de classificação emocional ou análise social. Contudo, esses modelos podem reproduzir vieses presentes nos dados de treinamento, exigindo auditoria e mitigação (BOLUKBASI et al., 2016; BENDER et al., 2021).

Em línguas de poucos recursos, o aprendizado auto-supervisionado é especialmente estratégico. Grandes quantidades de texto bruto podem ser usadas para pré-treinamento mesmo quando há poucos

rótulos. Essa possibilidade pode apoiar tecnologias linguísticas em idiomas regionais, línguas indígenas, dialetos e variedades não hegemônicas, desde que haja cuidado com soberania de dados e participação das comunidades (JOSHI et al., 2020; BOMMASANI et al., 2021).

A exploração textual auto-supervisionada, portanto, não apenas melhora tarefas computacionais, mas também reorganiza a gestão do conhecimento. Arquivos antes dispersos podem tornar-se pesquisáveis, comparáveis, resumíveis e integráveis por representações semânticas (DEVLIN et al., 2019; LEWIS et al., 2020).

## 9. APLICAÇÕES EM VISÃO COMPUTACIONAL, IMAGENS E VÍDEOS

Na visão computacional, dados visuais são abundantes, mas sua anotação é altamente custosa. Rotular imagens para classificação pode ser trabalhoso; segmentar objetos pixel a pixel é ainda mais caro; anotar vídeos exige tempo, precisão temporal e consistência. O aprendizado auto-supervisionado oferece alternativa ao aprender representações visuais a partir de imagens e vídeos não rotulados (CHEN et al., 2020; CARON et al., 2021; HE et al., 2022).

Métodos contrastivos como SimCLR e MoCo mostraram que aumentos de imagem podem criar pares positivos para treinamento sem rótulos. O modelo aprende a reconhecer que diferentes transformações da mesma imagem preservam sua identidade semântica, produzindo embeddings úteis para classificação, detecção e segmentação após adaptação (CHEN et al., 2020; HE et al., 2020).

DINO evidenciou propriedades emergentes em Vision Transformers treinados por auto-supervisão. Caron et al. (2021) observaram que

mapas de atenção de modelos auto-supervisionados podem capturar regiões semanticamente relevantes da imagem, sugerindo que o modelo aprende estrutura visual sem rótulos explícitos de segmentação (CARON et al., 2021).

MAE tornou-se referência em modelagem mascarada visual. Ao mascarar grande parte dos patches da imagem e reconstruir a entrada, o modelo aprende representações escaláveis e transferíveis. Esse método aproximou a visão computacional da lógica de masked modeling que já havia sido bem-sucedida em linguagem natural (HE et al., 2022).

Em imagens médicas, aprendizado auto-supervisionado pode reduzir dependência de radiologistas, patologistas ou especialistas para rotulação inicial. Grandes bases de imagens podem ser usadas para pré-treinamento e depois ajustadas para detecção de achados específicos. Contudo, aplicações clínicas exigem validação, explicabilidade, avaliação de viés e controle regulatório (AZIZPOUR et al., 2016; BOMMASANI et al., 2021).

Em sensoriamento remoto e imagens geoespaciais, a auto-supervisão pode apoiar classificação de uso do solo, detecção de desmatamento, monitoramento agrícola, análise urbana e identificação de desastres. Como a anotação geoespacial é cara e depende de contexto territorial, representações auto-supervisionadas podem acelerar aplicações ambientais e urbanas (JEAN et al., 2016; CHEN et al., 2025).

Em vídeos, o desafio é ainda maior porque há dimensão temporal. Tarefas auto-supervisionadas podem prever frames futuros, ordenar segmentos, contrastar clipes, reconstruir trechos mascarados ou

alinhar áudio e imagem. O aprendizado de movimento, causalidade visual e continuidade temporal é uma fronteira relevante da IA escalável (MISRA; ZITNICK; HEBERT, 2016; FEICHTENHOFER et al., 2021).

A visão computacional auto-supervisionada permite transformar acervos visuais brutos em conhecimento operacional. Indústrias podem monitorar defeitos, cidades podem analisar tráfego, hospitais podem apoiar triagem, pesquisadores podem explorar imagens científicas e organizações ambientais podem monitorar territórios com menor dependência de rotulação manual (CHEN et al., 2020; HE et al., 2022).

## 10. APLICAÇÕES EM FALA, ÁUDIO, SINAIS DIGITAIS E SÉRIES TEMPORAIS

O aprendizado auto-supervisionado em fala e áudio ganhou destaque porque transcrições são caras, especialmente em línguas de poucos recursos, contextos ruidosos ou domínios técnicos. wav2vec 2.0 demonstrou que modelos podem aprender representações de áudio bruto e atingir desempenho elevado em reconhecimento de fala com pouca transcrição rotulada (BAEVSKI et al., 2020).

A fala possui estrutura acústica, fonética, prosódica e temporal. Modelos auto-supervisionados podem explorar mascaramento, contraste, quantização e previsão de segmentos para aprender padrões relevantes. Essa estratégia pode apoiar reconhecimento de fala, identificação de locutor, diarização, tradução de fala e análise de emoções vocais (BAEVSKI et al., 2020; HSU et al., 2021).

Em sinais digitais e séries temporais, o aprendizado auto-supervisionado pode ser aplicado a sensores industriais, sinais biomédicos, dados financeiros, telemetria, redes elétricas, tráfego, clima e Internet das Coisas. Esses dados frequentemente apresentam padrões temporais complexos, ruído, sazonalidade, eventos raros e anomalias pouco rotuladas (FRANCESCHI; DIEULEVEUT; JAGGI, 2019; YUE et al., 2022).

Tarefas pretextuais para séries temporais podem incluir previsão de janelas futuras, reconstrução de trechos mascarados, ordenação temporal, contraste entre segmentos do mesmo sinal, detecção de transformações e alinhamento entre sensores. O objetivo é aprender representações que capturem regimes normais, transições, periodicidades e anomalias (FRANCESCHI; DIEULEVEUT; JAGGI, 2019; YUE et al., 2022).

Em saúde, sinais como ECG, EEG, pressão, oximetria e registros de dispositivos vestíveis podem ser explorados por auto-supervisão para reduzir dependência de rótulos clínicos extensos. Entretanto, esses usos exigem validação rigorosa, privacidade, segurança e interpretação por profissionais qualificados (RAJAGOPALAN et al., 2020; BOMMASANI et al., 2021).

Em indústrias, séries temporais de máquinas podem ser usadas para manutenção preditiva. Como falhas reais são raras e rotuladas com dificuldade, aprender o comportamento normal por auto-supervisão pode permitir detectar desvios, padrões anômalos e degradação progressiva (ZHANG et al., 2019; YUE et al., 2022).

Em ambientes urbanos e ambientais, sensores produzem fluxos contínuos sobre clima, poluição, drenagem, mobilidade, energia e

ruído. O aprendizado auto-supervisionado pode apoiar previsão, compressão, detecção de anomalias e descoberta de padrões em larga escala, reduzindo dependência de rotulação humana (KITCHIN, 2014; FRANCESCHI; DIEULEVEUT; JAGGI, 2019).

Assim, a auto-supervisão amplia a IA para domínios nos quais rótulos são raros, eventos críticos são incomuns e dados brutos são abundantes. Essa característica é decisiva para aplicações científicas, industriais e públicas (BAEVSKI et al., 2020; YUE et al., 2022).

## 11. APRENDIZADO MULTIMODAL E ALINHAMENTO ENTRE REPRESENTAÇÕES

O aprendizado multimodal busca integrar diferentes modalidades de dados, como texto, imagem, áudio, vídeo, sensores, tabelas e sinais. A realidade informacional é frequentemente multimodal: uma consulta médica envolve texto, imagem, áudio e sinais; um sistema urbano envolve vídeo, sensores, mapas e documentos; uma plataforma educacional envolve fala, texto, imagem e comportamento de navegação (BALTRUŠAITIS; AHUJA; MORENCY, 2019; RADFORD et al., 2021).

O aprendizado auto-supervisionado é particularmente útil nesse contexto porque modalidades diferentes podem supervisionar umas às outras. Uma legenda pode orientar representação visual; o áudio pode alinhar-se ao vídeo; sensores diferentes podem fornecer perspectivas complementares do mesmo evento; texto e imagem podem ser contrastados para aprender conceitos compartilhados (RADFORD et al., 2021; BAEVSKI et al., 2022).

CLIP demonstrou que pares imagem-texto coletados em larga escala podem treinar modelos capazes de associar conceitos visuais

a linguagem natural. Esse alinhamento permitiu classificação zero-shot, busca multimodal e generalização para múltiplas tarefas visuais sem treinamento específico em cada dataset (RADFORD et al., 2021).

O data2vec avançou em direção a uma estrutura comum para fala, visão e linguagem, buscando reduzir a fragmentação de objetivos auto-supervisionados específicos por modalidade. Essa linha aponta para modelos mais gerais, capazes de aprender representações abstratas a partir de diferentes formas de dados (BAEVSKI et al., 2022).

A multimodalidade também é estratégica para dados não estruturados corporativos. Uma organização pode ter documentos, imagens, vídeos, áudios de atendimento, registros de sensores e bancos de conhecimento. Representações multimodais permitem busca cruzada, sumarização, classificação, recomendação e apoio à decisão com base em múltiplas fontes (BALTRUŠAITIS; AHUJA; MORENCY, 2019; BOMMASANI et al., 2021).

O desafio do aprendizado multimodal está em alinhar modalidades com granularidades diferentes. Uma imagem pode corresponder a uma frase; um vídeo a múltiplos eventos; um áudio a uma transcrição; um sensor a uma condição física. O modelo precisa aprender correspondências sem reduzir indevidamente a complexidade de cada modalidade (BALTRUŠAITIS; AHUJA; MORENCY, 2019; RADFORD et al., 2021).

Há também riscos de viés. Pares imagem-texto coletados da internet podem reproduzir estereótipos, associações discriminatórias e desigualdades culturais. Modelos multimodais

podem aprender correlações problemáticas entre aparência, gênero, raça, território, classe social e linguagem. A avaliação ética deve ser parte do desenvolvimento (BENDER et al., 2021; BOMMASANI et al., 2021).

O aprendizado multimodal auto-supervisionado representa uma fronteira decisiva porque aproxima a IA da complexidade do mundo real. Sistemas inteligentes precisam compreender que dados raramente aparecem isolados; eventos, documentos, imagens, sinais e linguagem se complementam (RADFORD et al., 2021; BAEVSKI et al., 2022).

## 12. MODELO MIEAS-DNE: EXPLORAÇÃO AUTO-SUPERVISIONADA DE DADOS NÃO ESTRUTURADOS

Este artigo propõe o **MIEAS-DNE — Modelo Integrado de Exploração Auto-Supervisionada de Dados Não Estruturados.** O modelo busca orientar organizações, pesquisadores e desenvolvedores na construção de pipelines de aprendizado auto-supervisionado aplicados a grandes volumes de dados brutos, heterogêneos e pouco rotulados.

O MIEAS-DNE é estruturado em oito camadas: ingestão de dados, curadoria e qualidade, pré-processamento por modalidade, geração de tarefas pretextuais, pré-treinamento auto-supervisionado, avaliação de representações, adaptação downstream e governança responsável. Essas camadas operam de forma integrada e iterativa (BENGIO; COURVILLE; VINCENT, 2013; BOMMASANI et al., 2021).

A primeira camada, ingestão de dados, envolve coleta de textos, imagens, áudios, vídeos, sinais, logs e documentos. Essa etapa deve registrar origem, formato, licença, contexto, sensibilidade,

representatividade e limitações dos dados. A qualidade da IA começa antes do treinamento (MITCHELL et al., 2019; GEBRU et al., 2021).

A segunda camada, curadoria e qualidade, inclui deduplicação, remoção de ruído, detecção de dados corrompidos, filtragem de conteúdos sensíveis, balanceamento de domínios, documentação e análise de viés. Dados não estruturados em larga escala podem conter redundância, desinformação, linguagem ofensiva, dados pessoais e padrões discriminatórios (BENDER et al., 2021; GEBRU et al., 2021).

A terceira camada, pré-processamento por modalidade, adapta cada tipo de dado às exigências do modelo. Textos podem ser tokenizados; imagens podem ser redimensionadas e aumentadas; áudios podem ser segmentados; séries temporais podem ser normalizadas; documentos podem passar por OCR; e vídeos podem ser amostrados em frames ou clipes (DEVLIN et al., 2019; CHEN et al., 2020; BAEVSKI et al., 2020).

A quarta camada, geração de tarefas pretextuais, define a forma de supervisão interna. Para texto, pode-se usar mascaramento ou previsão autoregressiva. Para imagem, contraste ou reconstrução de patches mascarados. Para áudio, mascaramento de representações latentes. Para séries temporais, previsão de janelas ou contraste entre segmentos. Para multimodalidade, alinhamento entre pares (DEVLIN et al., 2019; CHEN et al., 2020; HE et al., 2022; RADFORD et al., 2021).

A quinta camada, pré-treinamento auto-supervisionado, envolve escolha de arquitetura, função de perda, infraestrutura

computacional, estratégia de otimização, regularização, batch size, taxa de aprendizado e monitoramento. Essa etapa pode exigir grande poder computacional, especialmente em modelos fundacionais (VASWANI et al., 2017; BROWN et al., 2020; SCHWARTZ et al., 2020).

A sexta camada, avaliação de representações, verifica se os embeddings aprendidos são úteis. Isso pode ocorrer por linear probing, fine-tuning, few-shot evaluation, clustering, recuperação semântica, transferência entre domínios, robustez a ruído e análise qualitativa de vizinhanças vetoriais (CHEN et al., 2020; HE et al., 2022).

A sétima camada, adaptação downstream, ajusta o modelo a tarefas específicas. Pode envolver classificação, detecção de anomalias, sumarização, busca semântica, reconhecimento de fala, segmentação, recomendação, previsão ou apoio à decisão. O objetivo é aproveitar o pré-treinamento para reduzir rótulos necessários e melhorar desempenho (DEVLIN et al., 2019; BAEVSKI et al., 2020).

A oitava camada, governança responsável, inclui privacidade, segurança, explicabilidade, documentação, auditoria de viés, eficiência computacional, controle de acesso e monitoramento pós-implantação. Essa camada é indispensável porque modelos auto-supervisionados podem aprender padrões problemáticos dos dados brutos em escala (MITCHELL et al., 2019; BENDER et al., 2021; BOMMASANI et al., 2021).

O MIEAS-DNE opera em ciclo contínuo. Novos dados podem atualizar representações; avaliações podem revelar vieses; tarefas downstream podem indicar lacunas; e auditorias podem exigir

curadoria adicional. Assim, a exploração auto-supervisionada não é etapa única, mas processo permanente de aprendizagem e governança.

## 13. AVALIAÇÃO, GENERALIZAÇÃO E TRANSFERÊNCIA DE APRENDIZAGEM

A avaliação de modelos auto-supervisionados é desafiadora porque o objetivo principal não é apenas bom desempenho na tarefa pretextual, mas qualidade da representação aprendida. Um modelo pode reconstruir entradas com precisão e ainda aprender representações pouco úteis para tarefas downstream. Por isso, a avaliação deve ir além da perda de pré-treinamento (BENGIO; COURVILLE; VINCENT, 2013; CHEN et al., 2020).

O linear probing é uma estratégia comum. Nela, o codificador pré-treinado é congelado e apenas um classificador linear é treinado sobre as representações. Se o desempenho for alto, entende-se que as representações já capturam informação relevante para a tarefa. SimCLR e outros métodos utilizaram essa estratégia em benchmarks de visão computacional (CHEN et al., 2020; HE et al., 2020).

O fine-tuning avalia a capacidade de adaptação do modelo. Após pré-treinamento, todos ou parte dos parâmetros são ajustados para uma tarefa específica. Essa avaliação reflete uso prático, especialmente quando há poucos rótulos. BERT, wav2vec 2.0 e MAE demonstraram forte desempenho por fine-tuning em tarefas downstream (DEVLIN et al., 2019; BAEVSKI et al., 2020; HE et al., 2022).

A avaliação few-shot e zero-shot tornou-se relevante com modelos fundacionais e multimodais. Em zero-shot, o modelo executa tarefa sem treinamento específico naquela base; em few-shot, adapta-se com poucos exemplos. CLIP popularizou avaliação zero-shot em visão por meio de descrições textuais de classes (RADFORD et al., 2021).

A generalização fora de distribuição é uma dimensão crítica. Modelos podem ter bom desempenho em benchmarks conhecidos, mas falhar em dados de novos domínios, populações, sensores, línguas ou condições ambientais. A robustez deve ser avaliada em cenários reais e variados (HENDRYCKS; DIETTERICH, 2019; BOMMASANI et al., 2021).

A interpretabilidade das representações também importa. Embora embeddings sejam vetores de alta dimensão, análises de vizinhança, visualização, clustering, probes linguísticos, mapas de atenção e testes contrafactuais podem ajudar a compreender o que foi aprendido. No entanto, essas técnicas têm limitações e não substituem avaliação empírica (RUDIN, 2019; JAIN; WALLACE, 2019).

A avaliação deve também considerar eficiência. Um modelo que obtém ganho marginal com enorme custo computacional pode não ser adequado em contextos de sustentabilidade, acesso democrático ou implantação em borda. A literatura de Green AI defende reportar custo computacional, energia e emissões junto aos resultados de desempenho (SCHWARTZ et al., 2020; STRUBELL; GANESH; MCCALLUM, 2019).

Portanto, avaliar aprendizado auto-supervisionado exige múltiplas métricas: desempenho downstream, transferência, robustez,

eficiência, viés, interpretabilidade e utilidade prática. A qualidade da representação não é propriedade abstrata; depende da tarefa, do domínio e dos riscos associados (BOMMASANI et al., 2021; MITCHELL et al., 2019).

## 14. DESAFIOS TÉCNICOS, LIMITAÇÕES E RISCOS

Apesar de seu potencial, o aprendizado auto-supervisionado apresenta desafios relevantes. O primeiro é o custo computacional. Pré-treinar modelos grandes em dados massivos pode exigir milhares de horas de GPU, infraestrutura especializada e alto consumo energético. Esse custo pode concentrar poder em poucas organizações com capacidade computacional elevada (STRUBELL; GANESH; MCCALLUM, 2019; SCHWARTZ et al., 2020).

O segundo desafio é a qualidade dos dados. Como o modelo aprende dos próprios dados, ruídos, vieses, duplicações, informações falsas, conteúdos tóxicos e desequilíbrios podem ser incorporados às representações. Dados não curados em larga escala podem gerar modelos que reproduzem estereótipos ou associações indesejadas (BENDER et al., 2021; GEBRU et al., 2021).

O terceiro desafio é a definição da tarefa pretextual. Uma tarefa pode induzir o modelo a aprender atalhos superficiais. Em visão, o modelo pode usar artefatos de augmentação; em texto, pode memorizar padrões espúrios; em séries temporais, pode capturar ruídos em vez de fenômenos relevantes. A tarefa deve ser alinhada ao tipo de estrutura que se deseja aprender (DOERSCH; GUPTA; EFROS, 2015; CHEN et al., 2020).

O quarto desafio é o colapso representacional. Em alguns métodos, especialmente sem exemplos negativos, existe risco de o modelo

produzir representações triviais ou indistinguíveis. Métodos como BYOL e SimSiam propuseram mecanismos arquiteturais para evitar esse problema, mas a estabilidade do treinamento continua sendo tema importante (GRILL et al., 2020; CHEN; HE, 2021).

O quinto desafio é a avaliação. Benchmarks podem não representar contextos reais. Um modelo excelente em ImageNet, GLUE ou LibriSpeech pode apresentar desempenho inferior em ambientes locais, dados ruidosos, línguas regionais, sensores específicos ou domínios especializados. A avaliação deve ser contextualizada (HENDRYCKS; DIETTERICH, 2019; BOMMASANI et al., 2021).

O sexto desafio é a privacidade. Pré-treinamento em grandes bases pode incluir dados pessoais, documentos sensíveis ou informações protegidas. Modelos podem memorizar trechos ou permitir inferências indesejadas. Técnicas de anonimização, filtragem, aprendizado federado, privacidade diferencial e governança de acesso tornam-se relevantes (CARLINI et al., 2021; DWORK; ROTH, 2014).

O sétimo desafio é a explicabilidade. Representações profundas são difíceis de interpretar, e decisões baseadas nelas podem afetar pessoas em contextos de crédito, saúde, segurança, educação ou justiça. A opacidade é especialmente problemática quando o modelo foi pré-treinado em dados amplos e adaptado a tarefas sensíveis (RUDIN, 2019; MITCHELL et al., 2019).

O oitavo desafio é o risco de uso indevido. Representações poderosas podem ser usadas para vigilância, manipulação, desinformação, reconhecimento biométrico abusivo ou automação discriminatória. A escalabilidade técnica deve ser acompanhada por

limites éticos e regulatórios (BENDER et al., 2021; BOMMASANI et al., 2021).

Assim, o aprendizado auto-supervisionado não é solução neutra. Ele amplia capacidades da IA, mas também amplia responsabilidades. Sua adoção deve ser acompanhada por governança técnica, social e ética (MITCHELL et al., 2019; SCHWARTZ et al., 2020).

## 15. RESULTADOS DA ANÁLISE

A análise da literatura permite identificar dez resultados principais.

O primeiro resultado é que o aprendizado auto-supervisionado reduz a dependência de rotulação manual, permitindo explorar grandes volumes de dados brutos em texto, imagem, áudio, fala, vídeo, sinais e dados multimodais (DEVLIN et al., 2019; CHEN et al., 2020; BAEVSKI et al., 2020).

O segundo resultado é que representações aprendidas por auto-supervisão podem ser transferidas para múltiplas tarefas downstream, melhorando desempenho em cenários de poucos rótulos e reduzindo custo de adaptação (DEVLIN et al., 2019; HE et al., 2022).

O terceiro resultado é que tarefas contrastivas são eficazes para aprender invariâncias e similaridades semânticas, especialmente em visão computacional e multimodalidade (CHEN et al., 2020; RADFORD et al., 2021).

O quarto resultado é que modelagem mascarada e reconstrução representam estratégias escaláveis e adaptáveis, com forte impacto

em linguagem, visão e fala (DEVLIN et al., 2019; HE et al., 2022; BAEVSKI et al., 2020).

O quinto resultado é que Transformers ampliaram a escalabilidade do aprendizado auto-supervisionado, permitindo modelar dependências longas e transferir arquiteturas entre modalidades (VASWANI et al., 2017; DOSOVITSKIY et al., 2021).

O sexto resultado é que modelos fundacionais dependem fortemente de pré-treinamento auto-supervisionado ou fracamente supervisionado em larga escala, servindo de base para múltiplas aplicações (BROWN et al., 2020; BOMMASANI et al., 2021).

O sétimo resultado é que dados não estruturados tornam-se mais exploráveis quando convertidos em representações vetoriais semanticamente organizadas, permitindo busca, agrupamento, classificação, recomendação e análise de similaridade (MIKOLOV et al., 2013; BENGIO; COURVILLE; VINCENT, 2013).

O oitavo resultado é que a eficiência computacional é condição estratégica, pois modelos auto-supervisionados em escala podem exigir alto consumo energético e infraestrutura especializada (STRUBELL; GANESH; MCCALLUM, 2019; SCHWARTZ et al., 2020).

O nono resultado é que qualidade e governança dos dados são fundamentais, pois modelos auto-supervisionados podem absorver ruídos, vieses e informações sensíveis presentes nas bases de pré-treinamento (BENDER et al., 2021; GEBRU et al., 2021).

O décimo resultado é que o MIEAS-DNE oferece estrutura conceitual para organizar pipelines de exploração auto-supervisionada,

integrando ingestão, curadoria, tarefas pretextuais, pré-treinamento, avaliação, adaptação e governança responsável.

## 16. DISCUSSÃO

A discussão central deste artigo é que o aprendizado auto-supervisionado representa uma mudança profunda na forma como a inteligência artificial aprende a partir de dados. Em vez de depender exclusivamente de rótulos humanos, os modelos passam a explorar regularidades internas dos próprios dados. Essa mudança é especialmente relevante porque a produção de dados não estruturados cresce em velocidade muito superior à capacidade humana de anotação (LECUN; MISRA, 2021; GOODFELLOW; BENGIO; COURVILLE, 2016).

O primeiro ponto de discussão refere-se à economia da rotulação. A IA supervisionada depende de bases anotadas, mas a anotação é um gargalo técnico, financeiro e epistemológico. O aprendizado auto-supervisionado não elimina completamente a necessidade de rótulos, mas desloca sua função. Rótulos passam a ser mais importantes para adaptação, validação e avaliação do que para todo o processo de aprendizagem inicial (DEVLIN et al., 2019; BAEVSKI et al., 2020).

O segundo ponto é a centralidade das representações. O valor do aprendizado auto-supervisionado está menos na tarefa pretextual em si e mais na representação que emerge dela. A tarefa de prever tokens mascarados, reconstruir patches ou contrastar visualizações é meio para aprender estruturas latentes úteis. Essa distinção é crucial para compreender por que modelos auto-supervisionados podem

ser aplicados a tarefas diferentes daquelas usadas no pré-treinamento (BENGIO; COURVILLE; VINCENT, 2013; HE et al., 2022).

O terceiro ponto envolve escalabilidade. O aprendizado auto-supervisionado permite treinar modelos com dados abundantes e rótulos escassos, tornando-se compatível com a escala da internet, de sensores, de arquivos corporativos, de imagens científicas e de dados multimodais. Essa escalabilidade explica seu papel nos modelos fundacionais, que funcionam como infraestruturas gerais de representação (BOMMASANI et al., 2021; BROWN et al., 2020).

O quarto ponto refere-se à transferência de aprendizagem. Modelos pré-treinados podem ser adaptados a tarefas específicas com menos dados rotulados. Essa característica democratiza parcialmente o uso da IA em domínios especializados, pois organizações menores podem aproveitar modelos pré-treinados em vez de treinar tudo do zero. Contudo, a dependência de modelos fundacionais desenvolvidos por poucas organizações também cria riscos de concentração tecnológica (BOMMASANI et al., 2021; SCHWARTZ et al., 2020).

O quinto ponto é a multimodalidade. O mundo não é unimodal, e a inteligência humana integra visão, linguagem, som, movimento e contexto. O aprendizado auto-supervisionado multimodal aproxima a IA dessa complexidade ao alinhar representações entre modalidades. CLIP, data2vec e modelos multimodais posteriores demonstram que a supervisão pode emergir de correspondências naturais entre dados (RADFORD et al., 2021; BAEVSKI et al., 2022).

O sexto ponto envolve a qualidade dos dados. A máxima “mais dados” não substitui “melhores dados”. Modelos auto-

supervisionados aprendem padrões presentes no corpus, incluindo ruído, preconceito, desinformação e redundância. A curadoria de dados, documentação e auditoria tornam-se tão importantes quanto arquitetura e poder computacional (BENDER et al., 2021; GEBRU et al., 2021).

O sétimo ponto é a eficiência. O sucesso dos métodos auto-supervisionados em larga escala pode estimular modelos cada vez maiores, mas esse crescimento possui custo energético, ambiental e econômico. A pesquisa precisa equilibrar desempenho com sustentabilidade computacional. Green AI e eficiência de modelos devem ser incorporadas como critérios de qualidade científica (STRUBELL; GANESH; MCCALLUM, 2019; SCHWARTZ et al., 2020).

O oitavo ponto é a governança ética. Modelos treinados em dados massivos podem ser aplicados em contextos sensíveis, como saúde, educação, crédito, segurança e justiça. A origem dos dados, a representatividade, a privacidade, o viés e a explicabilidade precisam ser avaliados antes da implantação. A auto-supervisão não dispensa responsabilidade humana; ao contrário, amplia a necessidade de governança (MITCHELL et al., 2019; BOMMASANI et al., 2021).

O nono ponto é que o aprendizado auto-supervisionado não substitui conhecimento especializado. Em domínios críticos, como medicina, engenharia, direito e políticas públicas, modelos podem apoiar análise, mas a validação deve envolver especialistas. Representações poderosas podem revelar padrões, mas sua interpretação exige contexto técnico e social (RUDIN, 2019; MITCHELL et al., 2019).

O décimo ponto é que a auto-supervisão abre novas formas de descoberta. Ao explorar padrões latentes em dados não estruturados, modelos podem sugerir agrupamentos, relações semânticas, anomalias, tendências e estruturas que não haviam sido rotuladas previamente. Essa capacidade pode apoiar pesquisa científica, inteligência organizacional e inovação em domínios complexos (BENGIO; COURVILLE; VINCENT, 2013; LECUN; MISRA, 2021).

Assim, o aprendizado auto-supervisionado não é apenas técnica de machine learning, mas mudança na infraestrutura cognitiva da IA. Ele altera a relação entre dado, rótulo, representação, generalização e escala. Seu futuro dependerá da capacidade de combinar desempenho, eficiência, governança e responsabilidade social (BOMMASANI et al., 2021; SCHWARTZ et al., 2020).

## 17. DIRETRIZES PARA APLICAÇÃO DO APRENDIZADO AUTO-SUPERVISIONADO

A partir da análise realizada, propõem-se diretrizes para aplicação responsável e eficaz do aprendizado auto-supervisionado em dados não estruturados.

Primeiro, definir claramente o domínio dos dados, sua origem, sua sensibilidade, sua qualidade e sua representatividade antes do pré-treinamento (GEBRU et al., 2021; MITCHELL et al., 2019).

Segundo, escolher tarefa pretextual compatível com a estrutura do dado. Texto, imagem, áudio, vídeo, sinais e séries temporais exigem estratégias diferentes de mascaramento, contraste, reconstrução ou previsão (DEVLIN et al., 2019; CHEN et al., 2020; BAEVSKI et al., 2020).

Terceiro, documentar o pipeline de dados, incluindo filtros, exclusões, deduplicação, anonimização, licenças, limitações e possíveis vieses (MITCHELL et al., 2019; GEBRU et al., 2021).

Quarto, utilizar avaliação downstream adequada, combinando linear probing, fine-tuning, few-shot evaluation, robustez fora de distribuição e análise qualitativa de representações (CHEN et al., 2020; HE et al., 2022).

Quinto, evitar avaliar modelos apenas pela tarefa pretextual. A qualidade da representação deve ser medida por sua utilidade em tarefas reais (BENGIO; COURVILLE; VINCENT, 2013; BOMMASANI et al., 2021).

Sexto, considerar eficiência computacional, energia, custo e emissões como critérios de desenvolvimento e comparação entre modelos (SCHWARTZ et al., 2020; STRUBELL; GANESH; MCCALLUM, 2019).

Sétimo, usar modelos pré-treinados quando apropriado, evitando treinar do zero sem necessidade. A reutilização responsável pode reduzir custo e acelerar aplicações (HOWARD; RUDER, 2018; BOMMASANI et al., 2021).

Oitavo, adaptar modelos ao domínio por fine-tuning, adapters, prompting, retrieval ou técnicas eficientes, conforme disponibilidade de rótulos e criticidade da tarefa (HOUlsBY et al., 2019; LEWIS et al., 2020).

Nono, implementar auditoria de viés, privacidade e segurança antes da implantação, especialmente em aplicações sensíveis (BENDER et al., 2021; CARLINI et al., 2021).

Décimo, manter humanos especialistas no ciclo de validação quando o modelo for aplicado a decisões de alto impacto (RUDIN, 2019; MITCHELL et al., 2019).

Décimo primeiro, monitorar desempenho pós-implantação, pois distribuição de dados pode mudar ao longo do tempo e degradar representações (HENDRYCKS; DIETTERICH, 2019).

Décimo segundo, evitar automatização sem explicabilidade em contextos críticos. Representações profundas devem ser acompanhadas por mecanismos de interpretação, avaliação e contestabilidade (RUDIN, 2019).

Décimo terceiro, tratar dados não estruturados como patrimônio informacional sensível. Sua exploração deve respeitar privacidade, direitos autorais, licenças, consentimento e finalidade legítima (BENDER et al., 2021; GEBRU et al., 2021).

Décimo quarto, incorporar multimodalidade quando houver ganho real de contexto, evitando integração artificial de modalidades sem necessidade (BALTRUŠAITIS; AHUJA; MORENCY, 2019; RADFORD et al., 2021).

Décimo quinto, compreender que aprendizado auto-supervisionado é processo contínuo de representação, avaliação e governança, não apenas uma etapa técnica de pré-treinamento.

## 18. CONSIDERAÇÕES FINAIS

O aprendizado auto-supervisionado consolidou-se como uma das principais fronteiras da inteligência artificial escalável. Em um cenário marcado por crescimento exponencial de dados digitais e

escassez relativa de rótulos qualificados, esse paradigma permite explorar textos, imagens, vídeos, áudios, sinais e dados multimodais por meio de representações aprendidas a partir da própria estrutura dos dados (DEVLIN et al., 2019; CHEN et al., 2020; BAEVSKI et al., 2020).

O estudo demonstrou que a auto-supervisão desloca a dependência da rotulação manual para a construção de tarefas pretextuais inteligentes. Mascaramento, reconstrução, contraste, previsão temporal, destilação e alinhamento multimodal tornam-se mecanismos pelos quais modelos aprendem padrões latentes sem supervisão humana explícita para cada amostra (HE et al., 2022; BAEVSKI et al., 2022; RADFORD et al., 2021).

A análise evidenciou que o aprendizado auto-supervisionado possui impacto significativo em linguagem natural, visão computacional, fala, séries temporais, sinais digitais e multimodalidade. Em todos esses domínios, a principal vantagem está na capacidade de aprender representações transferíveis, reduzindo necessidade de bases extensamente rotuladas e ampliando desempenho em cenários de poucos exemplos (DEVLIN et al., 2019; HE et al., 2022).

A proposta do MIEAS-DNE oferece uma estrutura conceitual para organizar pipelines de exploração auto-supervisionada de dados não estruturados. Suas camadas de ingestão, curadoria, pré-processamento, geração de tarefas pretextuais, pré-treinamento, avaliação, adaptação e governança permitem tratar o processo de forma sistêmica, indo além da escolha isolada de um algoritmo.

Conclui-se que o aprendizado auto-supervisionado pode ampliar profundamente a capacidade de extrair conhecimento de dados

não estruturados, apoiar modelos fundacionais, reduzir custos de rotulação, favorecer transferência entre domínios e acelerar aplicações de IA em ciência, indústria, saúde, educação, gestão pública e negócios.

Contudo, essa fronteira tecnológica exige cautela. Modelos auto-supervisionados podem ser computacionalmente caros, opacos, sensíveis a vieses, dependentes de dados de qualidade e potencialmente problemáticos em termos de privacidade e uso indevido. Por isso, seu avanço deve ser acompanhado por governança de dados, documentação, auditoria, eficiência computacional, avaliação robusta e compromisso ético.

A inteligência artificial escalável do futuro não será construída apenas pela ampliação de parâmetros e dados, mas pela capacidade de aprender de forma eficiente, generalizável, responsável e alinhada às necessidades humanas. O aprendizado auto-supervisionado representa um passo decisivo nessa direção.

## REFERÊNCIAS BIBLIOGRÁFICAS

AZIZPOUR, Hossein; RAZAVIAN, Ali Sharif; SULLIVAN, Josephine; MAKI, Atsuto; CARLSSON, Stefan. Factors of transferability for a generic convnet representation. **IEEE Transactions on Pattern Analysis and Machine Intelligence,** v. 38, n. 9, p. 1790-1802, 2016.

BAEVSKI, Alexei; HSU, Wei-Ning; XU, Qiantong; BABU, Arun; GU, Jiatao; AULI, Michael. data2vec: a general framework for self-supervised learning in speech, vision and language. In: **Proceedings of the 39th International Conference on Machine Learning.** PMLR, v. 162, p. 1298-1312, 2022.

BAEVSKI, Alexei; ZHOU, Henry; MOHAMED, Abdelrahman; AULI, Michael. wav2vec 2.0: a framework for self-supervised learning of speech representations. In: **Advances in Neural Information Processing Systems,** v. 33, p. 12449-12460, 2020.

BALTRUŠAITIS, Tadas; AHUJA, Chaitanya; MORENCY, Louis-Philippe. Multimodal machine learning: a survey and taxonomy. **IEEE Transactions on Pattern Analysis and Machine Intelligence,** v. 41, n. 2, p. 423-443, 2019.

BENDER, Emily M.; GEBRU, Timnit; MCMILLAN-MAJOR, Angelina; SHMITCHELL, Shmargaret. On the dangers of stochastic parrots: can language models be too big? In: **Proceedings of the ACM Conference on Fairness, Accountability, and Transparency.** New York: ACM, 2021. p. 610-623.

BENGIO, Yoshua; COURVILLE, Aaron; VINCENT, Pascal. Representation learning: a review and new perspectives. **IEEE Transactions on Pattern Analysis and Machine Intelligence,** v. 35, n. 8, p. 1798-1828, 2013.

BOLUKBASI, Tolga; CHANG, Kai-Wei; ZOU, James; SALIGRAMA, Venkatesh; KALAI, Adam. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: **Advances in Neural Information Processing Systems,** v. 29, 2016.

BOMMASANI, Rishi et al. On the opportunities and risks of foundation models. **arXiv preprint arXiv:2108.07258,** 2021.

BROWN, Tom B. et al. Language models are few-shot learners. In: **Advances in Neural Information Processing Systems,** v. 33, p. 1877-1901, 2020.

CARLINI, Nicholas et al. Extracting training data from large language models. In: **30th USENIX Security Symposium.** Berkeley: USENIX Association, 2021. p. 2633-2650.

CARON, Mathilde; TOUVRON, Hugo; MISRA, Ishan; JÉGOU, Hervé; MAIRAL, Julien; BOJANOWSKI, Piotr; JOUlin, Armand. Emerging properties in self-supervised vision transformers. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision.** Montreal: IEEE, 2021. p. 9650-9660.

CHEN, Ting; KORNBLITH, Simon; NOROUZI, Mohammad; HINTON, Geoffrey. A simple framework for contrastive learning of visual representations. In: **Proceedings of the 37th International Conference on Machine Learning.** PMLR, v. 119, p. 1597-1607, 2020.

CHEN, Xinlei; HE, Kaiming. Exploring simple siamese representation learning. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.** Nashville: IEEE, 2021. p. 15750-15758.

CHEN, Yile et al. Self-supervised representation learning for geospatial objects: a survey. **Information Fusion,** 2025.

CRESWELL, John W. **Research design: qualitative, quantitative, and mixed methods approaches.** 4. ed. Thousand Oaks: SAGE, 2014.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of NAACL-HLT 2019.** Minneapolis: Association for Computational Linguistics, 2019. p. 4171-4186.

DOERSCH, Carl; GUPTA, Abhinav; EFROS, Alexei A. Unsupervised visual representation learning by context prediction. In: **Proceedings of the IEEE International Conference on Computer Vision.** Santiago: IEEE, 2015. p. 1422-1430.

DOSOVITSKIY, Alexey et al. An image is worth 16x16 words: transformers for image recognition at scale. In: **International Conference on Learning Representations,** 2021.

DWORK, Cynthia; ROTH, Aaron. **The algorithmic foundations of differential privacy.** Hanover: Now Publishers, 2014.

FEICHTENHOFER, Christoph; FAN, Haoqi; XIONG, Bo; GIRSHICK, Ross; HE, Kaiming. A large-scale study on unsupervised spatiotemporal representation learning. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.** Nashville: IEEE, 2021. p. 3299-3309.

FRANCESCHI, Jean-Yves; DIEULEVEUT, Aymeric; JAGGI, Martin. Unsupervised scalable representation learning for multivariate time series. In: **Advances in Neural Information Processing Systems,** v. 32, 2019.

GEBRU, Timnit; MORGENSTERN, Jamie; VECCHIONE, Briana; VAUGHAN, Jennifer Wortman; WALLACH, Hanna; DAUMÉ III, Hal; CRAWFORD, Kate. Datasheets for datasets. **Communications of the ACM,** v. 64, n. 12, p. 86-92, 2021.

GIDARIS, Spyros; SINGH, Praveer; KOMODAKIS, Nikos. Unsupervised representation learning by predicting image rotations. In: **International Conference on Learning Representations,** 2018.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning.** Cambridge: MIT Press, 2016.

GRILL, Jean-Bastien et al. Bootstrap your own latent: a new approach to self-supervised learning. In: **Advances in Neural Information Processing Systems,** v. 33, p. 21271-21284, 2020.

HAN, Song; MAO, Huizi; DALLY, William J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. In: **International Conference on Learning Representations,** 2016.

HE, Kaiming; CHEN, Xinlei; XIE, Saining; LI, Yanghao; DOLLÁR, Piotr; GIRSHICK, Ross. Masked autoencoders are scalable vision learners. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.** New Orleans: IEEE, 2022. p. 16000-16009.

HE, Kaiming; FAN, Haoqi; WU, Yuxin; XIE, Saining; GIRSHICK, Ross. Momentum contrast for unsupervised visual representation learning. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.** Seattle: IEEE, 2020. p. 9729-9738.

HENDRYCKS, Dan; DIETTERICH, Thomas. Benchmarking neural network robustness to common corruptions and perturbations. In: **International Conference on Learning Representations,** 2019.

HINTON, Geoffrey E.; SALAKHUTDINOV, Ruslan R. Reducing the dimensionality of data with neural networks. **Science,** v. 313, n. 5786, p. 504-507, 2006.

HOUlsBY, Neil et al. Parameter-efficient transfer learning for NLP. In: **Proceedings of the 36th International Conference on Machine Learning.** PMLR, v. 97, p. 2790-2799, 2019.

HOWARD, Jeremy; RUDER, Sebastian. Universal language model fine-tuning for text classification. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.** Melbourne: ACL, 2018. p. 328-339.

HSU, Wei-Ning et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing,** v. 29, p. 3451-3460, 2021.

JAIN, Sarthak; WALLACE, Byron C. Attention is not explanation. In: **Proceedings of NAACL-HLT.** Minneapolis: ACL, 2019. p. 3543-3556.

JEAN, Neal et al. Combining satellite imagery and machine learning to predict poverty. **Science,** v. 353, n. 6301, p. 790-794, 2016.

JOSHI, Pratik et al. The state and fate of linguistic diversity and inclusion in the NLP world. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.** Online: ACL, 2020. p. 6282-6293.

KAPLAN, Jared et al. Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361,** 2020.

KINGMA, Diederik P.; WELLING, Max. Auto-encoding variational Bayes. In: **International Conference on Learning Representations,** 2014.

KITCHIN, Rob. The data revolution: big data, open data, data infrastructures and their consequences. London: SAGE, 2014.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature,** v. 521, p. 436-444, 2015.

LECUN, Yann; MISRA, Ishan. Self-supervised learning: the dark matter of intelligence. **Meta AI Research Blog,** 2021.

LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: **Advances in Neural Information Processing Systems,** v. 33, p. 9459-9474, 2020.

LIU, Yinhan et al. RoBERTa: a robustly optimized BERT pretraining approach. **arXiv preprint arXiv:1907.11692,** 2019.

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781,** 2013.

MISRA, Ishan; ZITNICK, C. Lawrence; HEBERT, Martial. Shuffle and learn: unsupervised learning using temporal order verification. In: **European Conference on Computer Vision.** Cham: Springer, 2016. p. 527-544.

MITCHELL, Margaret et al. Model cards for model reporting. In: **Proceedings of the Conference on Fairness, Accountability, and Transparency.** New York: ACM, 2019. p. 220-229.

NORTHCUTT, Curtis G.; JIANG, Lu; CHUANG, Isaac L. Confident learning: estimating uncertainty in dataset labels. **Journal of Artificial Intelligence Research,** v. 70, p. 1373-1411, 2021.

PETERS, Matthew E. et al. Deep contextualized word representations. In: **Proceedings of NAACL-HLT.** New Orleans: ACL, 2018. p. 2227-2237.

RADFORD, Alec et al. Language models are unsupervised multitask learners. **OpenAI Technical Report,** 2019.

RADFORD, Alec et al. Learning transferable visual models from natural language supervision. In: **Proceedings of the 38th International Conference on Machine Learning.** PMLR, v. 139, p. 8748-8763, 2021.

RAFFEL, Colin et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research,** v. 21, n. 140, p. 1-67, 2020.

RAJAGOPALAN, Shalini et al. Self-supervised learning for electrocardiogram classification. **Computing in Cardiology,** 2020.

RUDIN, Cynthia. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence,** v. 1, p. 206-215, 2019.

SCHWARTZ, Roy; DODGE, Jesse; SMITH, Noah A.; ETZIONI, Oren. Green AI. **Communications of the ACM,** v. 63, n. 12, p. 54-63, 2020.

STRUBELL, Emma; GANESH, Ananya; MCCALLUM, Andrew. Energy and policy considerations for deep learning in NLP. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.** Florence: ACL, 2019. p. 3645-3650.

TORRACO, Richard J. Writing integrative literature reviews. **Human Resource Development Review,** v. 4, n. 3, p. 356-367, 2005.

VASWANI, Ashish et al. Attention is all you need. In: **Advances in Neural Information Processing Systems,** v. 30, 2017.

VINCENT, Pascal; LAROCHELLE, Hugo; BENGIO, Yoshua; MANZAGOL, Pierre-Antoine. Extracting and composing robust features with denoising autoencoders. In: **Proceedings of the 25th International Conference on Machine Learning.** New York: ACM, 2008. p. 1096-1103.

WHITTEMORE, Robin; KNAFL, Kathleen. The integrative review: updated methodology. **Journal of Advanced Nursing,** v. 52, n. 5, p. 546-553, 2005.

YUE, Zhihan; WANG, Yujing; DUAN, Juanyong; YANG, Tianmeng; HUANG, Congrui; TONG, Yunhai; XU, Bixiong. TS2Vec: towards universal representation of time series. In: **Proceedings of the AAAI Conference on Artificial Intelligence,** v. 36, n. 8, p. 8980-8987, 2022.

ZHANG, J.; WANG, Y.; LIU, X.; CHEN, L. A survey on self-supervised learning: recent advances and applications. **Neurocomputing,** 2025.

ZHANG, Wei et al. Deep learning based prognostics and health management: a survey. **Reliability Engineering & System Safety,** v. 201, 106982, 2020.

---

[1] Doutor em Gestão de Projetos de Tecnologia da Informação e Doutorando em Engenharia Elétrica pela Universidade de Brasília (UnB).