

Secretaria
de Planejamento, Gestão
e Desenvolvimento
Regional



GOVERNO DE
**PER
NAM
BUCO**
ESTADO DE MUDANÇA

Fundamentos da Ciência de Dados

Curso Ciência de Dados para a Gestão Pública
Júlia Barrêto

12 de Maio de 2025





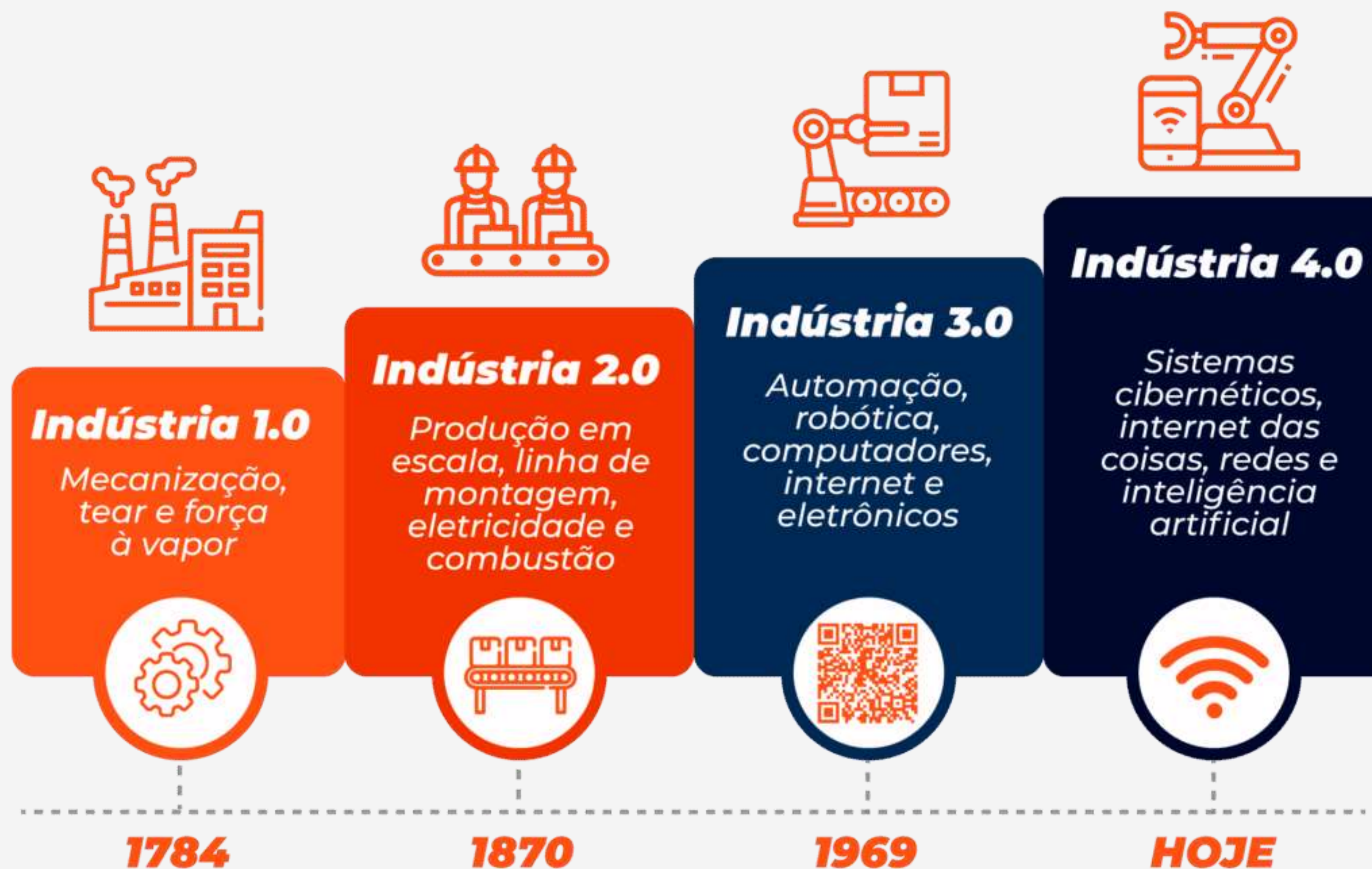
Fundamentos da Ciência de Dados

1. O que é Ciência de Dados?
2. Dado, Informação e Conhecimento
3. Relações entre Computação, Matemática e Negócio
4. Atividades, Papéis e Produtos em Ciência de Dados
5. Cultura Analítica, Capacidade Analítica e Maturidade Analítica

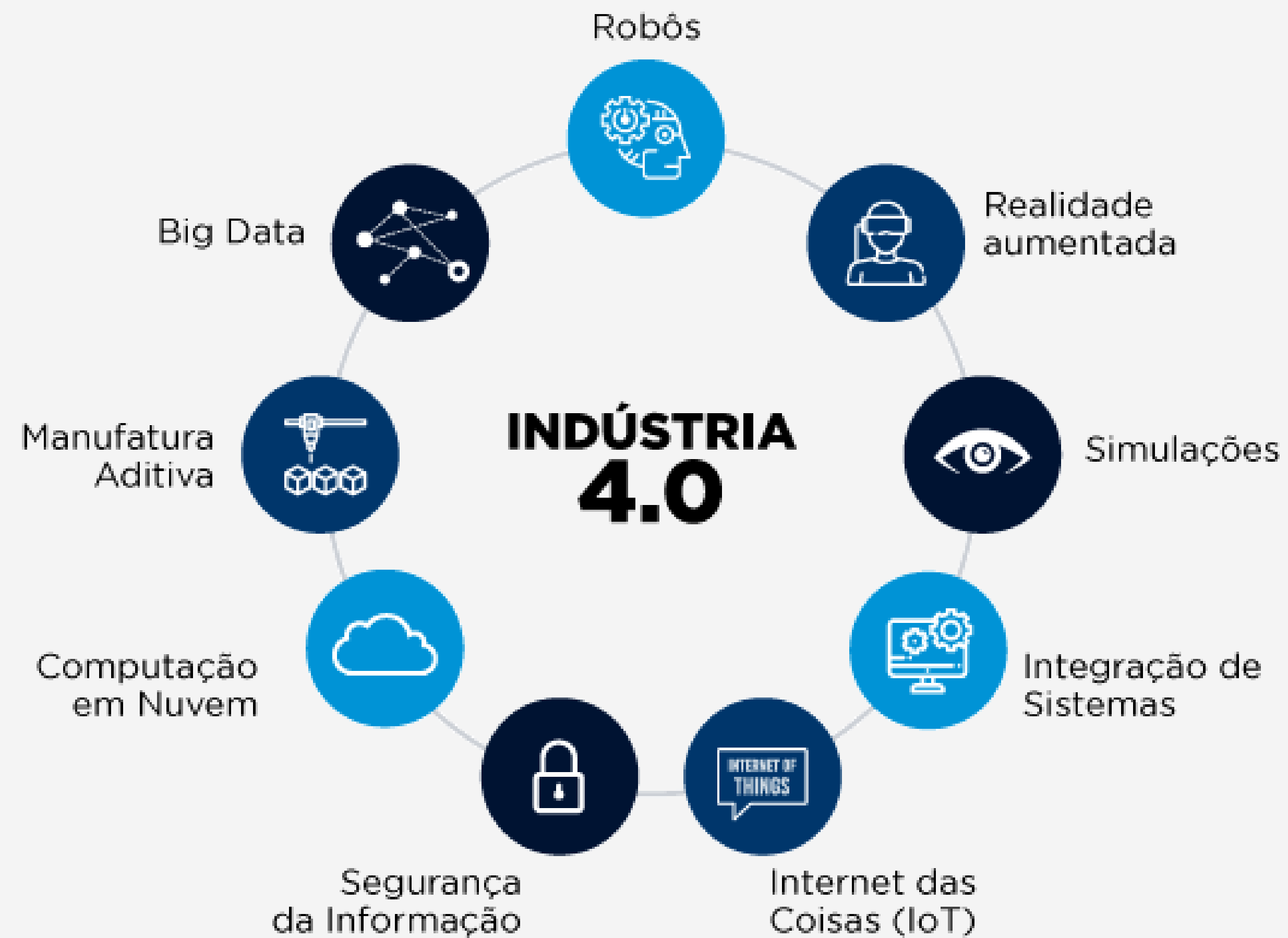


Momento atual

- Indústria 4.0 -> 4ª Revolução Industrial
- Fabricação inteligente
- Transformação digital do setor
- Tomada de decisões em tempo real
- Aumento da produtividade, flexibilidade e agilidade
- Melhoria no processo e resultado de produtos
- Fabricantes estão integrando novas tecnologias, incluindo Internet das coisas (IoT), computação e análise de dados em nuvem, IA e aprendizado de máquina
- Maior automação, manutenção preditiva, auto-otimização de melhorias de processos





Fonte: Reprodução da internet (<https://49educacao.com.br/industria-4-0/>)



Fonte: Reprodução da internet (<https://www.tecnicon.com.br/blog/476-4-exemplos-praticos-da-adocao-da-Industria-4-0-nas-fabricas>)



Dados: o novo petróleo

- Nova economia -> caracteriza-se pela criação de novos modelos de negócio, tecnologias disruptivas, foco na experiência do cliente e pela importância da colaboração e do conhecimento
 - Matéria-prima para tomada de decisão em grandes corporações
 - Precisa ser refinado para ser útil
 - Extremamente valioso
 - Dados são abundantemente e velozmente produzidos
 - Mineração dos dados: é o processo de extração de conhecimento em grandes quantidades de dados (Han e Kamber).
- 
- 

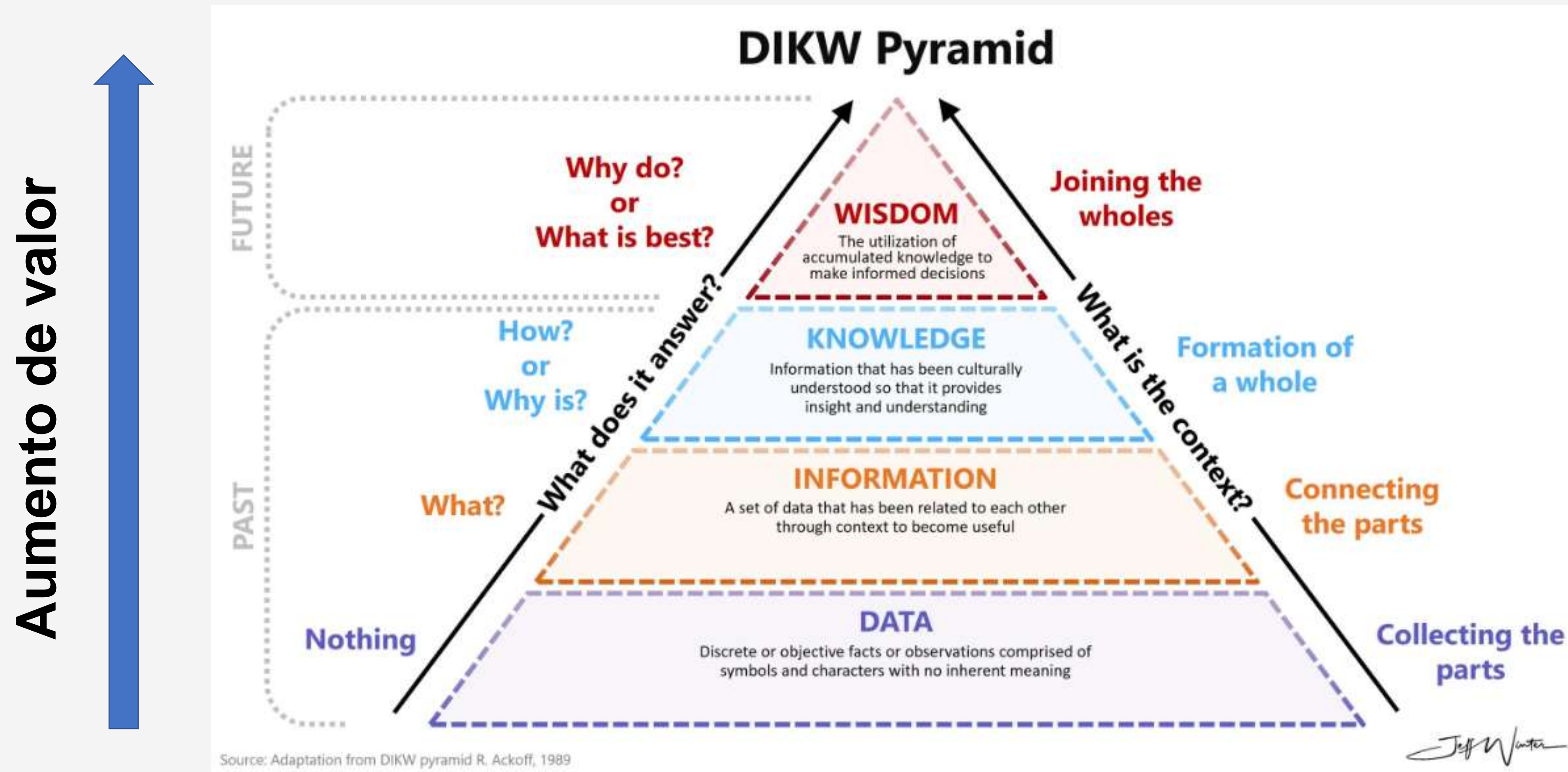


Dados: o novo petróleo

- Problema dos dados: extrair informações de poucos dados, mas hoje a dificuldade é saber o que é útil ou não
- Decompor problemas ajuda a entender quais dados são necessários para construir uma boa análise.
- A ubiquidade das oportunidades de dados (Provost, F. & Fawcett, T. (2013))
 - Evolução nas tecnologias
 - Aumento do volume de dados
 - Melhoria nas técnicas de mineração dos dados

Pirâmide DIKW

- Modelo que ilustra como dados, informações, conhecimento e sabedoria se relacionam e evoluem





Dado, informação e conhecimento

Dado

Elemento bruto, isolado, sem contexto ou interpretação. Sozinho, ele não tem significado claro. Ex: "150", "PE", "Janeiro"

Informação

Surge quando os dados são organizados e contextualizados, oferecendo um significado mais claro. Ex: "Em Pernambuco, no mês de janeiro, foram registrados 150 casos de MVI (Morte Violenta Intencional)"

Conhecimento

Aparece quando interpretamos a informação, cruzamos com outras informações e tiramos conclusões que podem orientar ações ou decisões. Adição de experiência. Ex: A taxa de 150 MVI em Pernambuco, em janeiro, representa uma queda em relação ao ano anterior, indicando que as políticas de segurança pública podem estar surtindo efeito."



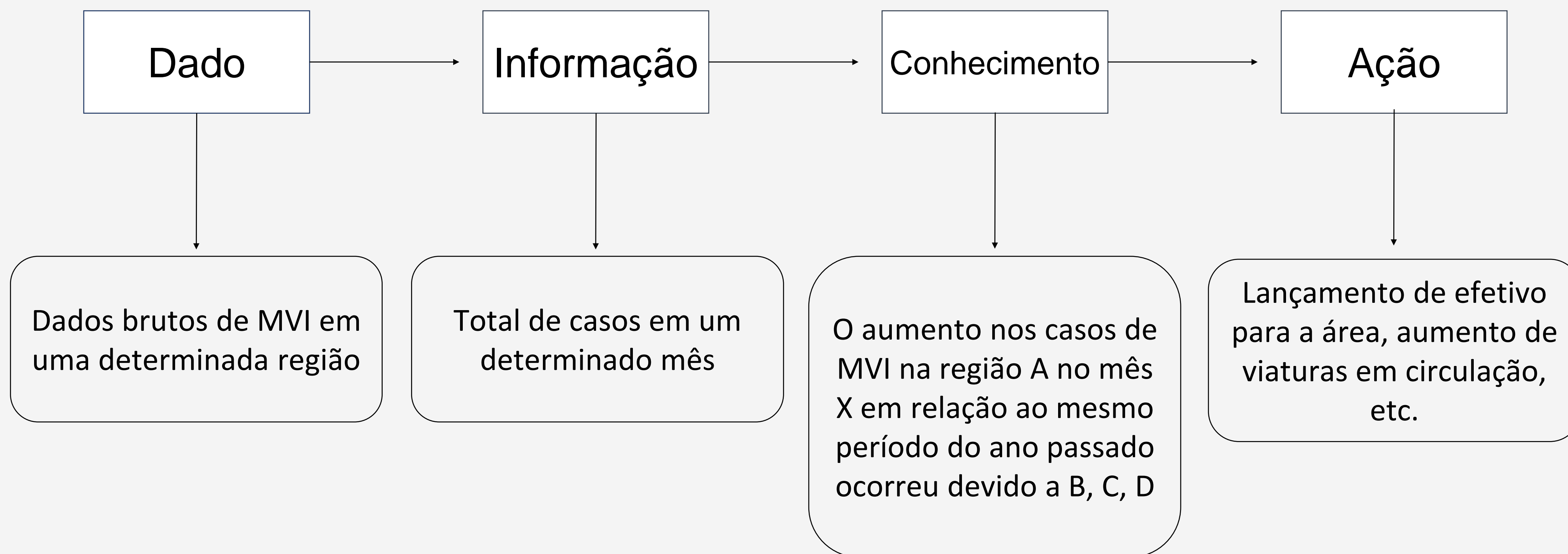
Wisdom (“sabedoria”)

Adição de ação ao conhecimento; subjetivo; voltado ao futuro



Dado, informação e conhecimento

- O valor cresce na medida em que essas transformações ocorrem





No setor privado

- Uso de dados no processo de tomada de decisão
- Ciência de Dados, Big Data e ferramentas de Business Intelligence (BI)
- Big Data: grandes volumes de dados complexos e diversificados, que incluem dados estruturados, não estruturados e semiestruturados
- Inteligência Artificial: simular o pensamento humano. Fazer máquinas inteligentes, com objetivo de fazer com que elas realizem tarefas que, se feitas por pessoas, exigiriam inteligência (John McCarthy).
- Obtenção de insights estratégicos, aumento da eficiência nas operações, melhora da experiência do cliente, identificação de padrões de comportamento do consumidor, análise de padrão de compras, etc.



No setor público

- Expansão para o setor público -> Eficiência, *accountability* e transparência
- Promover um alinhamento mais fácil com a estratégia, através do monitoramento e da avaliação de indicadores de resultado (Costa, 2012)
- Aplicações interativas, análises preditivas para o acompanhamento de indicadores
- Identificação de gargalos e atender demandas dos atores envolvidos





O que é Ciência de Dados?

- Envolve princípios, processos e técnicas para compreender fenômenos por meio de análise de dados automatizada (Provost & Fawcett).
- Conceito conexo à camada dos métodos, na qual os softwares são empregados para transformar dados em informação, resultando no apoio à tomada de decisão.
- Extração de informação útil a partir de imensas, complexas e dinâmicas bases de dados (Bugnion, Manivannan e Nicolas, 2017).
- Profissional deve entender o funcionamento de algoritmos de ML (subárea de AI) bem como saber interpretar os resultados
- Formulação de hipóteses e aquisição de informação para subsidiar o processo de decisão.



Conceito de Ciência de Dados

Ciência de Dados (Data Science) é uma área interdisciplinar que combina **métodos científicos, processos, algoritmos e sistemas** para extrair **conhecimento e insights** a partir de dados estruturados e não estruturados. Ela integra **estatística, ciência da computação e conhecimento do negócio**. O objetivo principal é apoiar a tomada de decisão baseada em evidências, criar produtos baseados em dados e gerar valor por meio de análises quantitativas.



Ciência de Dados

- Ciência utiliza métodos que garantam a replicação
- Dados estruturados: organizados em formatos predefinidos e fáceis de analisar (tabela, planilha, banco de dados)
- Dados não estruturados: não seguem um padrão fixo: texto, imagem, áudio, vídeo.
- Uso de métodos de aprendizado de máquina (supervisionado e não supervisionado) e estatística
- Algoritmos: conjunto de instruções passo a passo usado para resolver um problema ou executar uma determinada tarefa. É uma sequência lógica de operações que um computador pode seguir para transformar dados de entrada em resultados úteis.

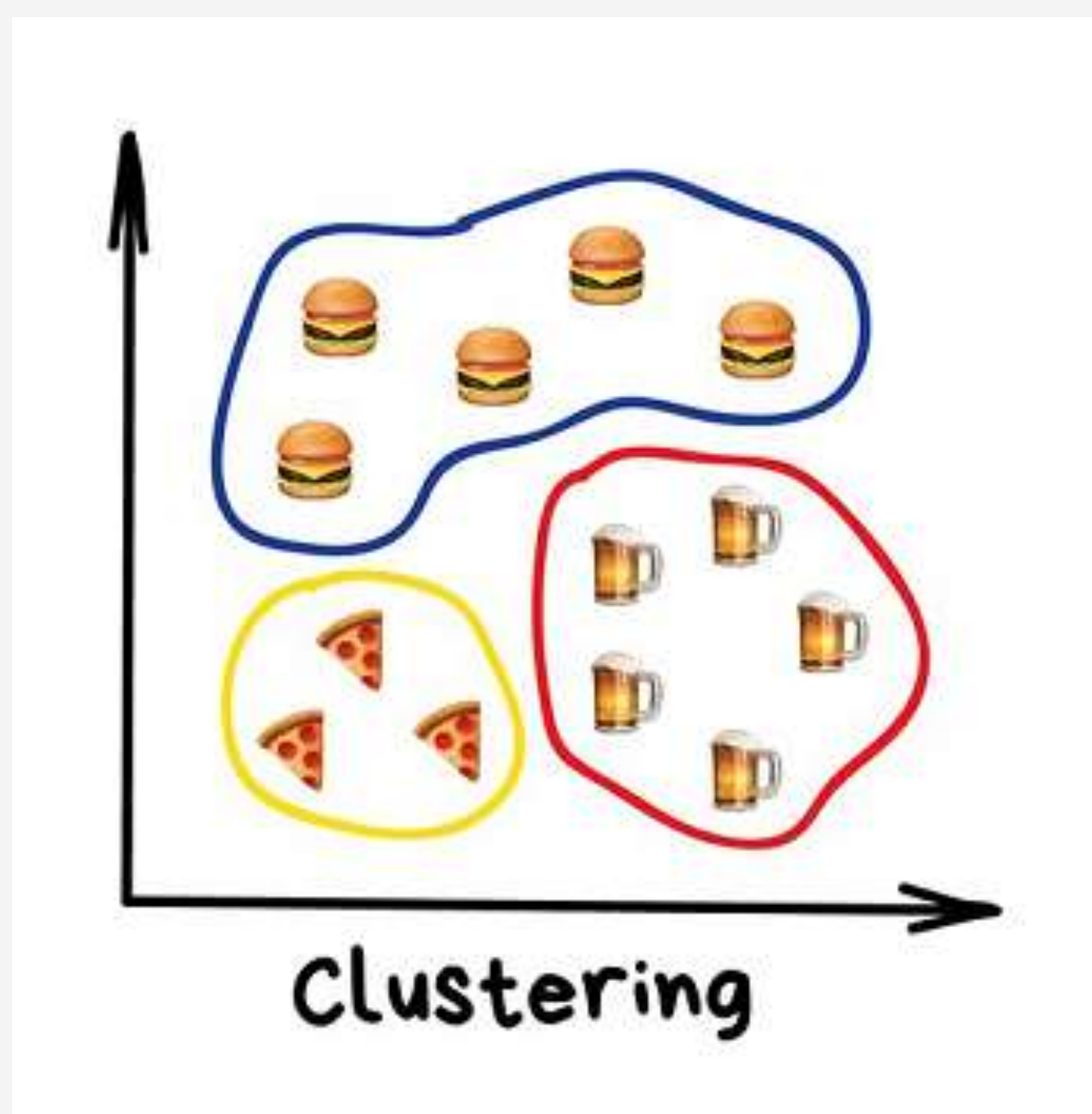


Aprendizado de máquina

- Campo de estudo localizado entre algoritmos, estatística, econometria e computação numérica. Em geral, os métodos de aprendizado de máquina estão preocupados com soluções algorítmicas para problemas numéricos baseados em dados e sua implementação eficiente na computação moderna.
- Aprendizagem Supervisionada: Métodos onde os rótulos dentro da amostra são conhecidos.
- Aprendizagem não supervisionada: Métodos onde os rótulos dentro da amostra não são conhecidos.

AM – não supervisionado

- Começam com apenas X e descobrem algum padrão latente dentro de X
- Ex: Clustering (agrupamento) -> definir essa métrica de similaridade.

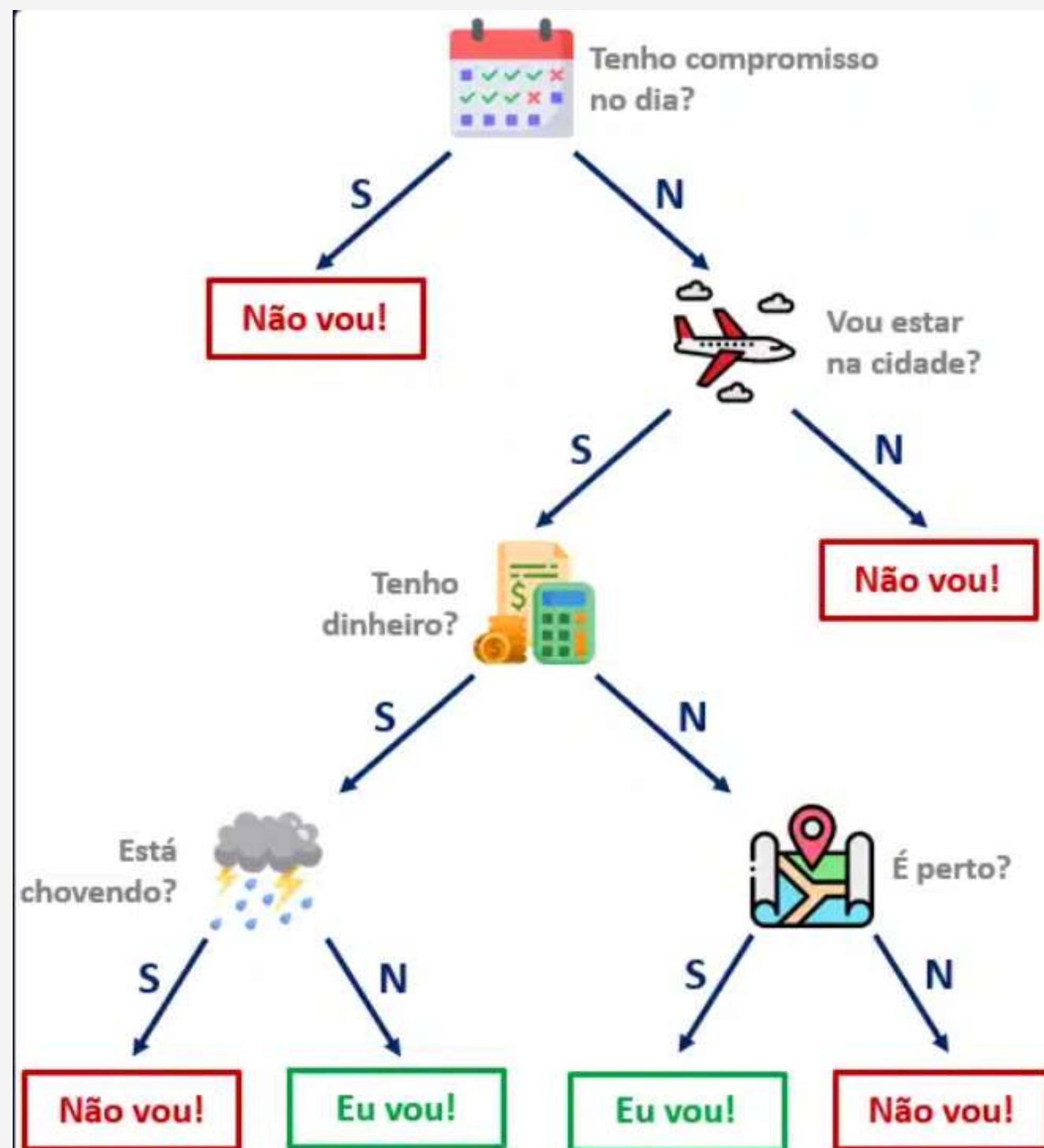


Fonte: Reprodução da Internet
(<https://www.datageeks.com.br/machine-learning/>)



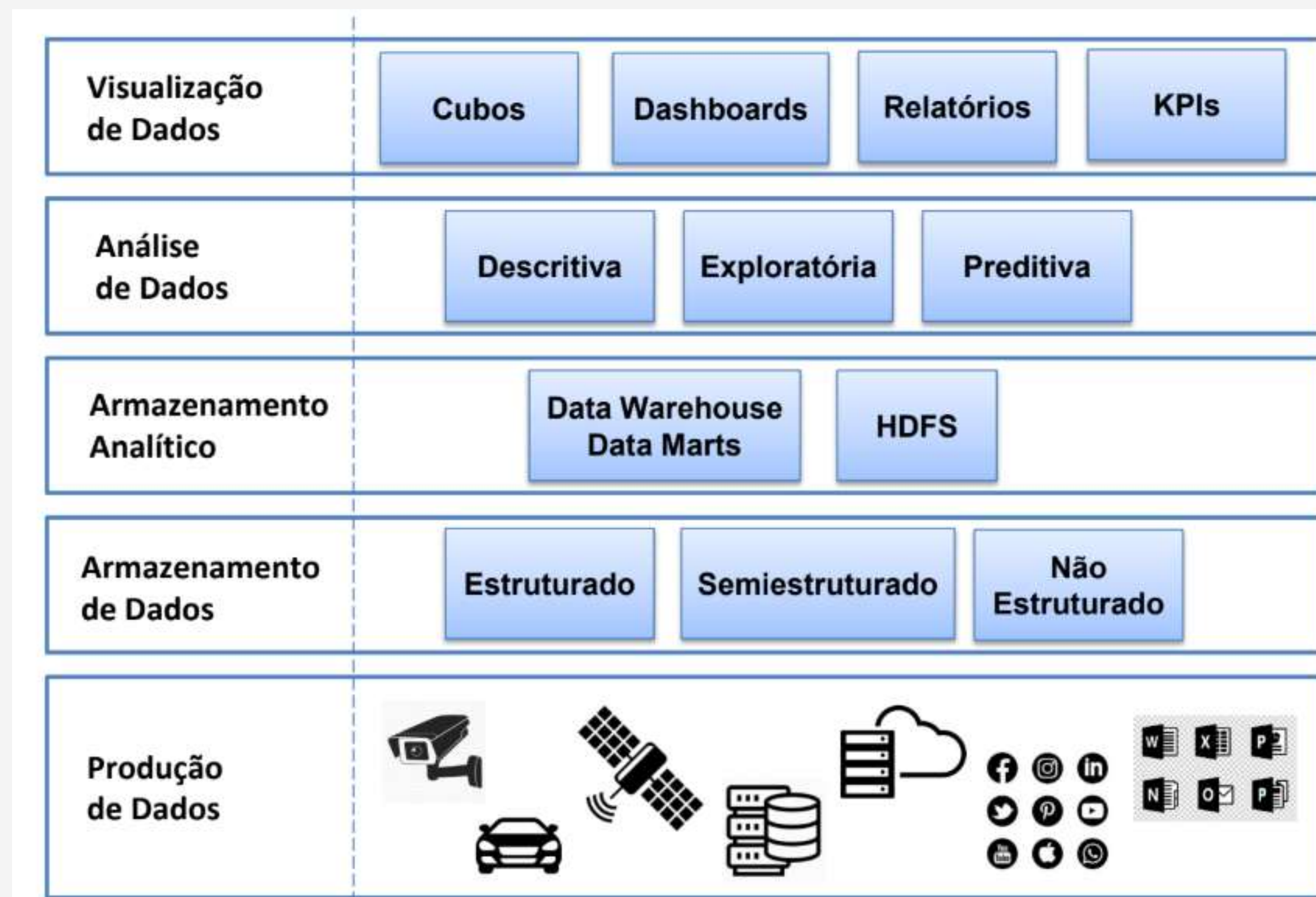
AM – supervisionado

- Começa com um conjunto de dados contendo features (X) e labels (y).
- Constrói-se uma “regra” relacionando X a y, de modo que, dada alguma combinação de valores para X, seja possível “prever” um valor de y.
- Ex: Árvores de decisão



Fonte: Reprodução da Internet
(<https://www.hashtagtreinamentos.com/arvore-decisao-ciencia-dados>)

Panorama da CD



Fonte: Maciel, 2025

Ciclo da CD



Figura 5. Ciclo de Vida da Ciência de Dados (Bugnion, Manivannan e Nicolas, 2017) [tradução dos autores]



Fonte: Mason and Wiggins, 2010



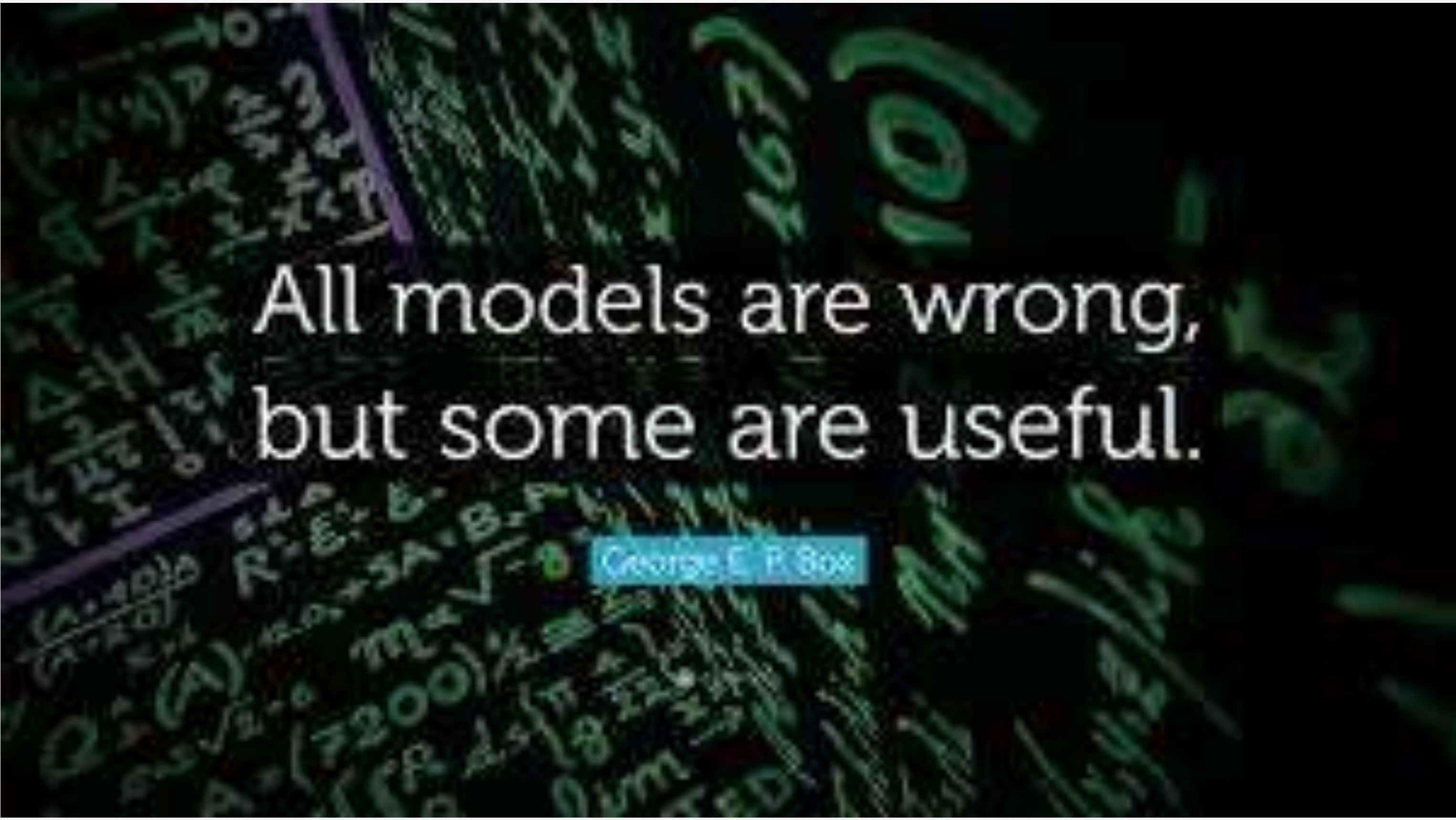

Ciclo da CD

- Obtenção dos Dados: seleção dos dados e metadados (dado sobre o dado) a partir do processamento de determinados arquivos, de dados da web, etc.
 - Em um arquivo de texto (como um Word ou PDF): Título do documento; Autor; Data de criação; Número de páginas; Palavras-chave associadas
 - Em uma foto digital: Tamanho do arquivo: 3 MB; Resolução: 1920x1080 pixels; Data e hora em que a foto foi tirada; Localização GPS (se o recurso estiver ativado); Modelo da câmera usada
- Ingestão dos Dados: transformação de dados de diferentes fontes em uma base centralizada. Utilização de softwares como R e Python
- Exploração dos Dados: Estudos preliminares para definir como os dados podem se transformar em informações relevantes



Ciclo da CD

- Definição dos Parâmetros: Escolhas necessárias para o emprego dos algoritmos de ML (construção do algoritmo)
- Implementação do Modelo: Estratégias de treinamento e testes dos algoritmos. Escolhe-se o melhor modelo
- Utilização do Modelo
- Tomada de decisão
- Necessidade de equilibrar as expectativas
- Modelos são representações imperfeitas da realidade e não devem ser levados como verdade absoluta



All models are wrong,
but some are useful.


George E. P. Box



Uso de Algoritmos - Exemplos

- Gestão e Otimização de Recursos: Alocar melhor policiais, ambulâncias ou professores com base em dados históricos e geolocalização.
- Previsões e Análises: Prever a demanda por serviços públicos como creches ou hospitais
- Óbitos em Acidentes de Transporte Terrestre (ATT)
 - https://segpr.seplag.pe.gov.br/docs/att_relatorio/
- Detecção de Irregularidades: Monitorar gastos públicos de forma automática
- Transparência e Acesso à Informação: Organizar e disponibilizar dados públicos com algoritmos que classificam e limpam as informações.
- Atendimento ao Cidadão: uso de ChatBots

Uso de Algoritmos

 Relatório ATT

Home 

Nota Técnica 4/2025

Informações Gerais

Objetivo

Metodologia

Ferramenta

Pernambuco

Nesta página

Informações Gerais

Objetivo


Metodologia

Ferramenta

Pernambuco

Checando o que acontece nas GERES

Secretaria de Planejamento, Gestão e Desenvolvimento Regional

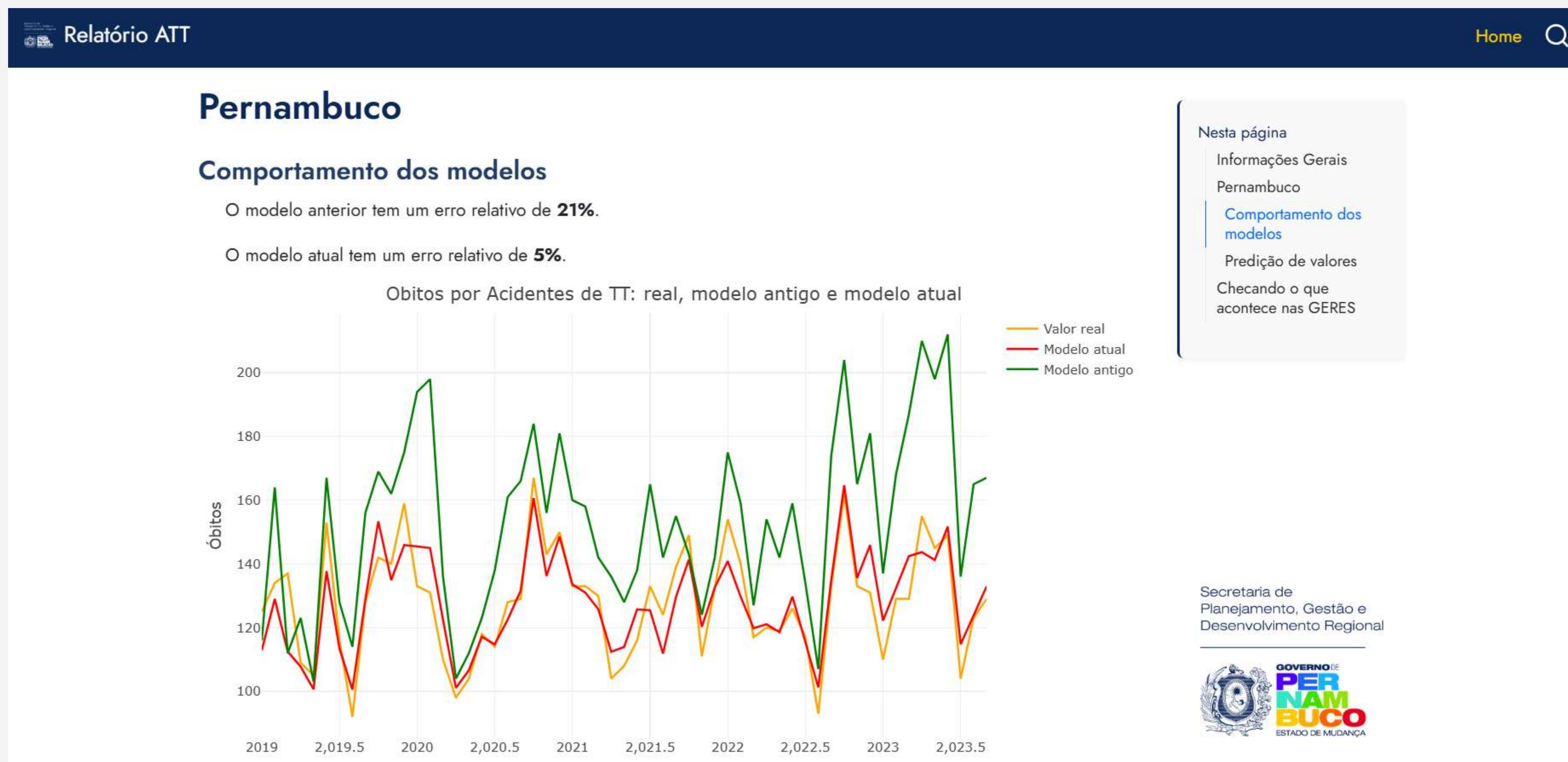
 GOVERNO DE PERNAMBUCO
ESTADO DE MUDANÇA

Analisar a série histórica de óbitos por acidente de transporte terrestre ocorridos em Pernambuco e nas Gerências Regionais de Saúde (GERES), a fim de criar novo fator de correção que estime de forma mais precisa os óbitos dos 12 (doze) últimos meses.

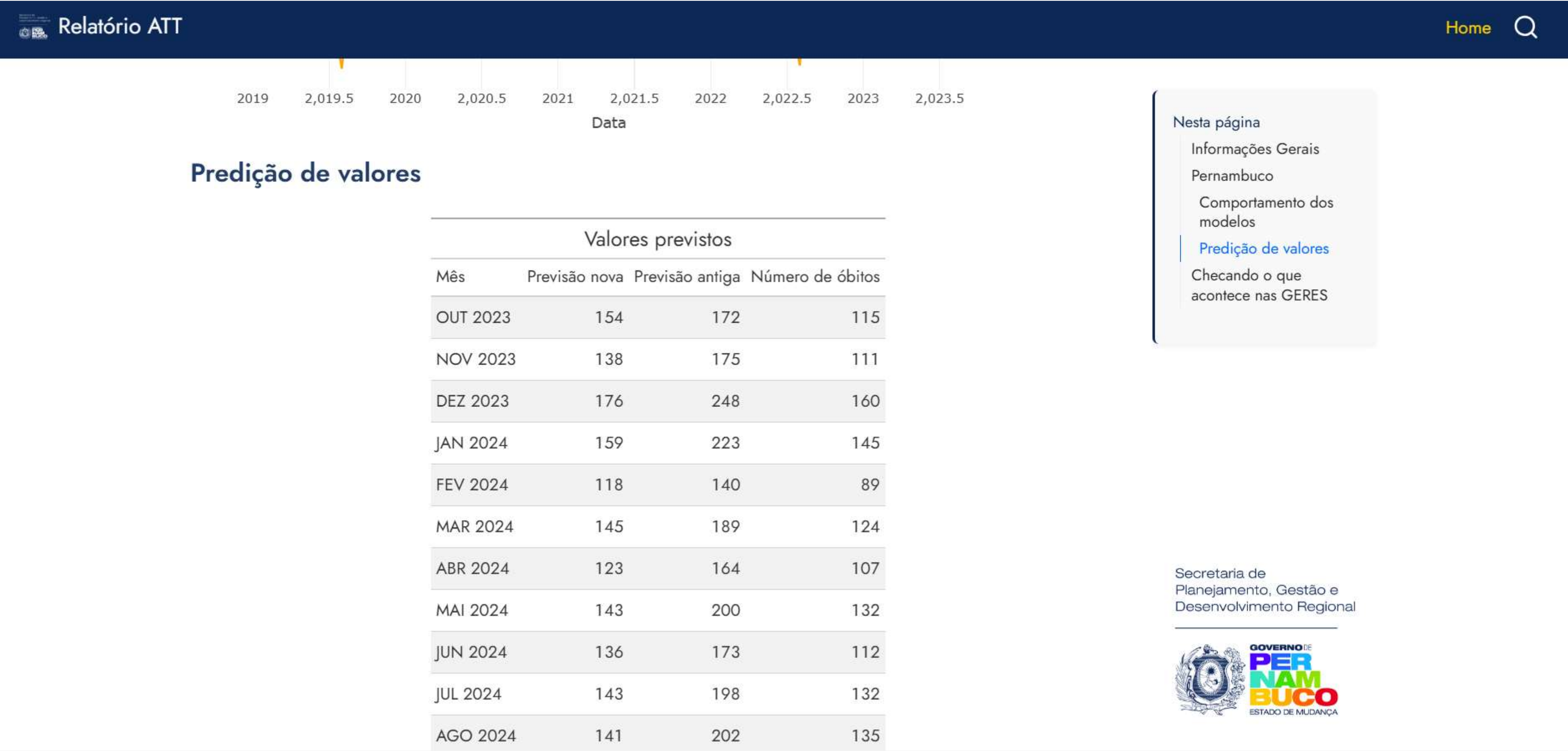
A análise de séries temporais observa o comportamento dos dados em relação ao passado, à tendência, à sazonalidade e ao ruído. Assim, é possível escolher a melhor estratégia para prever valores futuros.

Foi utilizada a linguagem R, juntamente com o pacote tslm.

Uso de Algoritmos



Uso de Algoritmos



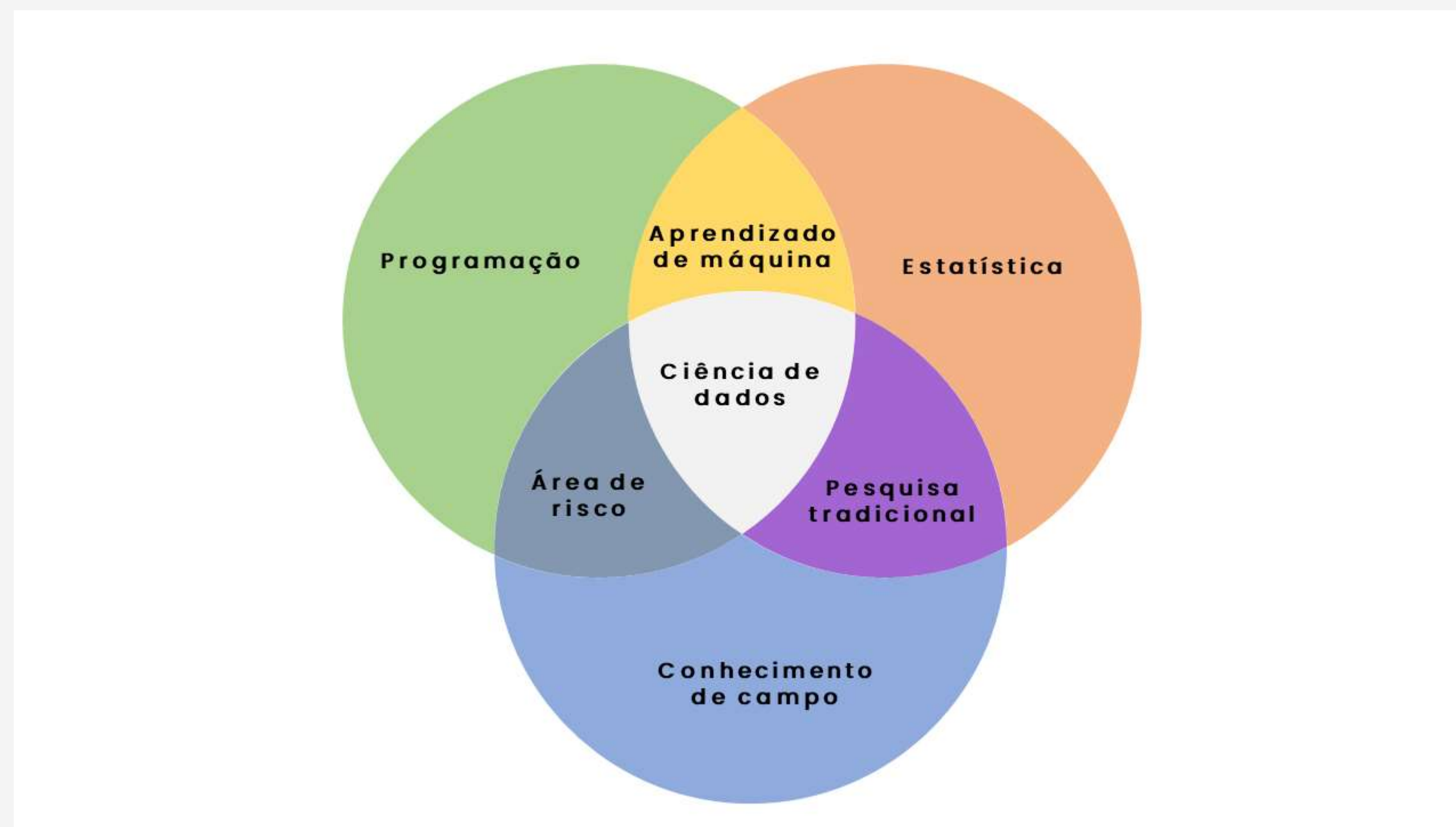


Construção do modelo

- Uso do R e Quarto
- Pacote tslm
- Usado para ajustar modelos lineares a séries temporais, incluindo componentes de tendência e sazonalidade

Elementos da Ciência de Dados

- Necessidade de integração dos 3 pilares (computação, matemática e negócio)



Fonte: Reprodução da internet (https://medium.com/@elieser_ribeiro/o-que-%C3%A9-ci%C3%Aancia-de-dados-5b2654b9fa08)



Elementos da Ciência de Dados

- Tríade da Ciência de Dados
- Cientista de Dados trafega pelos 3
- Estatística para inferir padrões
- Computação para processamentos dos dados em grande escala
- Conhecimento do negócio para a interpretação e aplicação dos resultados em determinado contexto
- No setor público auxilia no ciclo de políticas públicas, principalmente no monitoramento e avaliação.
- Facilita a gestão de recursos, aumento na produtividade e eficiência nos processos



Computação, Estatística e Negócio

- Extrair valor dos dados
- Matemática e Estatística: oferece as bases teóricas para entender padrões, modelar incertezas, otimizar processos e validar resultados.
 - Modelagem, inferência, regressão, testes de hipóteses, álgebra linear, cálculo.
- Computação: infraestrutura que torna viável coletar, armazenar, processar e modelar grandes volumes de dados.
 - Programação (Python, R, SQL), bancos de dados, big data, cloud, APIs, automação.
- Negócio/Área de Domínio: direciona o porquê de qualquer análise de dados.
 - Entendimento do problema, regras de negócio, impacto das decisões.

Computação, Estatística e Negócio

- Computação

Aspecto	Descrição	Exemplos
Programação	Transformar algoritmos em código eficiente.	Python, R, SQL, Scala
Banco de Dados	Gerenciar dados estruturados e não estruturados.	PostgreSQL, MongoDB, BigQuery
Engenharia de Dados	Construir pipelines para ETL (Extract, Transform, Load).	Airflow, Spark
APIs e Integrações	Consumir e integrar sistemas.	REST APIs, GraphQL
Computação em Nuvem	Escalar processamento.	AWS, Azure, GCP
MLOps	Deploy de modelos no ambiente de produção.	Docker, Kubernetes, MLflow

- Pipelines: sequência de processos que transferem dados de uma ou mais fontes para um destino



Computação, Estatística e Negócio

- Matemática

Aspecto	Descrição	Exemplos
Estatística Descritiva	Resumir dados (média, mediana, variância).	Descrição de comportamento de clientes.
Probabilidade	Modelar incertezas.	Modelos de previsão de risco.
Inferência Estatística	Tirar conclusões de amostras.	Testes de hipóteses A/B.
Álgebra Linear	Manipular matrizes, fundamental em ML.	Recomendação de produtos (SVD).
Cálculo	Otimizar funções (gradientes, loss functions).	Treinamento de redes neurais.
Otimização	Encontrar melhores soluções para problemas.	Ajustar hiperparâmetros de modelo.

Computação, Estatística e Negócio



- Negócio

Aspecto	Descrição	Exemplos
Entendimento de Problema	Traduzir problema de negócio para problema de dados.	"Como reduzir o churn em 10%?"
Métricas de Sucesso	Definir KPIs corretos.	Aumento de receita, redução de custo.
Comunicação com Stakeholders	Explicar insights técnicos em linguagem de negócios.	Apresentar modelo de churn para diretoria.
Tomada de Decisão Baseada em Dados	Influenciar decisões operacionais e estratégicas.	Decidir investimento em campanhas de marketing.
Ética e Governança de Dados	Garantir conformidade legal e social.	LGPD, GDPR.

- Key Performance Indicator* -> métrica que mede o progresso de um negócio ou projeto em direção a um objetivo estratégico



Exemplo aplicado

- Problema de Negócio: Identificar possíveis fraudes em pagamentos de auxílios (ex: auxílio emergencial, Bolsa Família).
 - Computação:
 1. Integrar bases públicas (Cadastro Único, Receita Federal, INSS).
 2. Usar bancos de dados grandes (PostgreSQL, BigQuery) para armazenamento e consulta.
 3. Automatizar verificações de inconsistências via scripts em R ou Python.
- 
- 





Exemplo aplicado

- Matemática:

1. Análise estatística para definir perfis de risco (clusters de beneficiários suspeitos).
2. Modelos supervisionados (ex: regressão logística, árvores de decisão) para prever chance de fraude.
3. Análise de anomalias usando técnicas de detecção de outliers.

- Negócio:

1. Reduzir perdas financeiras.
 2. Focar auditorias em casos de alta probabilidade de fraude (priorização inteligente).
 3. Economizar milhões de reais em fiscalizações.
- 
- 



Atividades, papéis e produtos da CD

- As soluções da CD auxiliam os gestores em suas Tarefas Intensivas em Conhecimento (Schreiber etl al., 2000)
 - Associação: estabelecimento de relações entre dois ou mais conjuntos de dados; identificação de padrões ou vínculos (não necessariamente causais).

"clientes que compram vinho tinto frequentemente também compram queijo brie".

Identificar que *estudantes de escolas com alto índice de evasão escolar também apresentam alto índice de vulnerabilidade social*. Essa associação pode servir para criar políticas públicas mais direcionadas.



Atividades, papéis e produtos da CD

- Avaliação: Analisar um objeto, situação ou processo com base em critérios definidos. Pode envolver julgamento de valor, qualidade ou conformidade.

Avaliar a qualidade de um produto com base em critérios como durabilidade, funcionalidade e design.

Avaliar o desempenho de um programa de saúde pública em relação a metas de cobertura e redução de internações hospitalares.



Atividades, papéis e produtos da CD

- Diagnóstico: Identificação de causas ou problemas com base em sintomas ou dados observáveis.

Diagnosticar uma falha em um equipamento com base em barulhos anormais e queda de performance.

Diagnosticar a causa do aumento da criminalidade em determinada região com base em dados socioeconômicos, patrulhamento e desemprego.



Atividades, papéis e produtos da CD

- Monitoramento: Observação contínua ou periódica de indicadores, eventos ou processos para acompanhar o estado ou desempenho.

Monitorar o tráfego em tempo real para gerenciar rotas alternativas em aplicativos de navegação.

Monitoramento de casos de dengue em diferentes bairros para ativar campanhas de prevenção em áreas críticas.



Atividades, papéis e produtos da CD

- Predição: Estimar eventos futuros com base em dados históricos e modelos.

Prever a demanda por energia elétrica nos próximos meses com base no histórico de consumo e clima.

Prever o número de matrículas escolares para o ano seguinte, ajudando no planejamento de infraestrutura e professores.



Atividades, papéis e produtos da CD

- Predição
- Coleta e Integração de Dados
- Limpeza e Transformação (Data Wrangling)
- Exploração e Análise Estatística (EDA)
- Modelagem Preditiva e Classificatória
- Validação e Interpretação de Modelos
- Visualização de Dados (DataViz)
- Geração de Insights e Tomada de Decisão
- Deploy e Monitoramento de Modelos (MLOps)



Produtos

- Modelos Preditivos
- Dashboards e Relatórios
- Sistemas de Segmentações e Perfis de
- Automação de Processos
- Assistentes Inteligentes e Chatbots
- Simulações e Otimizações
- Ferramentas de Detecção de Anomalias
- APIs de Dados



Papeis

- Cientista de Dados: Modelagem, análise estatística, predição, storytelling com dados
- Engenheiro de Dados: Coleta, tratamento, pipelines e armazenamento de grandes volumes de dados
- Analista de Dados: Relatórios, dashboards, visualizações, KPIs
- Engenheiro de Machine Learning: Deploy de modelos, automação, monitoramento
- Product Owner/Especialista de Domínio: Alinhamento dos modelos com as necessidades reais da organização



Cultura analítica, capacidade analítica e maturidade analítica.

- Analytics: uso amplo de dados, de análise estatística e quantitativa, de modelos descritivos e preditivos para orientar decisões e agregar valor (Davenport & Kim).
- Cultura analítica é um ambiente organizacional onde a análise de dados é integrada à tomada de decisões em todos os níveis, substituindo ou complementando a intuição e a experiência.
- Incentiva a análise dos dados de forma que eles se tornem informações de fácil leitura e compreensão
- Exige novos comportamentos por parte da equipe.
- As práticas devem ser incentivadas pela liderança



Como a cultura analítica ajuda?

- Cultura envolve valores compartilhados por uma organização
- O pensamento analítico tem a ver com as pessoas tomando decisões olhando para dados, não para tecnologia (Cappra)
- 4 pilares fundamentais: Pessoas, Processos, Regras e Tecnologia
- Mudança de padrão de pensamento
- Tomada de decisões mais assertivas
- Identificação do que está ou não funcionando
- Papel do gestor de dados é fundamental
- Necessidade de domínio de metodologias ágeis (Ex: Scrum)



Capacidade analítica e maturidade analítica.

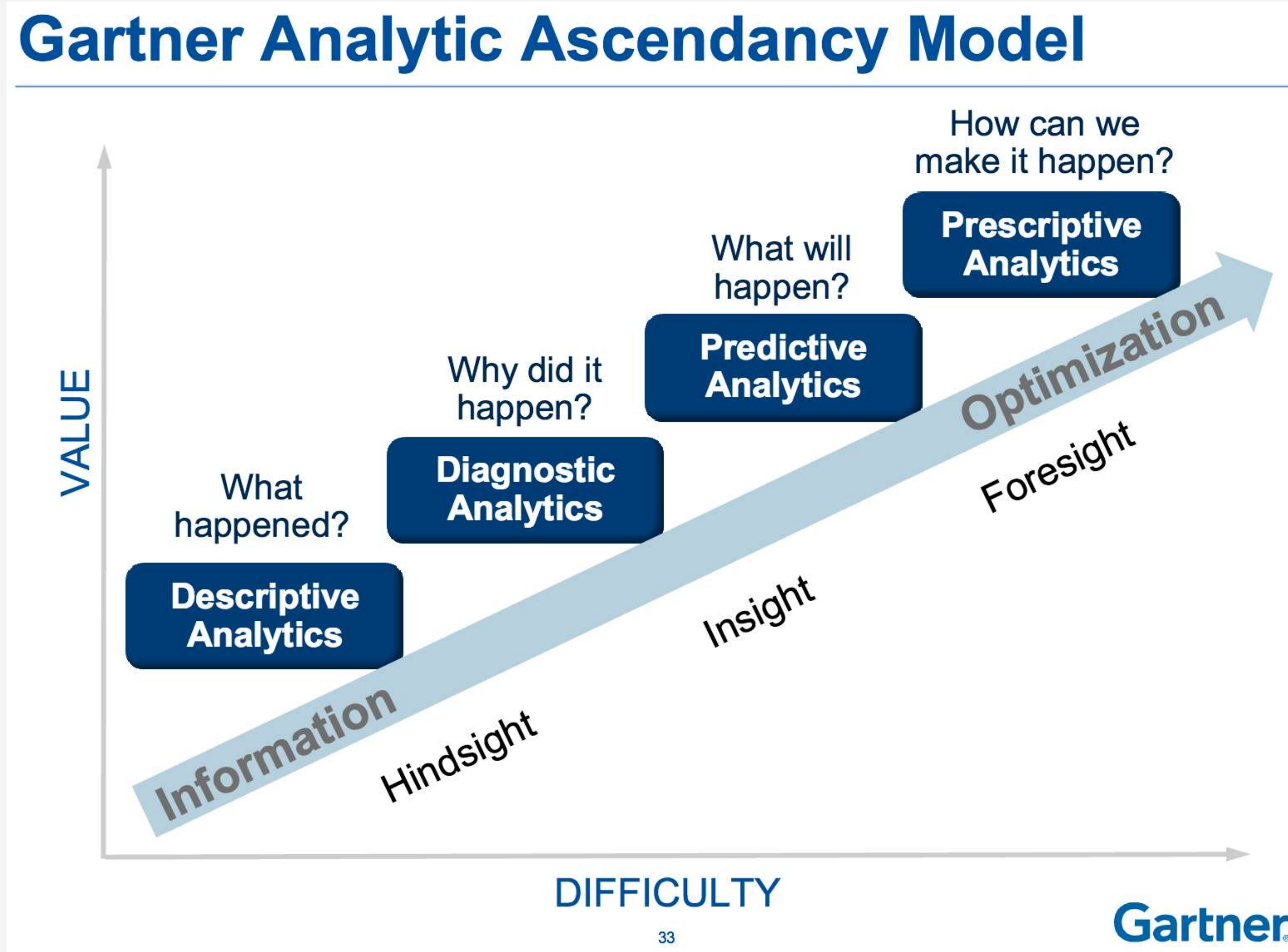
- Capacidade analítica: capacidade de analisar dados e informações, identificar padrões, resolver problemas e tomar decisões informadas.
- Pensar criticamente, interpretar informações, e utilizar essa compreensão para tomar decisões e resolver problemas complexos
- Requisito considerado básico nas organizações
- Maturidade analítica só se alcança com profissionais com alta capacidade analítica
- Essencial o entendimento do contexto externo, usuário final, processos internos, etc.
- Pessoas que têm capacidade analítica embasam suas decisões sempre que precisam dar passos importantes -> projetam cenários e interpretam dados



Maturidade analítica.

- Maturidade analítica: capacidade de uma organização ou indivíduo utilizar a análise de dados de forma eficaz para tomar decisões informadas e melhorar o desempenho.
- Envolve a coleta, análise e aplicação de dados para identificar oportunidades, reduzir riscos, otimizar processos e tomar decisões estratégicas.
- Só é alcançada quando há pessoas capazes de tomar decisões bem embasadas.
- Quando as organizações conseguem avançar em sua maturidade analítica, elas ganham vantagem competitiva.
- No setor público, promove uma maior eficiência nos processos

Gartner e o modelo de maturidade analítica



Fonte: Reprodução da Internet ([Link](#))



Modelo de Maturidade analítica.

- Análise descritiva: o que já aconteceu. (Ex: relatórios mensais e dashboards de KPIs (indicadores-chave de performance))
- Análise diagnóstica: o porquê de algo ter acontecido. Olhar com mais detalhes as informações disponíveis, a fim de responder perguntas e encontrar padrões.
- Análise preditiva: o que vai acontecer. Começa a ser mais proativa no uso de dados para a tomada de decisões (uso de Ciência de Dados para a criação de modelos estatísticos).
- Análise prescritiva: usar dados para saber com clareza o melhor curso de ação. (Ex: Netflix, Amazon. Uso de IA e ML para é aprender com os dados. Eles apontam o caminho a ser seguido. Unindo com a visão de negócio e experiência, melhores decisões são tomadas).



Estágios de maturidade analítica - Cappra

- Cappra Institute for Data Science
- Data-negation: organizações que não valorizam os dados em seus processos e que não acreditam no potencial analítico aplicado ao negócio
- Data-curious: organizações que fazem aplicações pontuais para uso de dados, sem constância.
- Data-try: organizações em busca de estabilização de uma operação orientada por dados, testando alternativas para reduzir o feeling nas decisões.



Estágios de maturidade analítica - Cappra

- Data-safety: organizações que utilizam os dados de forma estável, principalmente para justificar suas ações, e que se sentem seguras usando dados.
- Data-driven: organizações analíticas, com estratégia, processos, pessoas com capacidade analítica e espaços plenamente adequados para usar dados nos negócios.
- AI-driven



Inventário dos dados

- Necessário para a gestão de conhecimento e melhor entendimento do que se tem e o que se precisa
- A busca por soluções anteriores pode auxiliar na resolução de problemas atuais



GOVERNO DE
PERNAMBUCO
ESTADO DE MUDANÇA

