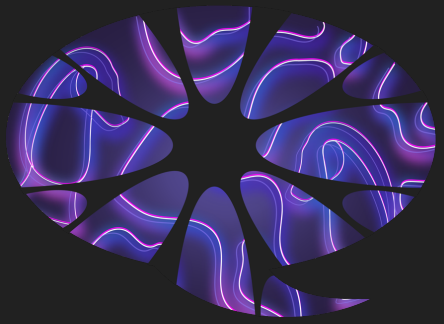


IBM Research Brasil

Colloquium 2019

Inteligência Artificial no Brasil

ProvLake on ML

Gerenciamento da Linhagem de Dados para o Ciclo de Vida de Machine Learning (ML) em Recursos Naturais

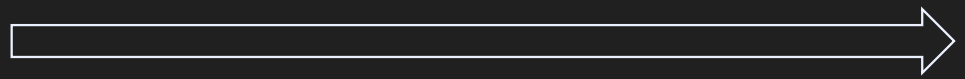
Renan Souza
Leonardo Guerreiro Azevedo
Raphael Thiago

IBM Research Brasil

Machine Learning (ou Aprendizado de Máquina)

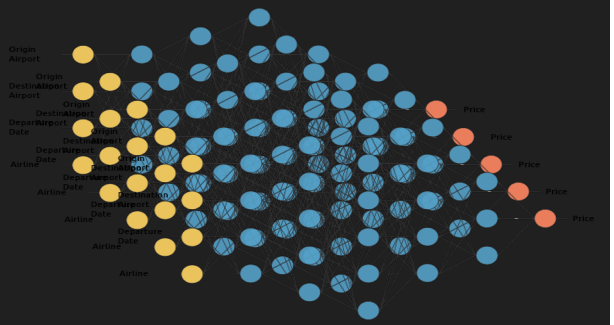
Constrói modelos a partir de dados de entrada para fazer previsões ou decisões guiadas

INPUT



OUTPUT

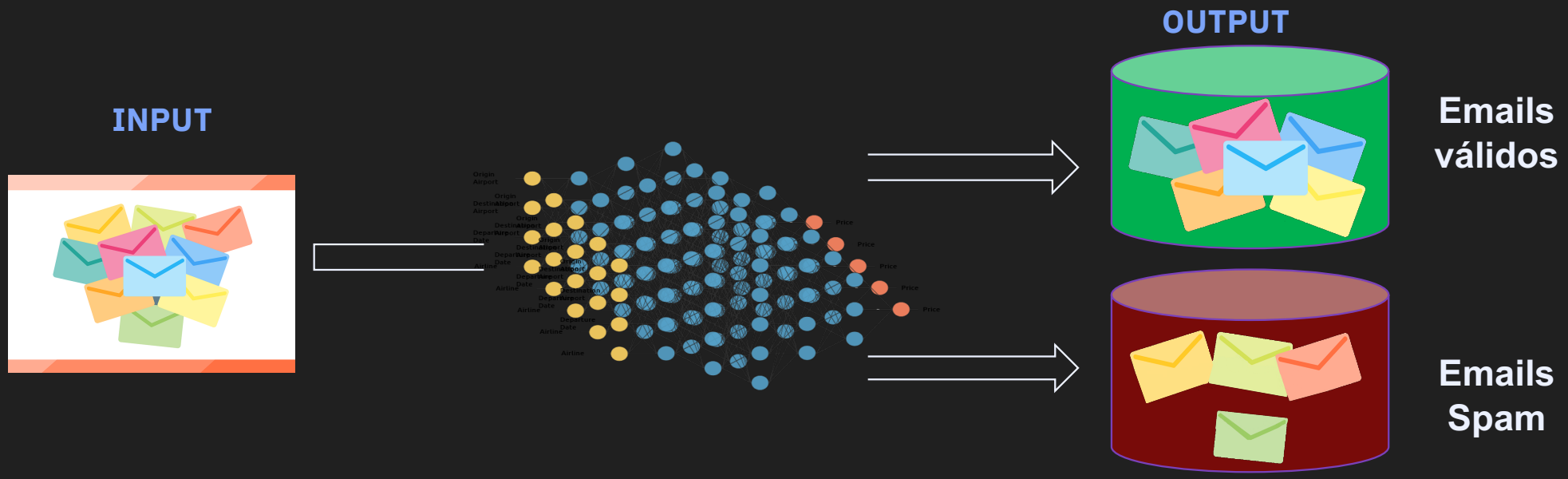
Modelos de ML
Apoiar a tomada de decisão no domínio



Exemplo: Filtrar emails

Machine Learning (ou Aprendizado de Máquina)

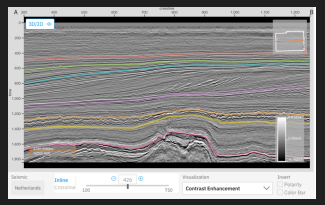
Modelos construídos são aplicados, fazendo as previsões ou decisões



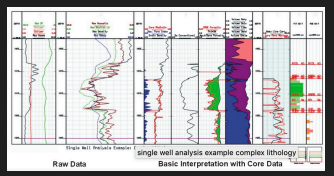
Contexto: Ciclo de Vida de Machine Learning para Óleo & Gás

INPUT

Dados Geológicos Brutos



Sísmicas

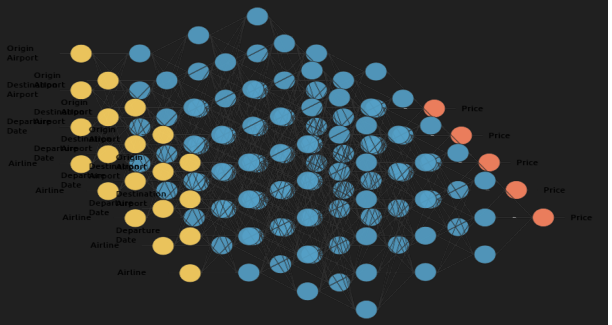


Well Logs

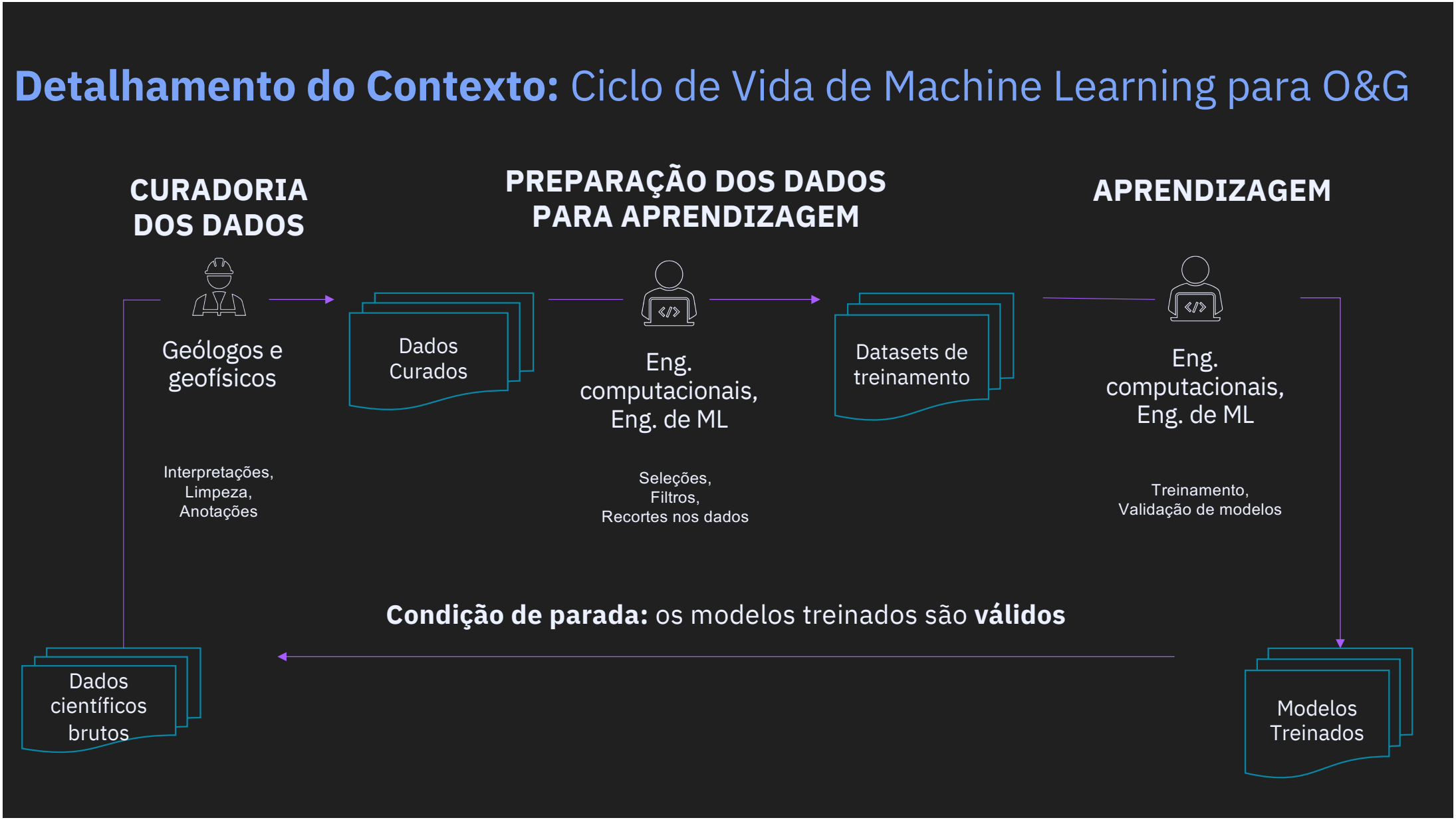
OUTPUT

Modelos de ML

Apoiar a tomada de decisão no domínio



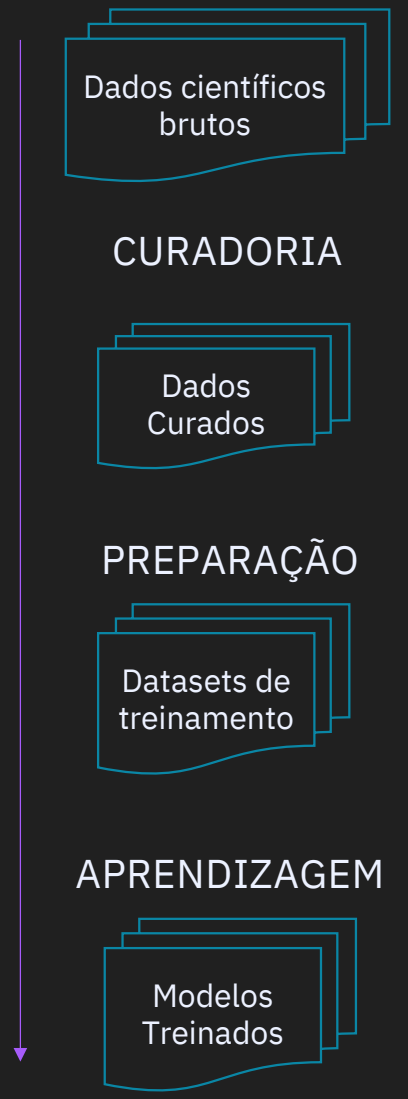
Detalhamento do Contexto: Ciclo de Vida de Machine Learning para O&G



Problema e Motivação

Problema: Como permitir entender as transformações dos dados que ocorrem no ciclo de vida de ML, de ponta-a-ponta, **desde os dados brutos até os modelos treinados?**

Motivação: Resolver esse problema é essencial para validar os modelos treinados e entender o quanto os modelos generalizam **respeitando as características do domínio**.

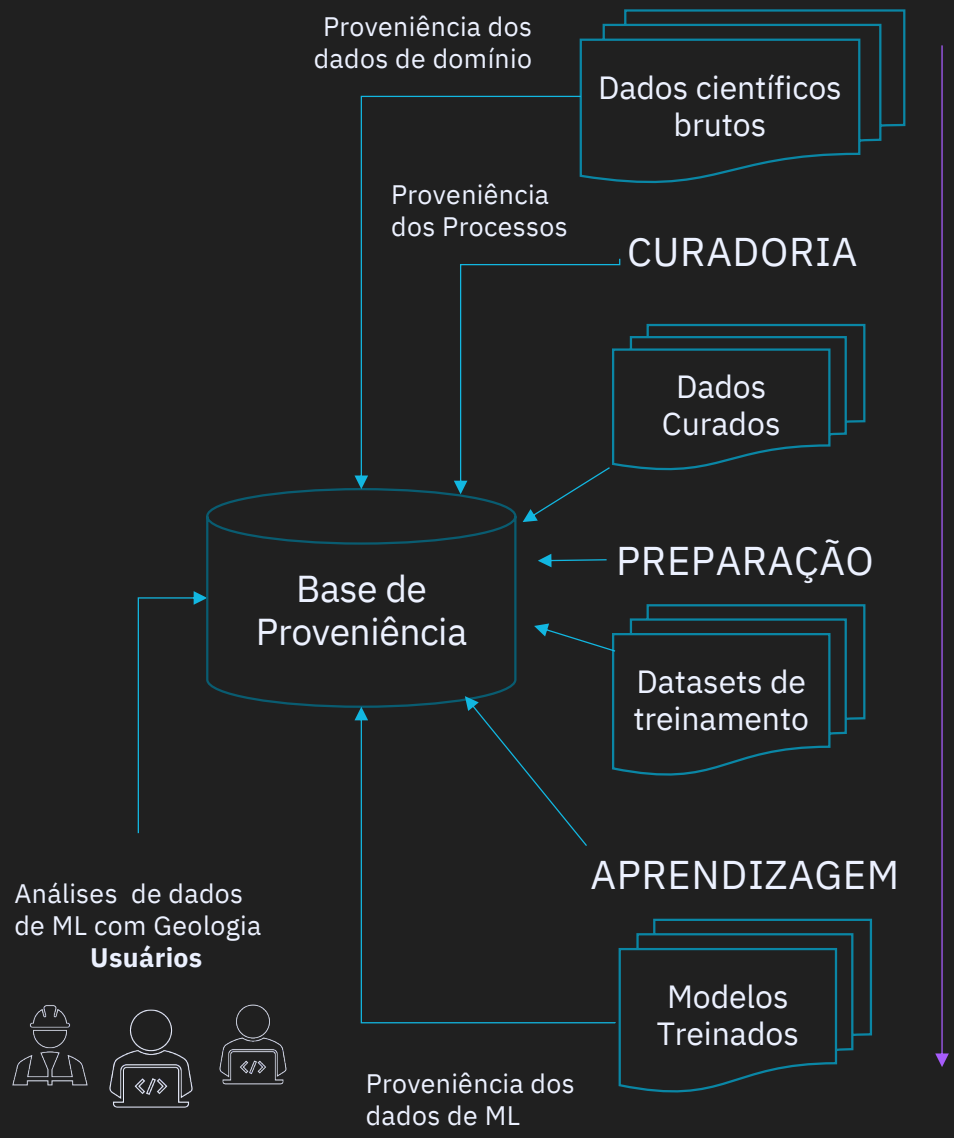


Solução

ProvLake on ML

Permite o rastreamento da linhagem dos dados ao longo do ciclo de vida de ML, armazenada numa única **base de proveniência** e, como consequência:

- Permite uma **visão integrada** dos dados geológicos até os dados de ML
- Facilita tomadas de **decisão** que analisam os processos computacionais, os dados e os modelos de ML gerados



Diferenciadores da solução

- Baixa sobrecarga de captura de proveniência em execuções de Computação de Alto Desempenho
- Captura de proveniência das três fases do ciclo de vida, especialmente a de curadoria de dados de domínio
- Visão integrada de dados de domínio e de ML processados em múltiplos workflows.
- Permite análises que incluem dados de domínio, execução e ML.

Data Lineage Management

ProvLake on ML

Gerenciamento da Linhagem de Dados para o Ciclo de Vida de Machine Learning (ML) em Recursos Naturais

Renan Souza
Leonardo Guerreiro Azevedo
Raphael Thiago