Machine learning para predições em saúde

Prof. Dr. Alexandre Chiavegatto Filho

FEDERATED LEARNING: CHALLENGES, METHODS, AND FUTURE DIRECTIONS

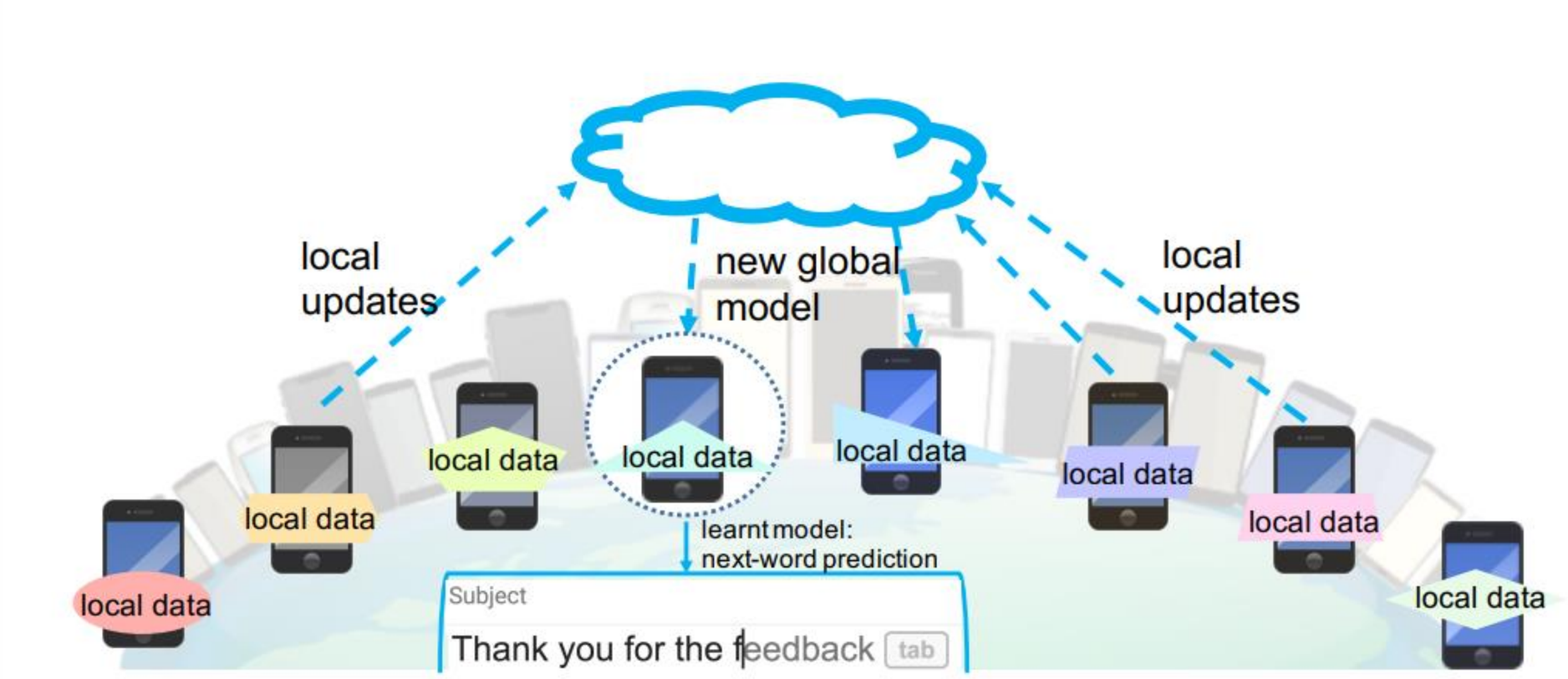
Authors: Tian Li, Anit Kumar Sahu, Ameet Talwalkar, Virginia Smith.

Resumo

- Treinamento em dispositivos remotos
- Desafios: heterogeneidade e grande quantidade de dados
- Dados locais
- Modelo global
- Principal objetivo: preservação da privacidade + aceitação de novas colaborações (sem perda de controle dos dados)

Contextualização

Exemplo: Aprendizado Federado para predição da palavra seguinte em dispositivos móveis.



Contextualização

- Outro exemplo importante são os Hospitais (dados dos pacientes)
- Hospitais operam com rigorosas práticas de privacidade (LGPD)
- Restrições legais e éticas
- Aprendizado federado aparece como uma solução promissora

Formulação do Problema

- Objetivo é aprender o modelo sob restrição de que o armazenamento e o processamento são em geral locais, com algumas atualizações intermediárias periódicas

Quatro desafios centrais

1.Comunicação em rede é cara

Precisa reduzir nº de transferências de informação e o tamanhos delas

2.Heterogeneidade dos sistemas

Diferente disponibilidade de hardware (memória, CPU), velocidade de internet e queda de energia.

Desafios Centrais

3.Heterogeneidade estatística

Precisa lidar com quantidades distintas de dados disponíveis dependendo da instituição.

4.**Preocupações com a privacidade**

Mesmo só a transferência de modelos pode reveler informação sensível (ex: o parâmetro de HIV/AIDS do modelo mudou).

Eficiência de comunicação

- Método de atualização local
- Esquemas de compressão
- Treinamento descentralizado

- Realizar algumas atualizações locais em paralelo antes do envio do modelo

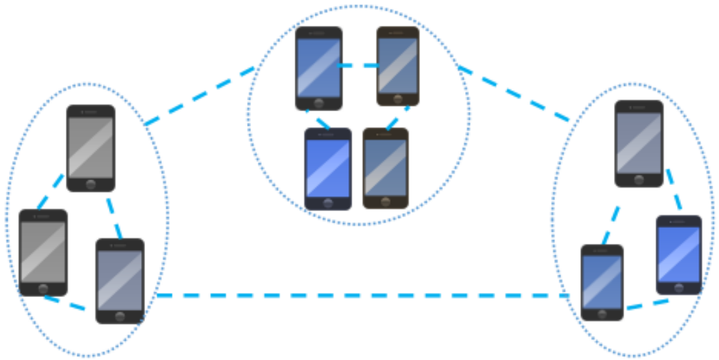
Eficiência de comunicação

- Método de atualização local
- Esquemas de compressão
- Treinamento descentralizado

- Diminuição do tamanho das comunicações via técnicas de esparsificação, subamostragem e quantização

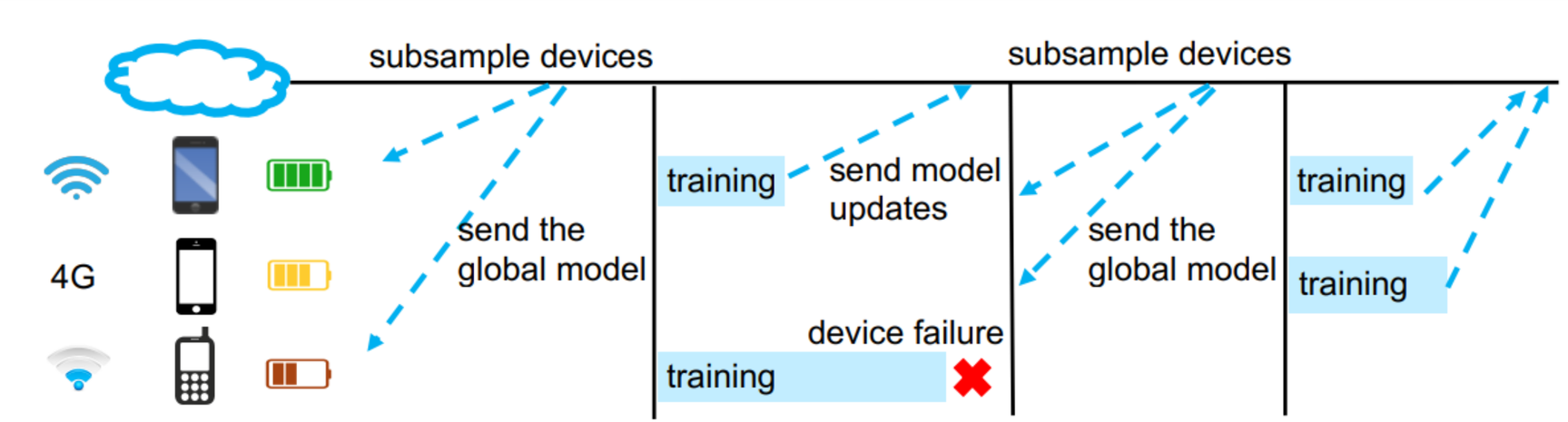
Eficiência de comunicação

- Método de atualização local
- Esquemas de compressão
- Treinamento descentralizado – se comunicar com vizinhos próximos.



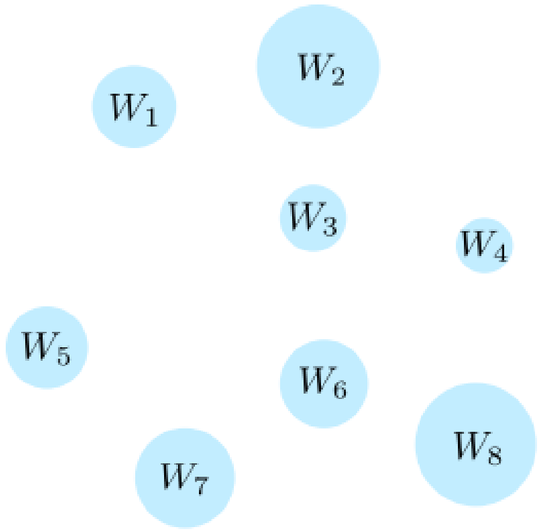
Sistemas heterogêneos

- Hardware (CPU, memória).
- Conectividade de rede de internet.
- Energia da bateria.

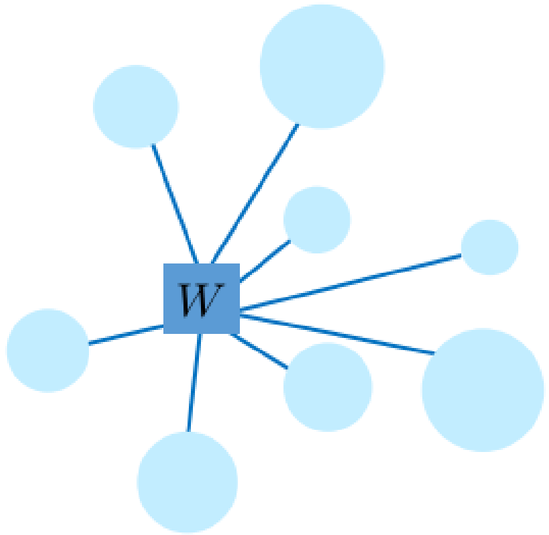


Heterogeneidade estatística

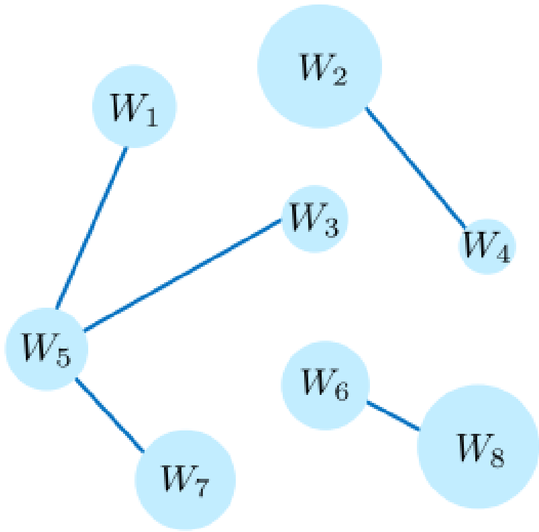
Desafio: dados não são distribuídos de forma idêntica entre os dispositivos



(a) Aprende modelos separados para cada dispositivo.



(b) Ajusta um modelo global para todos os dispositivos.



(c) Aprende modelos parecidos, mas distintos na rede.

Privacidade em machine learning

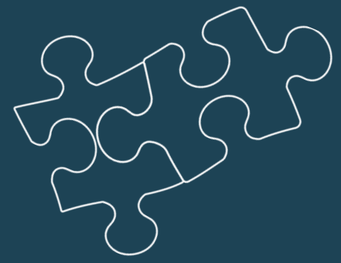
- Privacidade diferencial: adição de ruídos aleatórios nas comunicações
- Criptografia homomórfica: operações realizadas apenas em dados criptografados
- Limitações: custos computacionais e perda de performance.

Direções Futuras

- Esquemas de comunicações extremos (mínimos)
- Redução da quantidade de informação comunicada (métodos de compressão e treinamento local)
- Novos modelos assíncronos: centrados nos dispositivos, que decidem quando interagir com o servidor central
- Diagnóstico de heterogeneidade: métodos estatísticos para identificar dissimilaridade

Conclusão

- Visão geral de aprendizado federado
- Modelos treinados em redes distribuídas
- Foram apresentados os principais desafios
- Tem um longo caminho de desenvolvimento pela frente



Referência

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

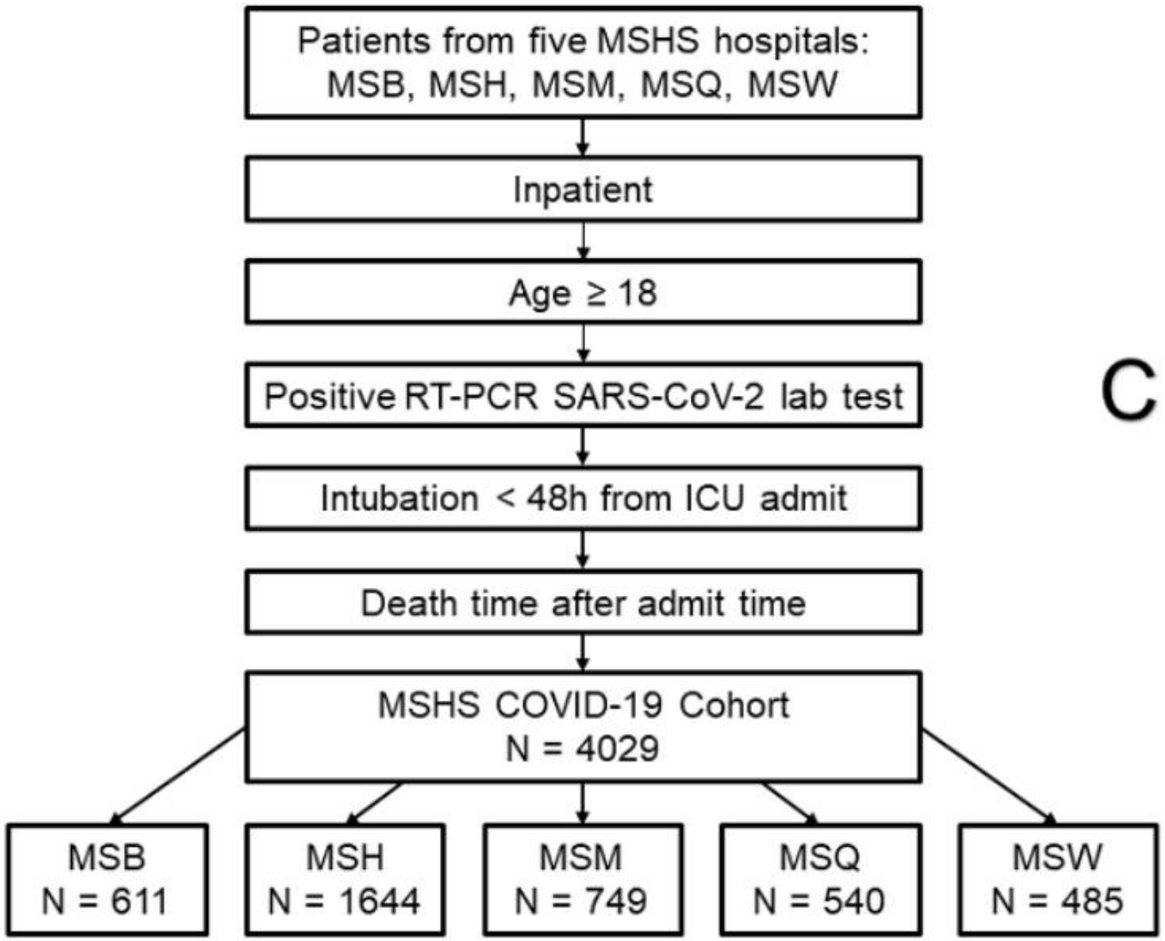
Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach

Authors: Akhil Vaidet al.

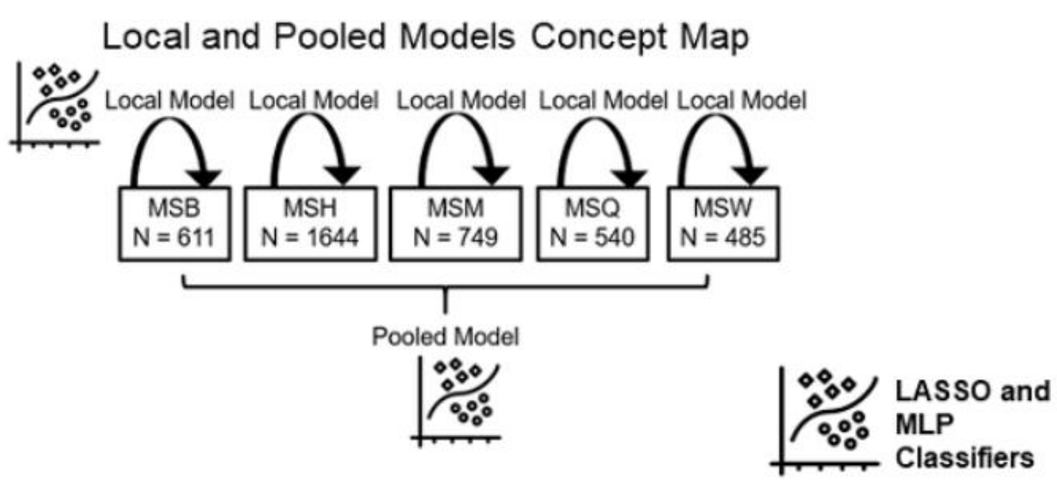
Métodos

- Agregador central foi usado para inicializar o modelo federado com parâmetros aleatórios.
  - Este modelo foi enviado para cada site e treinado por 1 época.
  - A média federada foi realizada de acordo com o número de pontos de dados disponíveis.
  - Parâmetros atualizados do agregador foram então enviados de volta para cada site.

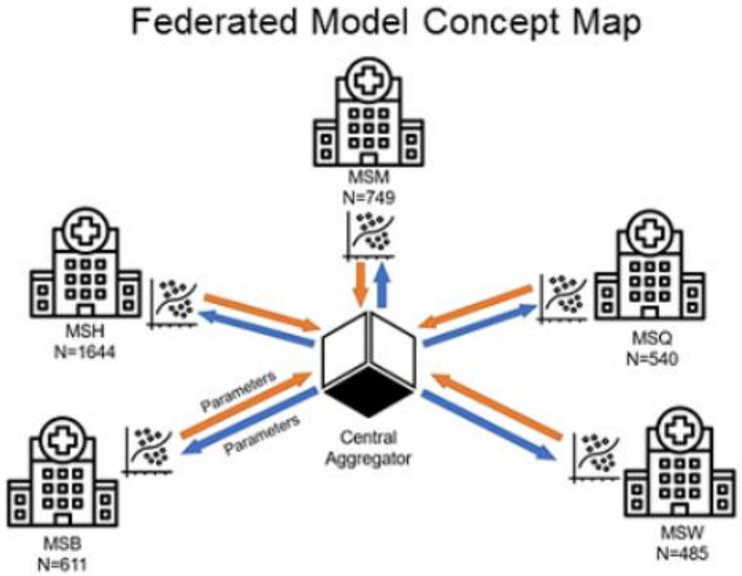
A


Patient Inclusion Criteria
Patients from five MSHS hospitals: MSB, MSH, MSM, MSQ, MSW
Inpatient
Age ≥ 18
Positive RT-PCR SARS-CoV-2 lab test
Intubation < 48h from ICU admit
Death time after admit time
MSHS COVID-19 Cohort N = 4029
MSB N = 611
MSH N = 1644
MSM N = 749
MSQ N = 540
MSW N = 485

B


Local and Pooled Models Concept Map
Local Model
Local Model
Local Model
Local Model
Local Model
MSB N = 611
MSH N = 1644
MSM N = 749
MSQ N = 540
MSW N = 485
Pooled Model
LASSO and MLP Classifiers

C


Federated Model Concept Map
MSM N=749
MSH N=1644
MSQ N=540
Parameters
Parameters
Central Aggregator
MSB N=611
MSW N=485

Resultados

- O modelo federado LASSO superou o modelo local LASSO em todos os hospitais, exceto nos hospitais Mount Sinai Brooklyn e Mount Sinai Queens.
- O modelo pooled de LASSO superou o LASSO federado em todos os hospitais.
- O modelo federado MLP superou o modelo local MLP em todos os hospitais.
- O modelo federado MLP superou o modelo pooled MLP nos hospitais Mount Sinai Morningside e Mount Sinai Queens.

ONLINE LEARNING: A COMPREHENSIVE SURVEY

Authors: Steven C. H. Hoi, Doyen Sahoo, Jing Lu, Peilin Zhao.

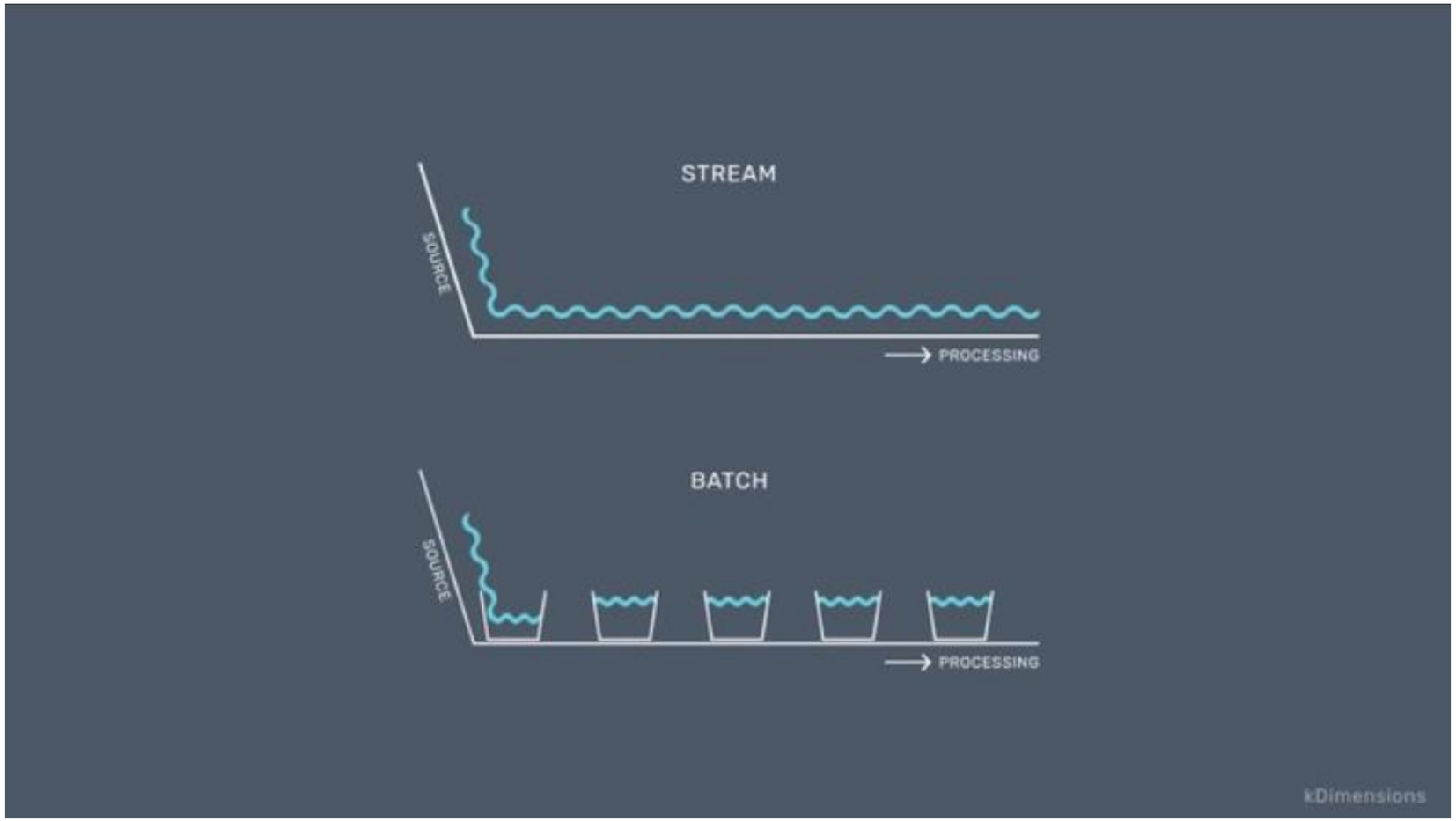
Resumo

- A aprendizagem ocorre a cada nova inclusão de dados
- O objetivo é conseguir predições corretas usando o conhecimento adquirido até então
- Online learning x ML tradicional em lotes (offline: algoritmo aprende com todos os dados – sem updates futuros)
- Foco em online learning supervisionado

O que é online learning?

- Aprendizado no qual os dados chegam sequencialmente, e o preditor é atualizado em cada etapa
- Atualização instantânea do modelo (menor custo)
- Bons para aplicações em grande escala (tamanho e velocidade)

Online learning X ML tradicional (offline)



Tarefas e Aplicações

- Aprendizado online supervisionado
  - classificação binária
  - classificação multiclasses (desfechos exclusivos)
  - classificação multilabel (intersecção de desfechos)
  - regressão

Tarefas e Aplicações

- Aprendizagem online não supervisionada
  - clustering
  - redução de dimensão
  - detecção de anomalias
- Aprendizagem online por reforço

Taxonomia

- Aprendizagem online supervisionada
  - feedback completo para o modelo ao fim de cada rodada
- Aprendizagem online com feedback limitado
  - feedbacks parciais durante o processo (ex. em multiclasse, é comunicado se acertou, mas não o resultado correto)
- Aprendizagem online não supervisionada
  - não há feedback sobre o desfecho durante a aprendizagem

Formulação do Problema

Considerando uma tarefa de classificação binária:

1º A cada rodada são recebidos os dados pelo modelo e realizada a predição

2º Na sequência, são fornecidos os feedback sobre a resposta correta

3º Com base no feedback é mensurada a perda originada pela predição

4º O modelo é atualizado para melhorar o desempenho em novos dados

Online Learning ativo

- Diminuição do custo computacional e de comunicação
- Observa uma sequência de observações não rotuladas
- Decide se deve ou não consultar o rótulo de entrada (com uma condição pré-determinada: como, por exemplo, no caso de alta incerteza) – consultar é caro
- Se consulta, atualiza o modelo
- Caso contrário, não altera o modelo

Online Learning não supervisionado

- **Clustering**: Particionado (Alguns algoritmos: k-Means, k-Medoids); baseado em densidade, micro-clustering e clustering baseado em grade (DBSCAN e seus variantes)

- **Redução de dimensão**: Aprendizagem Subespacial (PCA e ICA) e Aprendizagem múltipla (Multi-dimensional scaling, IsoMap, Locally Linear Embedding, Laplacian Eigenmaps)

Online Learning não supervisionado

- **Detecção de anomalias**: identificação de padrões não esperados nos dados. Agrupadas em: baseadas na distância; baseadas na densidade, baseadas em cluster, métodos estatísticos

Outras terminologias

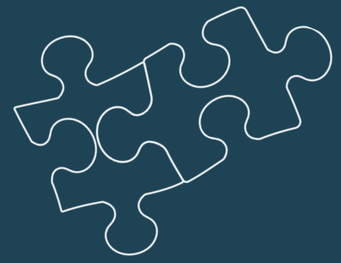
- **Aprendizagem incremental**: aprendizado a partir de fluxos de dados (pode ser online ou offline) que têm espaço e recursos computacionais limitados. Aproximar ao máximo o modelo completo

- **Aprendizagem sequencial**: aprendizado para gráficos ou sequência onde a ordem importa (pode ser online ou offline)

Outras terminologias

- **Aprendizagem interativa**: envolve o ser humano no loop (sistema de aprendizado + usuário). Incorcopa conhecimento de experts por meio de interações contínuas
- **Aprendizagem por reforço**: agente que procura maximizer a recompensa por meio da interação com ambiente (pode ser online ou em lotes)

Conclusão



Agrupamentos por tipo de feedback:

- Aprendizado online supervisionado :
  - feedback completo é sempre revelado
- Aprendizado online com feedback limitado
  - Tradeoff entre aprender com o que conhece e explorar novas informações desconhecidas
- Aprendizado online não supervisionado
  - Sem feedback

Futuro

- Pesquisas sobre “concept drift”: quando o desfecho a ser predito muda
- Aplicações para big data em tempo real: muito volume e velocidade
- Lidar com a variedade de dados (múltiplas fontes, estruturado e não estruturado, etc)
- Lidar com a qualidade dos dados (ruídos, missing, incompletude)

Futuro

- Online deep learning (encontrar equilíbrio entre precisão, eficiência computacional, escalabilidade e complexidade dos modelos)
- Aprendizagem contínua (o problema do esquecimento do aprendizado)

Referência

Hoi, S. C., Sahoo, D., Lu, J., & Zhao, P. (2018). Online learning: A comprehensive survey. arXiv preprint arXiv:1802.02871.