**Inteligência Artificial Geral (AGI)**  
*Definição:*  
Refere-se a um conceito amplo de inteligência artificial que visa construir sistemas capazes de realizar qualquer tarefa intelectual que um ser humano pode executar. É frequentemente usado para descrever objetivos de longo prazo de criação de sistemas de IA altamente autônomos e flexíveis.

*Conceito-chave:*  
IA com capacidade ampla para executar múltiplas tarefas em diferentes domínios, ao contrário de sistemas de IA especializados.

*Exemplos:*  
• Sistemas Especialistas (ex: MYCIN), que usam lógica e regras predefinidas para tomar decisões.

**Aprendizado de Máquina (ML)**  
*Definição:*  
Subconjunto da IA que permite que sistemas aprendam com dados e façam previsões ou decisões sem programação explícita.

*Conceito-chave:*  
IA que opera por meio de padrões de dados e treinamento, em vez de instruções codificadas manualmente.

*Exemplos:*  
• Regressão Linear  
• Árvores de Decisão  
• Máquinas de Vetores de Suporte (algoritmos de aprendizado supervisionado para tarefas preditivas).

**Aprendizado Profundo (Deep Learning)**  
*Definição:*  
Subconjunto do aprendizado de máquina que usa redes neurais com múltiplas camadas (daí o termo "profundo") para modelar padrões complexos em dados.

*Conceito-chave:*  
Eficaz para tarefas como reconhecimento de imagem, processamento de linguagem natural e sistemas autônomos.

*Exemplos:*  
• Redes Neurais Convolucionais (CNNs) para processamento de imagens.  
• Redes Neurais Recorrentes (RNNs) para dados sequenciais (ex: séries temporais ou modelos de linguagem).

**IA Generativa (Generative AI)**  
*Definição:*  
Subconjunto do aprendizado profundo focado na criação de novos conteúdos (texto, imagens, música) a partir de dados aprendidos. Modelos de IA generativa geralmente são baseados em **modelos de base (foundation models)**, que são modelos grandes e pré-treinados adaptáveis a diversas tarefas.

*Conceito-chave:*  
Sistemas de IA que geram saídas originais com base em padrões aprendidos durante o treinamento. Modelos de base fornecem versatilidade para ajustes em tarefas específicas.

*Exemplos:*  
• ChatGPT (geração de texto)  
• DALL·E (geração de imagens)  
• DeepDream (geração visual artística).

**Modelos de Linguagem de Grande Escala (LLMs)**  
*Definição:*  
Tipo de IA generativa focado em entender e gerar texto semelhante ao humano. LLMs são treinados em grandes volumes de dados textuais para aprender padrões linguísticos, gramática e contexto.

*Conceito-chave:*  
LLMs geram texto coerente e de alta qualidade com base em *prompts* de entrada e podem realizar tarefas como sumarização, tradução e resposta a perguntas.

*Exemplos:*  
• GPT-4  
• BERT  
• T5  
• Claude 2 (usados para geração de texto, análise de sentimentos, sumarização).

**Conceitos Avançados**

**Aprendizado em Contexto (In-Context Learning)**  
*Definição:*  
Método para melhorar modelos de IA generativa ao adicionar dados e exemplos adicionais ao *prompt*, ajudando o modelo a resolver tarefas com mais eficiência.

**Tipos de Prompt**  
• **Prompt de Poucos Exemplos (Few-Shot):** Fornece alguns exemplos para guiar o comportamento do modelo.  
• **Prompt Zero-Shot:** Não fornece exemplos, solicitando que o modelo execute a tarefa sem orientação.  
• **Prompt One-Shot:** Fornece exatamente um exemplo.  
• **Modelo de Prompt (Prompt Template):** Formato pré-definido para padronizar a interação com modelos de IA.  
• **Prompt com Cadeia de Pensamento (Chain-of-Thought):** Incentiva o modelo a decompor o raciocínio em etapas.  
• **Ajuste de Prompt (Prompt Tuning):** Ajuste de *prompts* para melhorar o desempenho em tarefas específicas.

**Espaço Latente (Latent Space)**  
*Definição:*  
Conhecimento ou padrões codificados capturados por LLMs que armazenam relações entre dados.

*Uso:*  
Representa o entendimento interno de linguagem ou dados que os modelos de IA usam para gerar saídas.

**Embeddings**  
*Definição:*  
Representações numéricas (vetoriais) de *tokens* (palavras, frases) que capturam seu significado semântico.

*Aplicações:*  
• Busca semântica (ex: encontrar palavras com significado similar).  
• Armazenados em **bancos de dados vetoriais** para busca e recuperação eficientes.

**Tokens**  
*Definição:*  
Unidades básicas de texto (palavras, subpalavras ou caracteres) processadas por modelos de linguagem.

**Contexto de Janela (Context-Window)**  
*Definição:*  
Número máximo de *tokens* que um modelo pode processar de uma vez (incluindo entrada e saída).  
*Uso:*  
Se excedido, partes do texto podem ser truncadas. Isso afeta tarefas como geração de texto longo ou análise de documentos.

**Alucinações (Hallucinations)**  
*Definição:*  
Ocorrem quando um modelo gera informações incorretas ou sem sentido, mas que parecem plausíveis.

**Mitigações**  
• **Geração Aumentada por Recuperação (RAG):** Combina geração com dados externos para garantir precisão.  
• **Ajuste Fino (Fine-Tuning):** Treina o modelo com dados mais relevantes.  
• **Humano no Loop (HITL):** Inclui revisão humana em áreas críticas.

**Modelos Multimodais**  
*Definição:*  
Modelos que trabalham com múltiplos tipos de dados (texto, imagem, áudio) em um espaço compartilhado.

*Exemplos:*  
• Geração de legendas para imagens.  
• Criação de imagens a partir de descrições textuais.

**Métodos de Busca**  
• **Busca por Palavra-Chave:** Correspondência exata de termos.  
• **Busca Semântica:** Usa *embeddings* para entender o significado da consulta.

**Bancos de Dados Vetoriais na AWS**  
• **Amazon OpenSearch Service:** Suporta busca k-NN para vetores (útil para análise de logs e monitoramento em tempo real).  
• **Amazon Aurora PostgreSQL:** Compatível com a extensão pgvector para armazenamento eficiente de *embeddings*.  
• **Amazon Neptune ML:** Usa GNNs para previsões em dados gráficos.  
• **Amazon MemoryDB:** Armazenamento de vetores com resposta em milissegundos.  
• **Amazon DocumentDB:** Busca vetorial compatível com MongoDB.

**Pipeline de Aprendizado de Máquina (ML)**  
Processo sistemático para construir, treinar e implantar modelos de ML:

1. **Identificar Objetivo de Negócio**
2. **Definir Problema de ML**
3. **Coletar Dados**
4. **Pré-Processar Dados**
5. **Engenharia de Features**
6. **Treinar, Ajustar, Avaliar**
7. **Implantar**
8. **Monitorar**

**1. Identificar Objetivo de Negócio**  
*Descrição:*  
Definir critérios de sucesso e alinhar stakeholders para garantir que o projeto de ML atenda aos objetivos do negócio.

*Atividades-chave:*  
• Estabelecer métricas claras de sucesso.  
• Alinhar com stakeholders em toda a organização.

**2. Definir o Problema de ML**  
*Descrição:*  
Definir o problema de ML, entradas, saídas e métricas, considerando viabilidade e análise de custo-benefício.

*Atividades-chave:*  
• Identificar entradas, saídas, requisitos e métricas de desempenho.  
• Realizar análise de custo-benefício para avaliar viabilidade.

*Opções de Modelo:*  
• **Serviço de IA/ML Hospedado** (ex: AWS Comprehend, Forecast, Personalize): Sem necessidade de treinamento.  
• **Modelos Pré-treinados** (ex: Amazon Bedrock, SageMaker JumpStart): Ajuste fino com seus dados.  
• **Modelo Totalmente Personalizado:** Construído e treinado do zero.

**3. Coletar Dados**  
*Descrição:*  
Coletar e preparar os dados necessários para treinar o modelo.

*Atividades-chave:*  
• Identificar fontes de dados (ex: bancos de dados, data lakes, APIs externas).  
• Ingerir e rotular dados.

*Onde os Dados são Armazenados:*  
• Dados coletados para ML são armazenados no **Amazon S3** (incluindo dados de sistemas internos e de terceiros via AWS Data Exchange).

*Ferramentas:*  
• **AWS Glue:** Processos ETL para transformar dados antes do armazenamento no S3.  
• **SageMaker Ground Truth:** Rotulagem humana de dados ambíguos.  
• **AWS Data Exchange:** Acesso seguro a dados de terceiros.

**4. Pré-processar Dados**  
*Descrição:*  
Limpar e preparar dados para treinamento.

*Atividades-chave:*  
• Realizar Análise Exploratória de Dados (EDA).  
• Limpar dados (remover duplicatas, preencher valores ausentes, anonimizar PII).  
• Dividir dados em conjuntos de treinamento (80%), validação (10%) e teste (10%).

*Como Limpar Dados:*  
• **AWS Glue:** Transformações personalizadas com Python/Spark.  
• **Macie para PII:** Detecta e anonimiza dados sensíveis no S3.  
• **AWS Glue DataBrew:** Preparação visual de dados com regras reutilizáveis.  
• **SageMaker Canvas:** Transformações sem código com fluxo visual.

**5. Engenharia de Features**  
*Descrição:*  
Selecionar e criar features para melhorar o desempenho do modelo.

*Atividades-chave:*  
• **Seleção de Features:** Identificar as mais relevantes com base no domínio.  
• **Criação de Features:** Transformar dados brutos em novas features (ex: escalonamento, codificação categórica).  
• **Redução de Dimensionalidade:** Técnicas como PCA para simplificar dados.

*Ferramentas:*  
• **SageMaker Feature Store:** Armazenar e gerenciar features centralizadas.

**6. Treinar, Ajustar e Avaliar o Modelo**  
*Descrição:*  
Treinar o modelo, ajustar hiperparâmetros e avaliar o desempenho.

*Atividades-chave:*  
• Treinar iterativamente e ajustar parâmetros.  
• Avaliar métricas como MSE, Precisão, Recall e F1-Score.

*Parâmetros de Inferência (Amazon Bedrock):*  
• **Temperatura:** Controla aleatoriedade (valores altos = criatividade; baixos = previsibilidade).  
• **Top-P:** Limita tokens por probabilidade cumulativa.  
• **Comprimento da Resposta:** Define tamanho máximo da saída.

*Métricas de Erro:*  
• **MSE/RMSE:** Para problemas de regressão (ex: previsão de preços).  
• **Precisão/Recall:** Para classificação (ex: detecção de spam).  
• **F1-Score:** Equilíbrio entre Precisão e Recall.

*Problemas no Treinamento:*  
• **Overfitting:** Solução = Dados mais diversificados.  
• **Underfitting:** Solução = Mais épocas de treinamento.  
• **Viés:** Solução = Dados representativos e restrições de fairness.

*Ajuste Fino (Fine-Tuning):*  
• **Bedrock:** Adaptação de modelos pré-treinados a tarefas específicas.  
• **Domain Adaptation:** Personalização para domínios como jurídico ou médico.

*Ferramentas:*  
• **SageMaker Training Jobs:** Gerenciar processos de treinamento.  
• **Automatic Model Tuning (AMT):** Ajuste automático de hiperparâmetros.

**7. Implantar o Modelo**  
*Descrição:*  
Implantar o modelo treinado para fazer previsões.

*Opções de Implantação (SageMaker):*  
• **Inferência em Tempo Real:** Para baixa latência e tráfego sustentado.  
• **Batch Transform:** Processamento assíncrono de grandes volumes.  
• **Serverless Inference:** Escalabilidade automática para tráfego intermitente.

*Tipos de Instâncias:*  
• **Inf1:** Custo-efetivo para inferência com chips AWS Inferentia.  
• **P4:** Alta performance com GPUs NVIDIA A100.

*Ferramentas:*  
• **SageMaker Endpoints:** Disponibilizar modelos para inferência.  
• **AWS Lambda:** Integração com Bedrock Agents para automação.

**8. Monitorar o Modelo**  
*Descrição:*  
Monitorar continuamente o desempenho e detectar *drift* (mudanças nos dados ou conceitos).

*Tipos de Drift:*  
• **Data Drift:** Mudança nos dados de entrada (ex: nova demografia).  
• **Concept Drift:** Mudança na relação entrada-saída (ex: novos padrões de fraude).

*Ferramentas:*  
• **SageMaker Model Monitor:** Alertas automatizados via CloudWatch.

**MLOps e Automação**  
*Descrição:*  
Aplicar princípios de DevOps para gerenciar modelos de ML com automação e versionamento.

*Atividades-chave:*  
• Automatizar implantação, monitoramento e retreinamento.  
• Implementar CI/CD para atualizações de modelos.

*Ferramentas:*  
• **SageMaker Pipelines:** Gerenciar fluxos de ML de ponta a ponta.  
• **AWS CodePipeline:** Automatizar build, teste e deploy.

**Governança de Modelos e Explicabilidade**

**Descrição:**  
Garantir transparência, responsabilidade e conformidade regulatória para modelos de ML, tornando seu comportamento interpretável para as partes interessadas.

**Atividades-chave:**  
• Implementar estruturas de governança para rastrear uso e linhagem de modelos.  
• Usar ferramentas para detectar e mitigar viés, garantir justiça e explicar previsões.  
• Fornecer documentação clara do histórico e desempenho do modelo para auditorias.

**Ferramentas:**  
• **SageMaker Clarify:** Detecta viés, explica previsões e aumenta a transparência.  
• **SageMaker Model Cards:** Cria documentação para modelos treinados, incluindo métricas de desempenho e uso pretendido.  
• **ML Governance do SageMaker:** Oferece ferramentas para controle e visibilidade de modelos de ML.  
• **SageMaker ML Lineage Tracking:** Captura todo o fluxo de trabalho para rastrear linhagem de modelos.  
• **Glue DataBrew:** Simplifica governança de dados com preparação visual e regras de qualidade.  
• **AWS Audit Manager:** Automatiza auditorias de serviços AWS para conformidade contínua.  
• **AWS Artifact:** Fornece acesso sob demanda a relatórios de conformidade.  
• **AWS AI Service Cards:** Melhora a transparência com informações sobre práticas de IA responsável.

**Otimização de Custo e Desempenho**

**Descrição:**  
Otimizar uso de recursos e desempenho do modelo sem aumentar custos.

**Atividades-chave:**  
• Usar treinamento com spot gerenciado para reduzir custos.  
• Selecionar tipos de instância adequados e usar autoescalamento.  
• Monitorar recursos para detectar ineficiências.

**Ferramentas:**  
• **AWS Trusted Advisor:** Recomenda melhorias de custo/desempenho.  
• **SageMaker Managed Spot Training:** Reduz custos de treinamento usando capacidade ociosa do EC2.  
• **SageMaker Profiler:** Identifica uso ineficiente de recursos durante o treinamento.  
• **Amazon Inspector:** Automatiza avaliações de segurança em aplicações de ML.

**Aprendizado Contínuo e Retreinamento**

**Descrição:**  
Retreinar modelos continuamente para adaptar-se a novos dados e evitar degradação de desempenho.

**Atividades-chave:**  
• Agendar retreinamento regular com base em novos dados.  
• Usar ferramentas para detectar quedas de desempenho e iniciar fluxos de retreinamento automático.  
• Lidar com *drift* de modelo monitorando desvio de conceito e dados.

**Ferramentas:**  
• **SageMaker Model Monitor:** Detecta desvio de dados e aciona retreinamento.  
• **SageMaker Pipelines:** Automatiza processos de retreinamento de ponta a ponta.

:

**Segurança**

**Descrição:**  
Implementar práticas de segurança para proteger modelos de ML, dados e infraestrutura.

**Atividades-chave:**  
• **Princípio do Menor Privilégio:** Garantir que políticas e funções do IAM concedam apenas as permissões necessárias.  
• **PrivateLink e VPC Endpoints:** Restringir o acesso à internet do SageMaker usando endpoints privados.  
• **Criptografia em Repouso e em Trânsito:** Criptografia padrão via AWS KMS (Key Management Service).  
• **S3 Block Public Access:** Bloquear acesso público a dados de modelos no Amazon S3.  
• **AWS IAM Identity Center:** Centralizar gerenciamento de identidades em múltiplas contas AWS.

**Ferramentas:**  
• **AWS Config:** Monitora e registra alterações de configuração em recursos AWS.  
• **AWS CloudTrail:** Registra chamadas de API e atividades de usuários para auditoria.  
• **Amazon Inspector:** Verifica vulnerabilidades em ambientes de ML.  
• **AWS Audit Manager:** Automatiza auditorias e gera relatórios de conformidade.  
• **AWS Artifact:** Fornece acesso a documentos de conformidade (ex: PCI-DSS, HIPAA).  
• **SageMaker Role Manager:** Gerencia permissões de recursos do SageMaker.

**Serviços de IA Gerenciados da AWS**

**Visão Computacional:**  
• **AWS Rekognition:**

* Comparação facial, detecção de objetos e moderação de conteúdo.
* Caso de uso: Identificação de material violento ou inadequado.

**Análise de Texto e Documentos:**  
• **AWS Textract (OCR):**

* Converte imagens digitalizadas em texto (ex: documentos escaneados).  
  • **Amazon Comprehend (NLP):**
* Extrai frases-chave, entidades, sentimentos e detecta dados PII.  
  • **AWS Intelligent Document Processing (IDP):**
* Combina serviços como Textract e Comprehend para processar documentos não estruturados (PDFs, faturas, contratos).

**IA de Linguagem:**  
• **Amazon Lex:**

* Cria chatbots com interfaces de voz e texto.  
  • **Amazon Transcribe:**
* Converte fala em texto (ex: legendas de áudio).  
  • **Amazon Polly:**
* Converte texto em fala natural para engajamento do usuário.

**Experiência do Cliente:**  
• **Amazon Kendra:**

* Busca inteligente em documentos com base em IA.  
  • **Amazon Personalize:**
* Oferece recomendações personalizadas (ex: "você também pode gostar").  
  • **Amazon Translate:**
* Tradução de idiomas para interação global.

**Métricas de Negócio:**  
• **Amazon Forecast:**

* Previsão de séries temporais (ex: estoque, demanda).  
  • **Amazon Fraud Detection:**
* Detecta fraudes em transações online e criação de contas.

**Amazon Q:**  
• **Amazon Q Business:**

* Assistente de IA generativo para automação de fluxos de trabalho empresariais.  
  • **Amazon Q Developer:**
* Gera código, realiza varredura de segurança e auxilia desenvolvedores.  
  • **Amazon Q in QuickSight:**
* Permite consultas em linguagem natural no Amazon QuickSight.  
  • **Amazon Q in Connect:**
* Automatiza atendimento ao cliente no Amazon Connect.  
  • **Amazon Q in AWS Supply Chain:**
* Otimiza gestão da cadeia de suprimentos com insights baseados em IA.

**Amazon SageMaker**

**Descrição:**  
Serviço de machine learning integrado que permite que desenvolvedores e cientistas de dados criem, treinem e implantem modelos de ML em escala. Os usuários podem criar modelos personalizados do zero ou usar e ajustar modelos existentes por meio do SageMaker JumpStart. Oferece mais controle do que serviços de IA de alto nível (como AWS Rekognition), permitindo personalização detalhada para atender requisitos específicos.

**SageMaker Studio**

* Ambiente de desenvolvimento integrado (IDE) para machine learning.
* Oferece uma única interface para preparar dados, construir modelos, treinar, ajustar e implantar.
* Recursos:  
  • Notebooks Jupyter para desenvolvimento de código.  
  • Ferramentas de depuração integradas.  
  • Gerenciamento de experimentos.  
  • Colaboração em tempo real.  
  • Acesso a capacidades do SageMaker (monitoramento de modelos, preparação de dados).

**Processo de Treinamento**

Elementos-chave para configurar e gerenciar trabalhos de treinamento:

* **Local dos Dados de Treinamento:** Armazenados no Amazon S3 (acessados via URLs do S3).
* **Instâncias de Computação ML:** Usa instâncias EC2 para escalabilidade.
* **Imagens de Treinamento:** Contêineres Docker projetados para machine learning.
* **Hiperparâmetros:** Parâmetros que guiam o aprendizado (ex: taxa de aprendizado, tamanho do *batch*).
* **Bucket de Saída S3:** Armazena artefatos do modelo treinado.

**Recursos do SageMaker**

* **SageMaker Feature Store:** Repositório central para armazenar, recuperar e compartilhar *features*.
* **SageMaker Model Registry:** Gerencia versões e metadados de modelos.
* **SageMaker Pipelines:** Orquestração de fluxos de trabalho repetíveis para construção, treinamento e implantação.
* **SageMaker Model Monitor:**  
  • Detecta *drift* de dados usando regras pré-definidas ou personalizadas.  
  • Envia resultados para o CloudWatch.  
  • Automatiza medidas corretivas.
* **SageMaker Ground Truth:** Rotulagem de dados por humanos para garantir qualidade.
* **SageMaker Canvas:** Ferramenta visual sem código para construção de modelos.
* **SageMaker JumpStart:** Acesso a modelos pré-treinados e implantação simplificada.
* **SageMaker Clarify:** Detecta viés e explica previsões para transparência.
* **SageMaker Role Manager:** Gerencia permissões de recursos e serviços do SageMaker.
* **SageMaker Model Cards:** Cria documentação transparente para modelos treinados.
* **SageMaker ML Lineage Tracking:** Rastreia a linhagem de modelos para governança e reprodução.
* **SageMaker Model Dashboard:** Interface unificada para gerenciar e monitorar atividades de modelos.
* **SageMaker Data Wrangler:** Simplifica preparação de dados (limpeza, transformação, visualização).
* **SageMaker Experiments (MLflow com Amazon SageMaker):**  
  • Rastreia, organiza e compara experimentos iterativos de ML.
* **SageMaker Autopilot:** Automatiza construção, treinamento e ajuste de modelos com visibilidade total.
* **Amazon Augmented AI (A2I):** Adiciona revisão humana a previsões de baixa confiança.
* **SageMaker Managed Spot Training:** Reduz custos usando capacidade ociosa do EC2.
* **SageMaker Profiler:** Identifica ineficiências de recursos em trabalhos de treinamento.

**Amazon Bedrock**

**Descrição:**  
Serviço gerenciado e *serverless* que fornece acesso a modelos de base (*foundation models*) de empresas líderes via API. Focado em aplicativos de IA generativa, com segurança e privacidade.

**Preços**

* **Pagamento conforme o uso:** Cobrado por tokens de entrada/saída durante inferência.
* **Modelos ajustados:** Custo adicional por *Provisioned Throughput* (capacidade reservada).

**Recursos**

* **Escolha de Modelos:**  
  • Modelos da AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI e **Amazon Titan** (exclusivos da AWS).
* **Modelos Titan:**  
  • Modelos de base pré-treinados e de alto desempenho para diversos casos de uso.
* **Personalização:**  
  • Ajuste fino (*fine-tuning*) e RAG (*Retrieval Augmented Generation*) para melhorar relevância.
* **Avaliação de Modelos:**  
  • Compare modelos com métricas personalizadas (ex: precisão, robustez, toxicidade).  
  • Integração com SageMaker Clarify para verificar viés e explicabilidade.
* **Bedrock Knowledge Bases:**  
  • Usa RAG para buscar dados de fontes privadas, melhorando respostas.  
  • Gerencia ingestão de dados, recuperação e aumento de *prompts*.
* **Bedrock Agents:**  
  • Cria agentes para tarefas multietapa (ex: consultas de clientes, processamento de pedidos).
* **Serverless:**  
  • Elimina gerenciamento de infraestrutura.
* **Proteções de Segurança:**  
  • Restringe conteúdos inadequados (ex: recomendações financeiras).
* **PartyRock:**  
  • Ambiente para criar aplicativos de IA generativa sem código.

**AWS Glue**

**Descrição:**  
Serviço ETL gerenciado e otimizado para nuvem, usado para preparar dados para análise e IA.

**Recursos**

* **AWS Glue ETL Service:**  
  • Serviço ETL em nuvem com transformações integradas (ex: remover duplicatas, preencher valores ausentes).  
  • **Exemplo de fluxo:**  
  AWS Kinesis Data Streams → Trabalho ETL do AWS Glue → CSV para Parquet → S3 Data Lake.
* **AWS Glue Data Catalog:**  
  • Repositório centralizado para metadados (esquemas, não dados) com classificadores integrados.
* **AWS Glue DataBrew:**  
  • Ferramenta visual para preparação de dados e definição de regras de qualidade (*recipes*).
* **AWS Glue Data Quality:**  
  • Detecta anomalias e recomenda regras para garantir dados limpos.

**Aprendizado de Máquina Tradicional vs. Aprendizado Profundo**

| **Categoria** | **Aprendizado de Máquina Tradicional** | **Aprendizado Profundo** |
| --- | --- | --- |
| **Complexidade da Tarefa** | Tarefas bem definidas | Tarefas complexas |
| **Tipo de Dados** | Dados estruturados/rotulados | Dados não estruturados (imagens, vídeo, texto) |
| **Metodologia** | Resolve problemas por estatística e matemática | Utiliza redes neurais |
| **Tratamento de Features** | Seleção e extração manual de features | Features aprendidas automaticamente pelo modelo |
| **Custo** | Menos custoso | Custos mais altos devido a demandas computacionais |

**Tipos de Aprendizado de Máquina**

| **Tipo de Aprendizado** | **Descrição** | **Desafios** | **Ferramentas AWS** | **Casos de Uso Comuns** |
| --- | --- | --- | --- | --- |
| **Aprendizado Supervisionado** | Usa dados pré-rotulados para treinar modelos. | Rotular dados pode ser desafiador. | SageMaker Ground Truth | Classificação de imagens, detecção de spam |
| **Aprendizado Não Supervisionado** | Trabalha com dados não rotulados para encontrar padrões. | Requer métodos para interpretar padrões. | Nenhuma específica | Clusterização, detecção de anomalias, treinamento inicial de LLMs |
| **Aprendizado por Reforço** | Aprende por tentativa e erro para maximizar recompensas. | Requer ambiente para interação do agente. | AWS DeepRacer | Jogos, robótica, decisões em tempo real |

**Tipos de Modelos de Difusão**

| **Tipo de Modelo de Difusão** | **Descrição** | **Quando Usar** | **Notas** |
| --- | --- | --- | --- |
| **Difusão Direta** | Adiciona ruído aos dados progressivamente. | Pouco utilizado (adiciona ruído). |  |
| **Difusão Reversa** | Reconstrói dados originais a partir do ruído. | Criação de imagens detalhadas a partir de entradas distorcidas. | Ferramentas de restauração de imagem |
| **Difusão Estável** | Opera em espaço latente reduzido, não diretamente em pixels. | Mais eficiente que a Difusão Reversa. | Midjourney, DALL-E |

**Métodos de Inferência do Amazon SageMaker**

| **Tipo de Inferência** | **Implantado em** | **Características** |
| --- | --- | --- |
| **Batch** | EC2 | Custo-efetivo para trabalhos grandes e infrequentes. |
| **Assíncrono** | EC2 | Adequado para aplicações não sensíveis ao tempo e cargas úteis grandes. |
| **Serverless** | Lambda | Tráfego intermitente, períodos sem tráfego, autoescalonamento integrado. |
| **Tempo Real** | EC2 | Previsões ao vivo, tráfego sustentado, baixa latência, desempenho consistente. |

**Tipos de Dados de Treinamento para ML/IA**

| **Tipo de Dados** | **Exemplo de Fonte de Dados AWS** | **Exemplo Prático** |
| --- | --- | --- |
| **Estruturado** | Dados SQL armazenados no RDS e movidos para S3 | Informações de clientes em tabelas relacionais |
| **Semi-estruturado** | Dados no DynamoDB ou DocumentDB movidos para S3 | Logs JSON de atividades de usuário |
| **Não estruturado** | Objetos e arquivos armazenados diretamente no S3 | Imagens, vídeos, documentos PDF |
| **Séries Temporais** | Dados com carimbo de tempo armazenados no S3 | Dados de dispositivos IoT, dados do mercado de ações |

**Métricas de Desempenho para IA Generativa**

| **Métrica** | **Explicação** |
| --- | --- |
| **ROUGE (Recall Oriented Understudy for Gisting Evaluation)** | Mede sobreposição entre texto gerado e referência, útil para resumos. |
| **BLEU (Bilingual Evaluation Understudy)** | Avalia precisão de tradução comparando n-grams entre saídas e referências. |
| **GLUE (General Language Understanding Evaluation)** | Avalia desempenho do modelo em múltiplas tarefas de compreensão de linguagem. |
| **HELM (Holistic Evaluation of Language Models)** | Oferece avaliação ampla e específica de capacidades de modelos de linguagem. |
| **MMLU (Massive Multitask Language Understanding)** | Testa conhecimento do modelo em diversos domínios e tópicos. |
| **BIG-bench (Beyond the Imitation Game Benchmark)** | Avalia modelos em tarefas criativas e complexas não cobertas por benchmarks padrão. |
| **Perplexidade** | Mede quão bem o modelo prevê a probabilidade do próximo token ou palavra. |

**Modelos de IA Generativa**

| **Modelo de IA Generativa** | **Exemplos** | **Caso de Uso/Principais Aplicações** |
| --- | --- | --- |
| **GANs (Redes Adversárias Generativas)** | StyleGAN, CycleGAN, ProGAN | Geração de imagens, síntese facial, criação de vídeo |
| **VAEs (Autoencoders Variacionais)** | VAE de Kingma & Welling, Beta-VAE | Remoção de ruído em imagens, detecção de anomalias |
| **Transformers** | GPT-4, BERT, T5 | Geração de texto, tradução, criação de conteúdo |
| **RNNs (Redes Neurais Recorrentes)** | LSTMs, GRUs | Dados sequenciais, previsão de séries temporárias |
| **Aprendizado por Reforço para Tarefas Generativas** | AlphaGo, DQN, OpenAI Five | IA para jogos, controle autônomo, otimização de tarefas |
| **Difusão** | Stable Diffusion, DALL·E 2, Imagen | Geração de imagens e texto-para-imagem |
| **Modelos Baseados em Fluxo** | Glow, RealNVP | Geração de imagens de alta qualidade, estimativa de densidade |

Aqui está a tabela traduzida e formatada no estilo de um documento Word (modo de exibição Markdown para visualização aqui). Também adicionei o conceito de **Top-K** no início da tabela:

**Folha de Estudo - Conceitos de IA/ML**

| **Termo/Conceito** | **Descrição** |
| --- | --- |
| **Top-K** | Limita as escolhas do modelo aos *K* tokens mais prováveis. Reduz a aleatoriedade, mantendo apenas as opções mais prováveis no momento da geração. |
| **Top-P** | Valor entre 0 e 1 — considera a menor quantidade de tokens cuja soma das probabilidades ultrapassa P. Controla aleatoriedade mantendo naturalidade. |
| **Temperatura** | Controla a aleatoriedade — valores maiores geram saídas mais criativas; valores baixos tornam a resposta mais previsível. |
| **Épocas (Epochs)** | Número de passagens completas pelo conjunto de dados de treino — mais épocas podem melhorar o modelo, mas exageros causam sobreajuste. |
| **AWS Rekognition** | Serviço de visão computacional — usado para detecção de objetos, análise facial e moderação de conteúdo. |
| **AWS Textract** | Serviço de OCR — extrai texto e estrutura de documentos escaneados (PDFs, formulários, etc). |
| **Amazon Comprehend** | Extrai frases-chave, entidades, sentimentos e dados pessoais (PII) de textos. |
| **Processamento Inteligente de Documentos (IDP)** | Automatiza o processamento de documentos não estruturados combinando Textract, Comprehend e A2I. |
| **Amazon Lex** | Construtor de chatbots com entrada por texto ou voz em linguagem natural. |
| **Amazon Transcribe** | Serviço de transcrição automática de voz para texto. |
| **Amazon Polly** | Converte texto em fala com vozes naturais. |
| **Amazon Kendra** | Motor de busca inteligente com compreensão semântica de documentos corporativos. |
| **Amazon Personalize** | Gera recomendações personalizadas de produtos. |
| **Amazon Translate** | Serviço de tradução automática de idiomas. |
| **Amazon Forecast** | Realiza previsões de séries temporais — útil para prever demanda, estoques etc. |
| **Amazon Fraud Detection** | Detecta atividades fraudulentas como transações suspeitas e invasões de conta. |
| **Amazon Q Business** | Assistente de IA generativa para dados e fluxos de trabalho empresariais. |
| **Amazon Macie** | Detecta e anonimiza dados sensíveis (PII) para segurança de dados. |
| **SageMaker Clarify** | Detecta viés e fornece explicações para modelos de machine learning. |
| **SageMaker Ground Truth** | Realiza rotulagem humana de dados para treinar modelos. |
| **Amazon Augmented AI (A2I)** | Permite revisão humana de previsões com baixa confiança durante a inferência. |
| **AWS Data Exchange** | Acesso seguro a conjuntos de dados de terceiros. |
| **Transformações AWS Glue** | ETL — remove duplicatas, preenche valores ausentes e transforma dados. |
| **SageMaker JumpStart** | Catálogo com modelos pré-treinados e implantações com um clique. |
| **SageMaker Canvas** | Ferramenta sem código para treinar e implantar modelos de machine learning. |
| **Fine-Tuning (Ajuste Fino)** | Ajusta os pesos de um modelo com dados rotulados para melhorar a performance em tarefas específicas. |
| **Ajuste Fino com Adaptação de Domínio** | Personaliza o modelo para setores como jurídico ou médico usando pequenos conjuntos de dados. |
| **Ajuste Fino com Instruções** | Treina o modelo para melhorar seu desempenho em tarefas como classificação, resumo, etc. |
| **Continued Pretraining** | Expande a base de conhecimento de um modelo com dados não rotulados. |
| **Tuning Automático de Modelos (AMT)** | Otimiza hiperparâmetros automaticamente para melhorar desempenho. |
| **Data Drift (Deslocamento de Dados)** | Os dados mudam, mas a relação entre entrada e saída permanece (ex: novos perfis de usuários). |
| **Concept Drift (Deslocamento de Conceito)** | A relação entre entrada e saída muda — o modelo passa a errar, como em novos tipos de fraude. |
| **AWS Trusted Advisor** | Recomendações para reduzir custos e melhorar segurança, desempenho e confiabilidade. |
| **Amazon Inspector** | Avaliação automatizada de segurança para cargas de trabalho na AWS. |
| **AWS PrivateLink** | Cria conexões privadas e seguras entre serviços na nuvem. |
| **AWS Config** | Monitora alterações de configuração em recursos da AWS para conformidade. |
| **AWS CloudTrail** | Registra chamadas da API para auditoria e rastreabilidade. |
| **Bedrock Guardrails** | Restringe saídas inadequadas de modelos de fundação e controla conteúdos sensíveis. |
| **Postgres (Aurora ou RDS)** | Banco de dados relacional com suporte a vetores para buscas por similaridade. |
| **Amazon DocumentDB** | Armazena documentos JSON, compatível com MongoDB e com busca vetorial. |
| **Amazon Neptune** | Banco de dados de grafos com suporte a busca vetorial. |
| **Amazon Neptune ML** | Usa redes neurais em grafos (GNNs) para prever resultados a partir de dados conectados. |
| **Amazon MemoryDB** | Banco de dados em memória com busca vetorial ultrarrápida. |
| **Amazon OpenSearch Service** | Serviço de busca com suporte a vetores para encontrar itens similares. |
| **MSE (Erro Quadrático Médio)** | Média das diferenças ao quadrado entre valores reais e previstos — quanto menor, melhor. |
| **RMSE (Raiz do Erro Quadrático Médio)** | Raiz quadrada do MSE — mais fácil de interpretar; valores menores indicam melhor desempenho. |