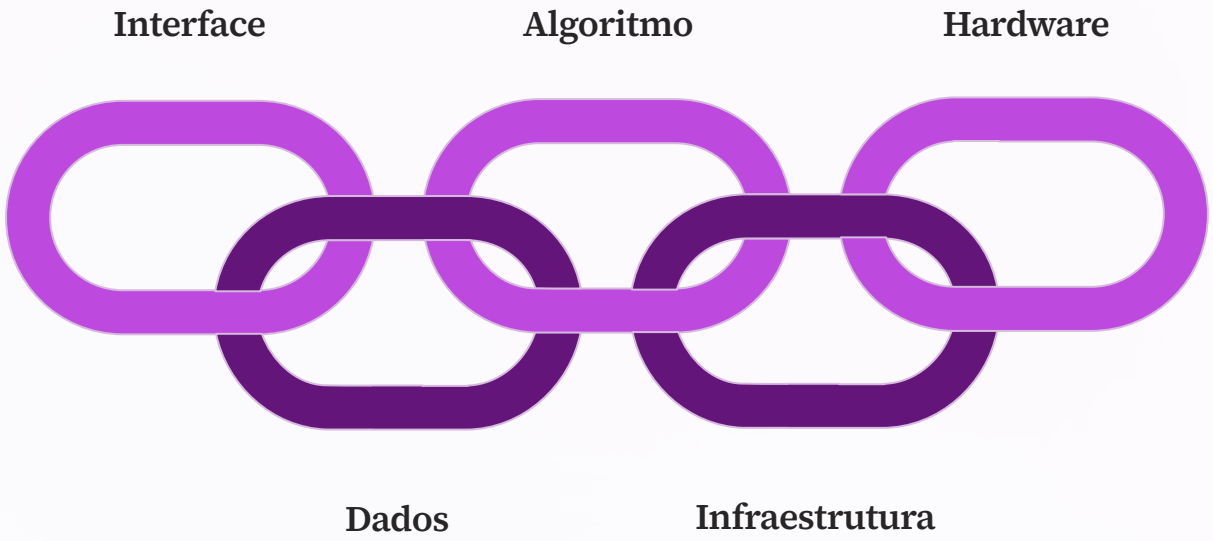
Design de Sistemas ML



A maioria dos cursos de ML foca apenas nos algoritmos, ignorando os outros componentes essenciais.

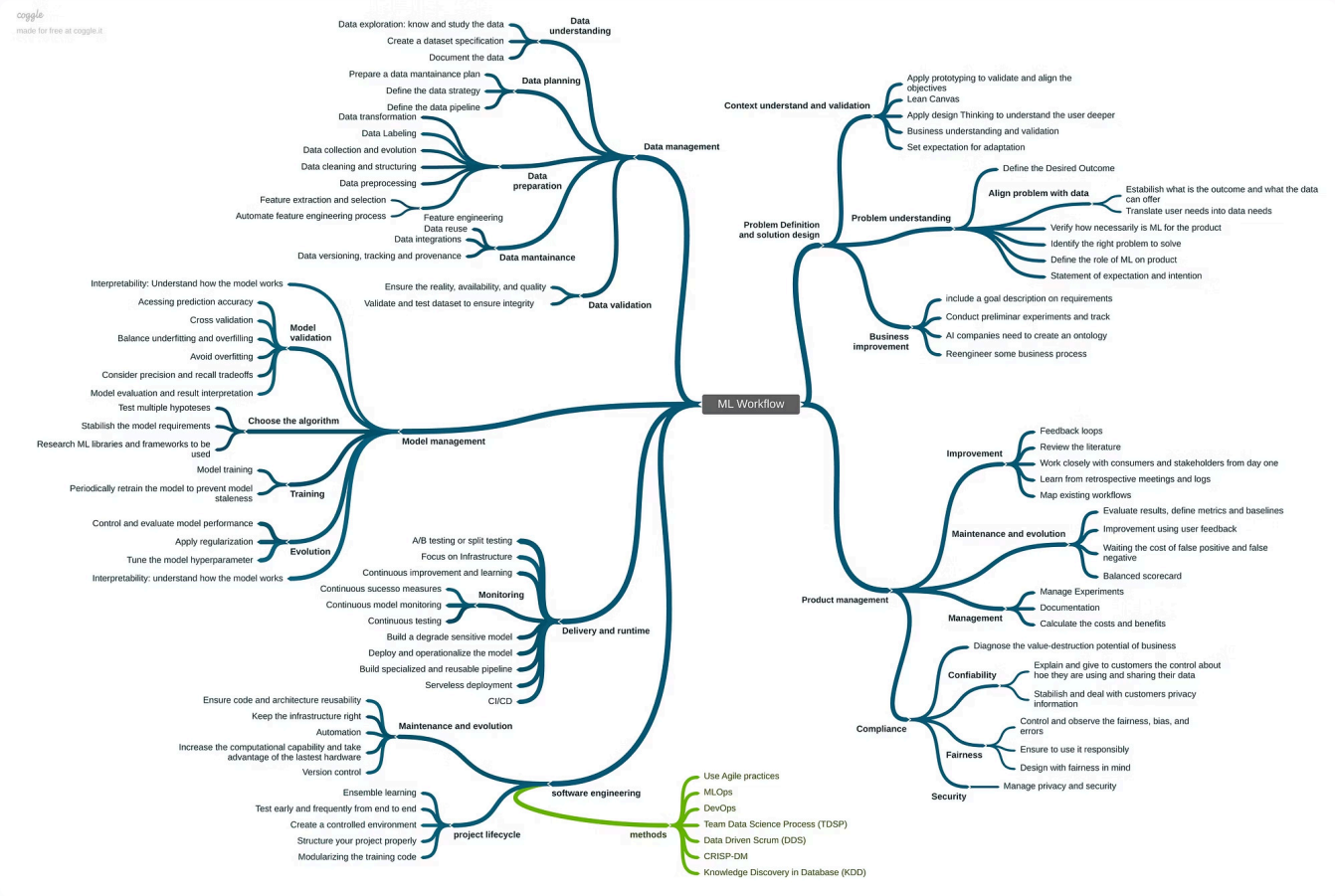
Lab Livre

UnB

Quem somos nós?

O Que é Design de Sistemas ML?

O processo de definir a **interface, algoritmos, dados, infraestrutura e hardware** para um sistema de machine learning que satisfaça **requisitos específicos**.



Requisitos Fundamentais

Confiabilidade

O sistema deve funcionar corretamente mesmo em condições adversas

Escalabilidade

Capacidade de lidar com crescimento em volume de dados e tráfego

Manutenibilidade

Facilidade para atualizar e modificar o sistema ao longo do tempo

Adaptabilidade

Capacidade de se ajustar a mudanças nos dados e no ambiente

Perguntas que Este Curso Ajudará a Responder

- Você treinou um modelo, e agora?
- Quais são os diferentes componentes de um sistema ML?
- Como fazer engenharia de dados e features?
- Como avaliar seus modelos, offline e online?
- Como monitorar e implantar mudanças continuamente?

Este Curso Não Ensinará...

Algoritmos de ML/DL

Sistemas Computacionais

Design UX

Machine Learning: Expectativa

Esta classe não ensinará como fazer isso

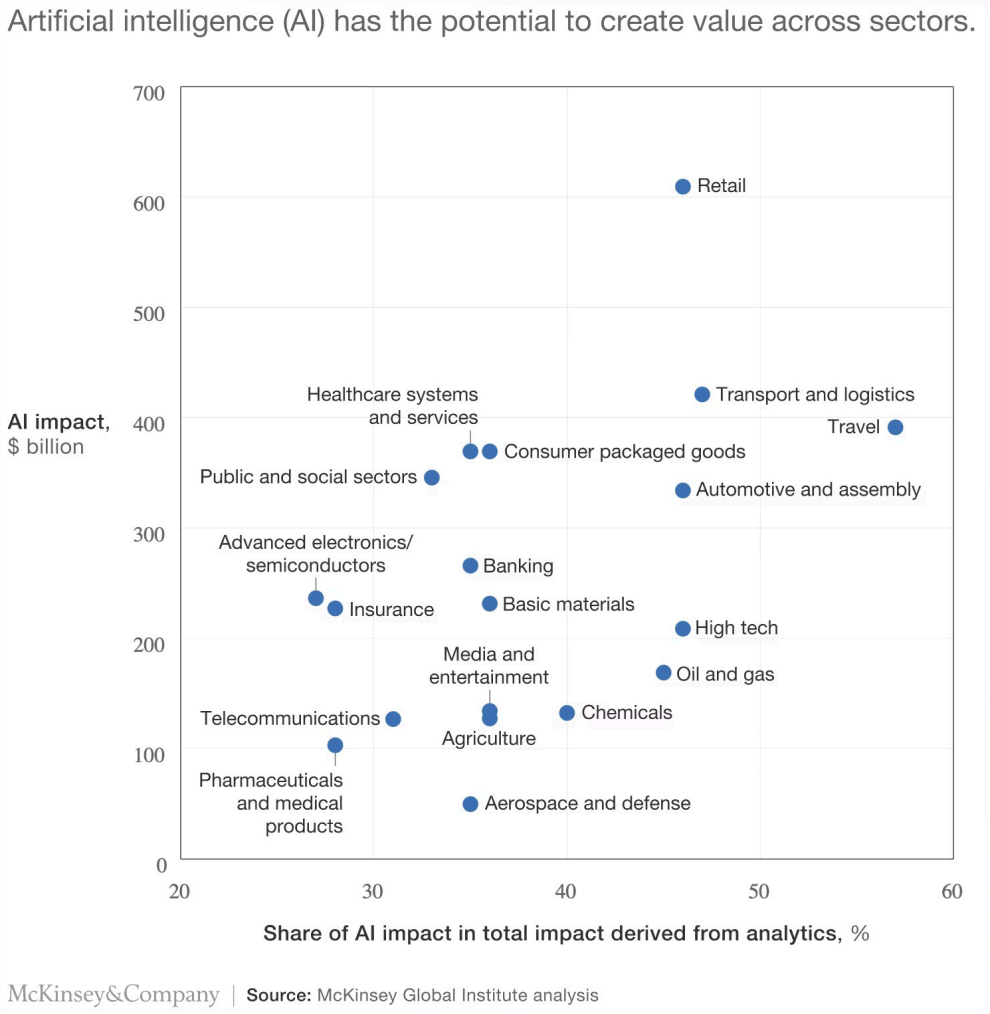
Machine Learning: Realidade

Você provavelmente construirá algo como isto (com bugs, mas legal)

Pré-requisitos

- Conhecimento de princípios de programação/estrutura de dados
- Entendimento de algoritmos ML
- Familiaridade com pelo menos um framework como TensorFlow, PyTorch, JAX
- Conhecimentos básicos de teoria da probabilidade

Valor da IA até 2030



13 trilhões USD

A maior parte estará fora da indústria de internet para consumidores

Precisamos de mais pessoas de áreas não-CS na IA!

Presencial ou Remoto?

Aulas

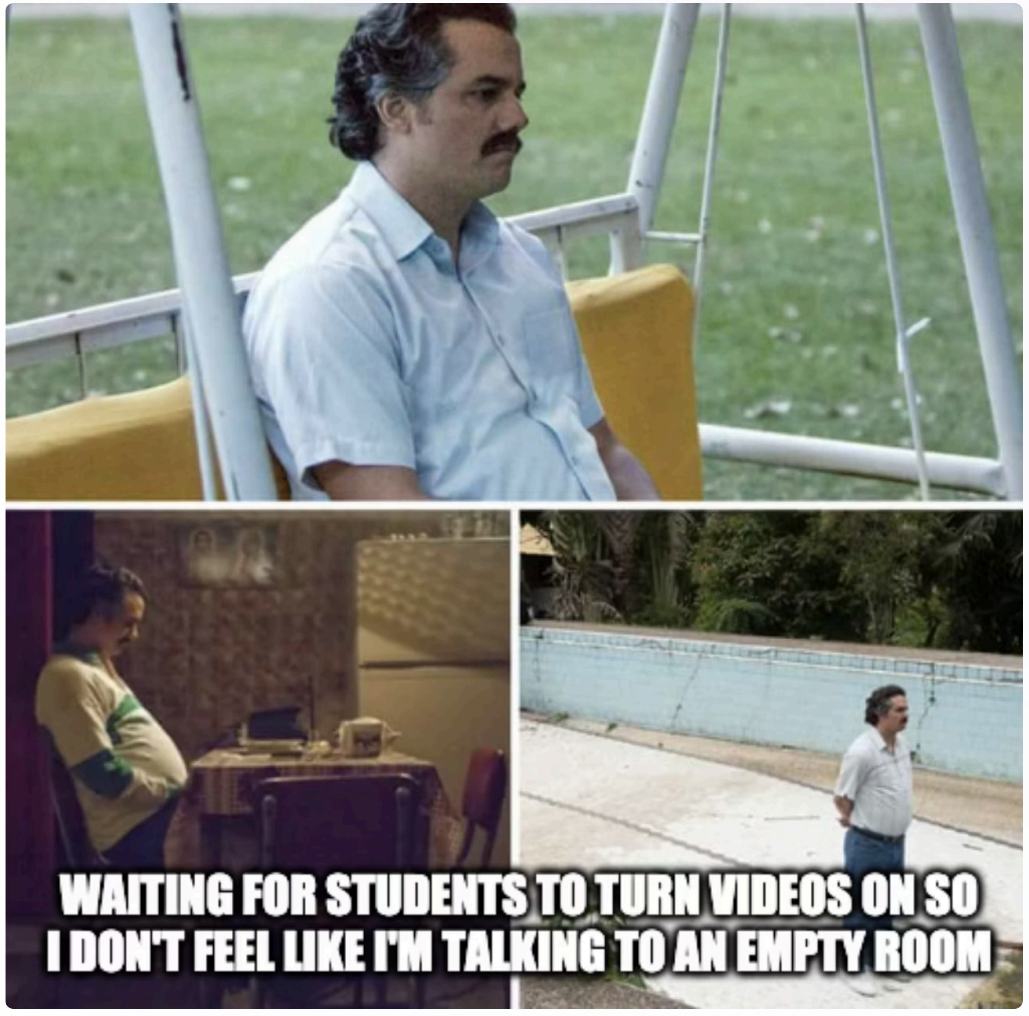
Presenciais

Horários de Atendimento

Mistura de remoto + presencial

Avise-nos se tiver dúvidas ou preocupações

Colabore e Participe das Aulas



Agradecemos se você interagir nas aulas

- Mais feedback visual para ajustarmos o material
- Melhor ambiente de aprendizagem
- Melhor noção de quem está na aula

Avaliação

30%

Tarefas

3 tarefas ao longo do curso

65%

Projeto Final

Aplicação com ML

5%

Participação

Mais informações: https://unb-sistemas-de-machine-learning.github.io/Disciplina/

Projeto Final

Construir Aplicação

Crie uma aplicação movida por ML

Trabalho em Grupo

Grupos de três pessoas obrigatório

Demonstração + Relatório

Formatos criativos encorajados

Procurando colegas para o projeto final?

Sessão de Projetos

Próxima semana: sessão para discutir ideias de projetos e encontrar potenciais colegas de equipe!

Código de Honra

Permissivo mas rigoroso - não nos teste ;)

1

OK pesquisar e perguntar publicamente sobre os sistemas. Cite todas as fontes.

2

NÃO OK pedir a alguém para fazer tarefas/projetos para você.

3

OK discutir questões com colegas. Divulgue seus parceiros de discussão.

4

NÃO OK copiar soluções de colegas.

Equipe do Curso

Carla Rocha, Isaque Alves e Guilherme

⚠️ Trabalho em Andamento ⚠️

Primeira vez que o curso é oferecido

O assunto é novo, não temos todas as respostas

- Nós também estamos aprendendo!

Agradecemos seu:

- **entusiasmo** para experimentar coisas novas
- **paciência** com coisas que não funcionam perfeitamente
- **feedback** para melhorar o curso

Recursos

Site do Curso

unb-sistemas-de-...

SistemaML / Optativa – UnB

– SistemaML / Optativa

Discussões

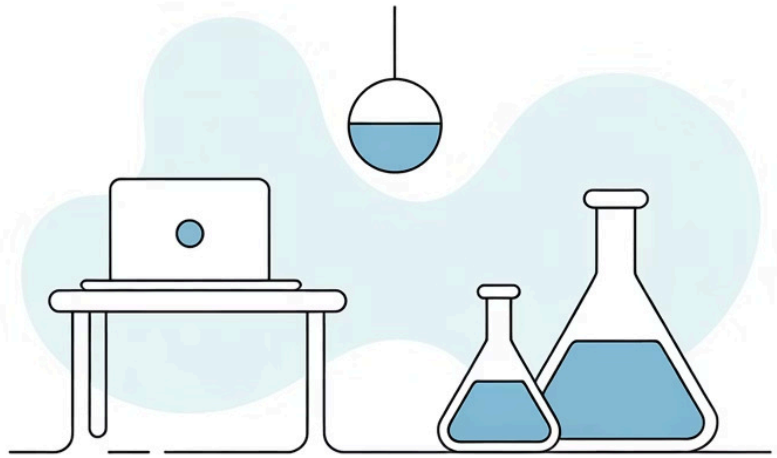
Grupo no telegram

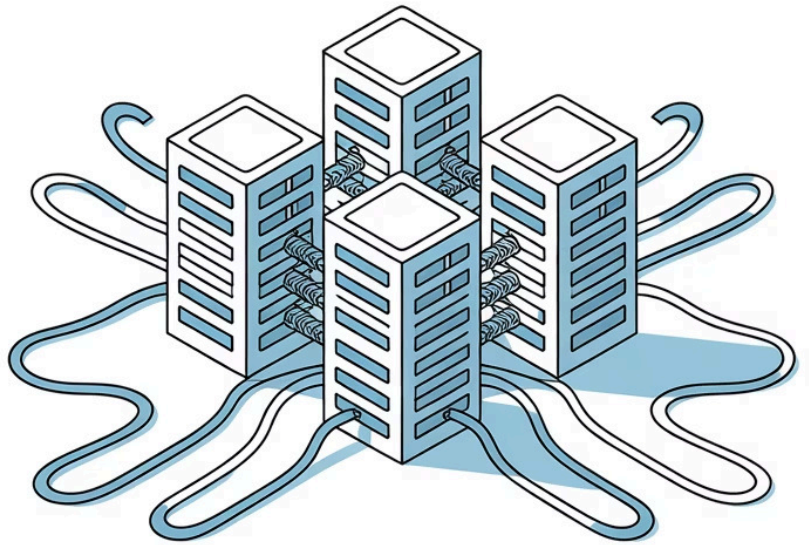
Horários de Atendimento

Começam na próxima semana

Se você se inscreveu sem enviar uma inscrição, envie-nos um e-mail! caguiar@unb.br

2. ML em Pesquisa vs. ML em Produção





Objetivos Diferentes

Pesquisa

Desempenho do modelo*

Produção

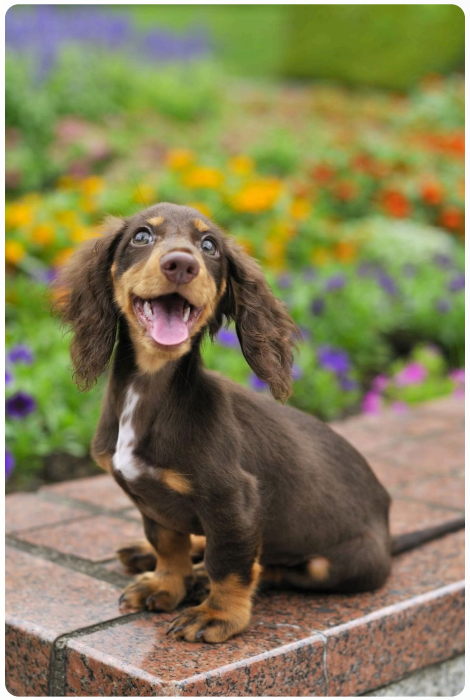
Diferentes stakeholders têm objetivos diferentes

*Está sendo ativamente trabalhado. Veja "Utility is in the Eye of the User: A Critique of NLP Leaderboards" (Ethayarajh e Jurafsky, EMNLP 2020)

Objetivos dos Stakeholders

Equipe ML

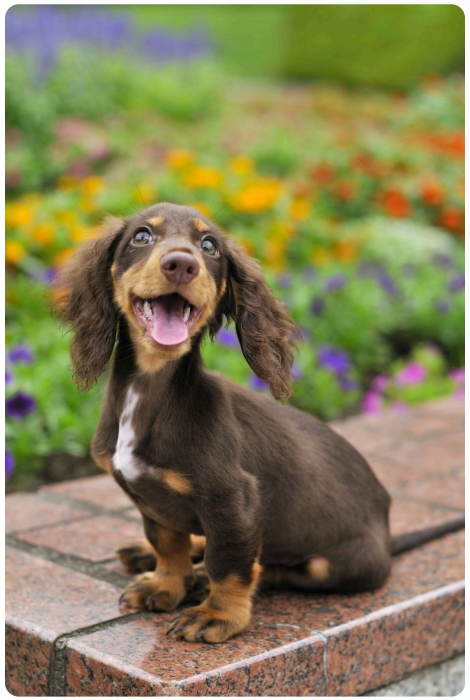
Maior precisão



Diferentes Prioridades

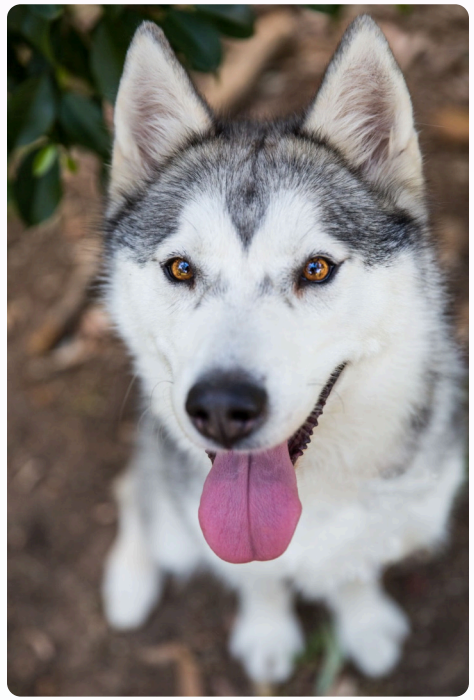
Equipe ML

Maior precisão



Vendas

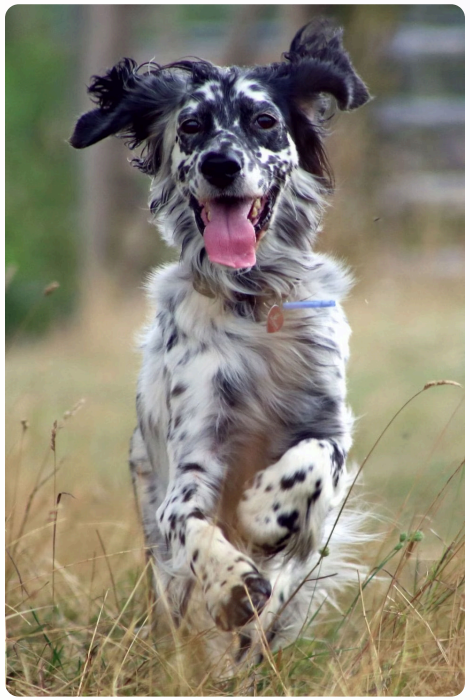
Vende mais anúncios



Objetivos Conflitantes

Produto

Inferência mais rápida



Gestão

Maximiza lucro = demitir equipes ML

ML Estilo Leaderboard

Função de utilidade abrangente

- Desempenho do modelo
- Latência
- Custo de predição
- Interpretabilidade
- Robustez

Adaptável a diferentes casos de uso

Em vez de um ranking para cada conjunto de dados/tarefa, adapta-se às necessidades de cada empresa

Conjuntos de dados dinâmicos

Mudanças realistas de distribuição com diferentes tipos de alterações

Latência vs. Throughput

Latência

Tempo para mover uma folha

Throughput

Quantas folhas em 1 segundo

Prioridade Computacional

Pesquisa

Treinamento rápido, alto throughput

Produção

Inferência rápida, baixa latência ao gerar predições

A Latência Importa

7%

Queda nas conversões

100ms de atraso podem reduzir taxas de conversão em 7% (Estudo Akamai '17)

0.5%

Perda em conversões

30% de aumento na latência custa 0.5% na taxa de conversão (Booking.com '19)

53%

Abandono de página

Usuários abandonam uma página que leva >3s para carregar (Google '16)

Tempo Real vs. Lote

Tempo Real

Baixa latência = alto throughput

Em Lote (Batch)

Alta latência, alto throughput

Dados

Pesquisa

- Limpos
- Estáticos
- Principalmente dados históricos

Produção

- Desordenados
- Constantemente mudando
- Históricos + streaming
- Enviesados, e você não sabe o quanto
- Preocupações com privacidade e regulações

Conceito de "Time Travel" em ML

Extremamente difícil garantir a correção ao longo do tempo

ML em Pesquisa vs. em Produção

Pesquisa

Dados estáticos

Produção

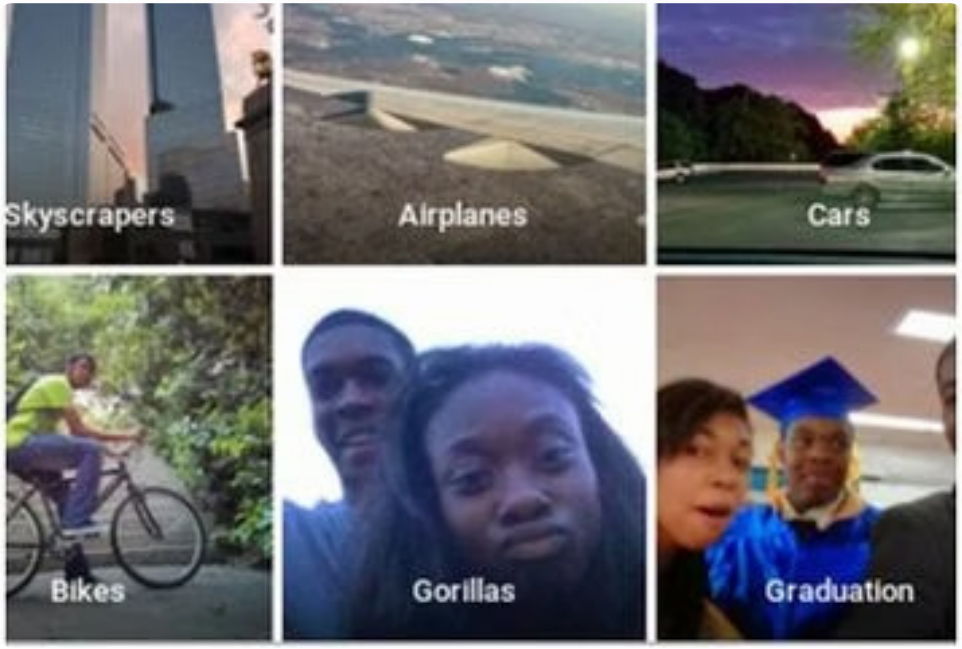
Dados constantemente mudando

Equidade (Fairness)

Bom ter (infelizmente)

Importante

Problemas de Equidade (Fairness)



between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

Interpretabilidade

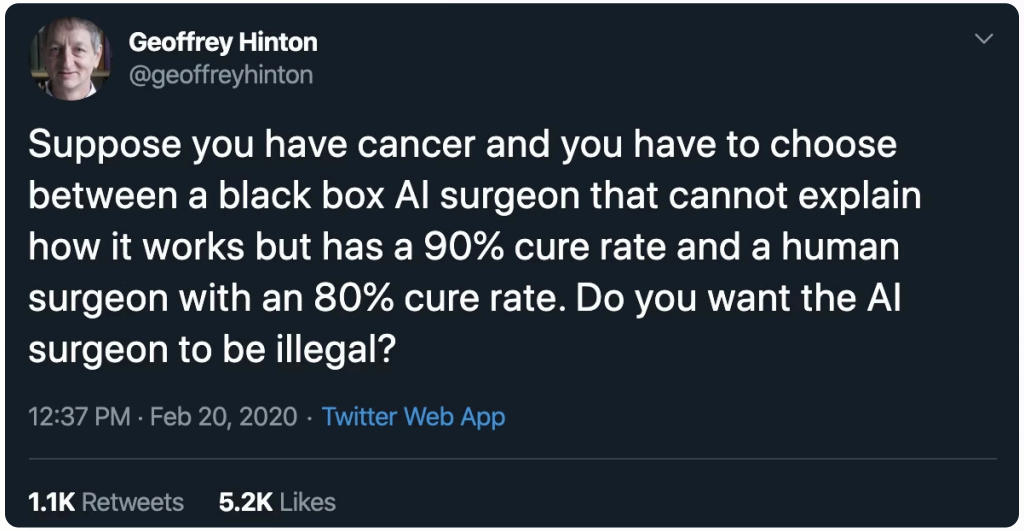
Pesquisa

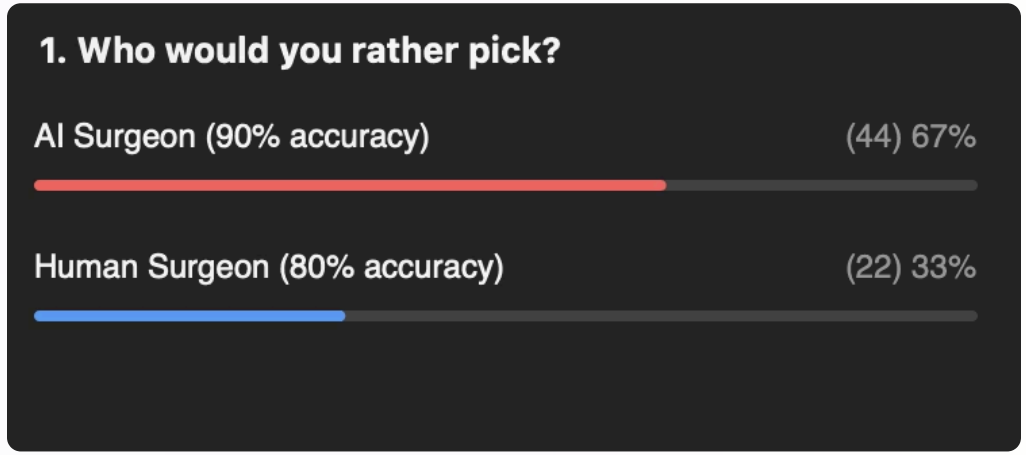
Bom ter

Produção

Importante

Quem Você Escolheria como Cirurgião?





Resultado da pesquisa do Zoom do ano passado

Resumo: ML em Pesquisa vs. em Produção

Objetivos

Desempenho do modelo vs. objetivos de diferentes stakeholders

Prioridade Computacional

Treinamento rápido vs. inferência rápida

Dados

Estáticos vs. constantemente mudando

Equidade e Interpretabilidade

Bom ter vs. essencial

3. Exercício em Grupo

Cada aula, você será designado aleatoriamente para um grupo

Desta vez: 5 pessoas por grupo

8 Minutos - Conhecendo Uns aos Outros

1

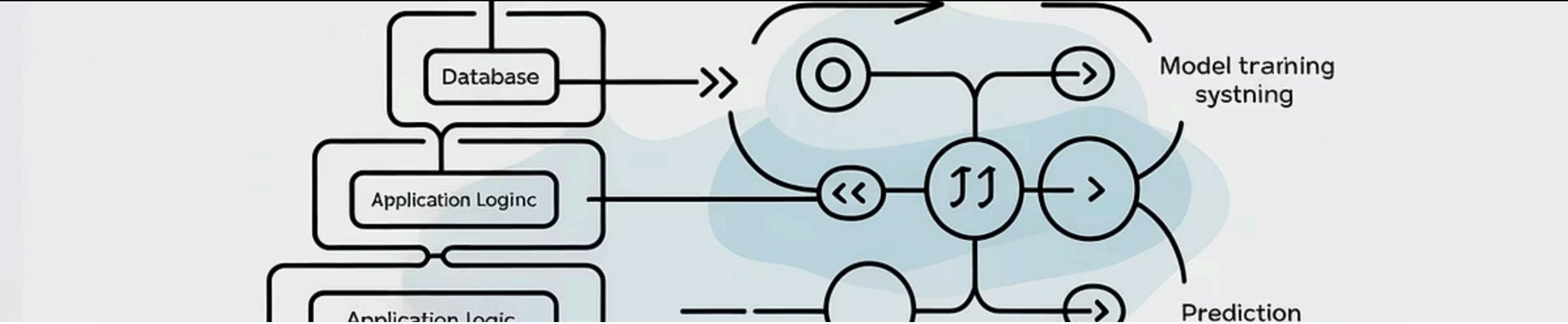
Apresente-se

- De onde você está participando?
- Qual é seu ano/especialização?
- Do que você mais tem medo neste curso?

2

Projetos Finais

- Está procurando colegas para projetos finais?
- O que você gostaria de fazer no projeto final?
- Alguma preocupação sobre seu projeto final?

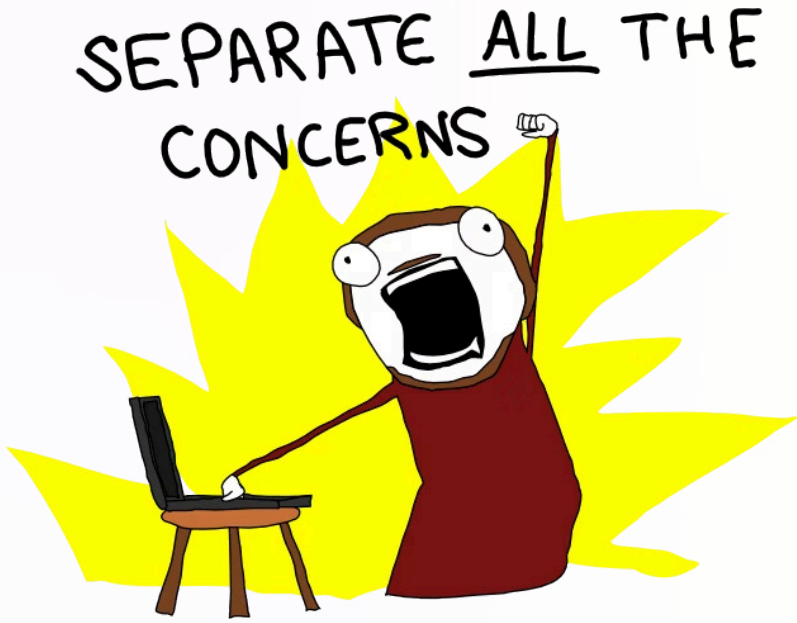


4. Sistemas ML vs. Software Tradicional

Software Tradicional

Separação de Preocupações é um princípio de design para separar um programa de computador em seções distintas, cada uma abordando uma preocupação separada

- Código e dados são separados
- Entradas no sistema não devem alterar o código subjacente



Sistemas ML

Código e Dados Acoplados

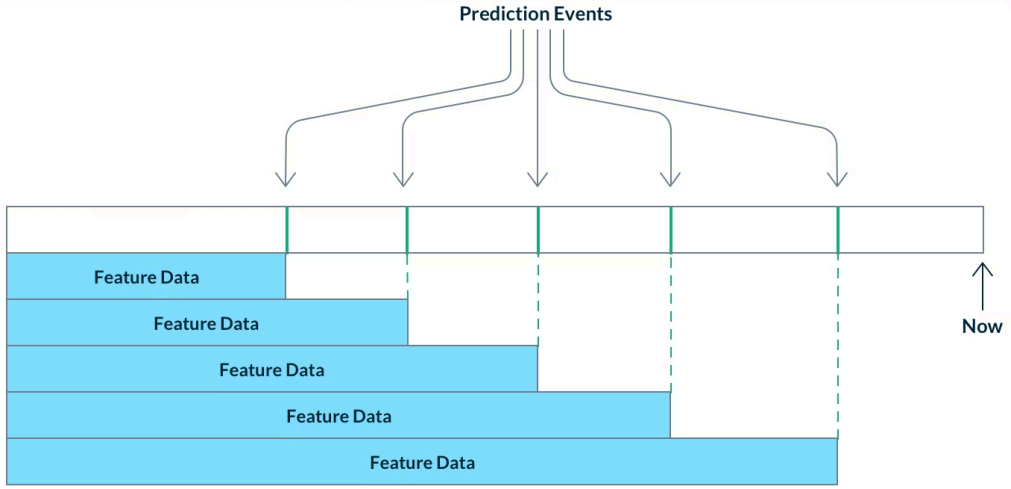
Sistemas ML são parte código, parte dados

Desafio

Não apenas testar e versionar código, mas também testar e versionar dados

A parte difícil!

Testar e Versionar Dados



Timestamp	Label	User ID	Feature Value
2:00	1	1	5
3:00	0	1	19
3:30	0	1	21
5:00	1	1	27
6:00	1	1	42
7:30	0	1	55

Extremamente Difícil

Garantir a correção ao longo do tempo

Não Entre em Pânico

Revisitaremos isso mais tarde!

Versionamento de Dados

- Diffs linha por linha como Git não funcionam com conjuntos de dados
- Não podemos criar ingenuamente múltiplas cópias de grandes conjuntos de dados
- Como mesclar alterações?

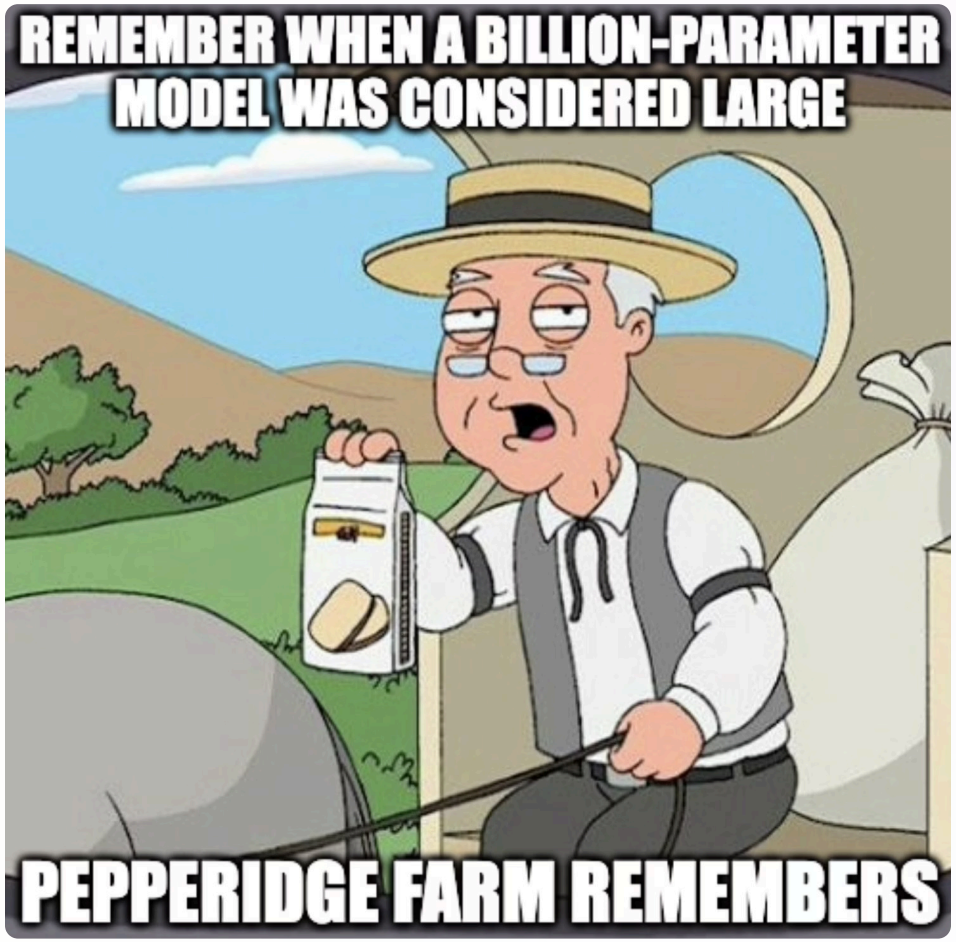
Desafios dos Dados

1. Como validar a correção dos dados?
2. Como testar a utilidade das features?
3. Como detectar mudanças na distribuição dos dados?
4. Como saber se as mudanças são ruins para os modelos sem rótulos verdadeiros?
5. Como detectar dados maliciosos?

Ataques de Envenenamento de Dados

Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning (Chen et al., 2017)

Vulnerabilidades nos Tweets



SWITCH TRANSFORMERS: SCALING TO TRILLION PARAMETER MODELS WITH SIMPLE AND EFFICIENT SPARSITY

William Fedus*
Google Brain
liamfedus@google.com

Barret Zoph*
Google Brain
barretzoph@google.com

Noam Shazeer
Google Brain
noam@google.com

Desafios de Engenharia com Grandes Modelos ML

Muito grandes para caber em dispositivos

Consomem muita energia para funcionar em dispositivos

Muito lentos para serem úteis

- Autocompletar é inútil se demorar mais para fazer uma previsão do que para digitar

Se testes unitários/CI levarem horas, os ciclos de desenvolvimento estagnarão

5. Mitos da Produção ML

Mito #1: Implantar é Difícil

Realidade

Implantar modelos ML está ficando cada vez mais fácil com plataformas modernas

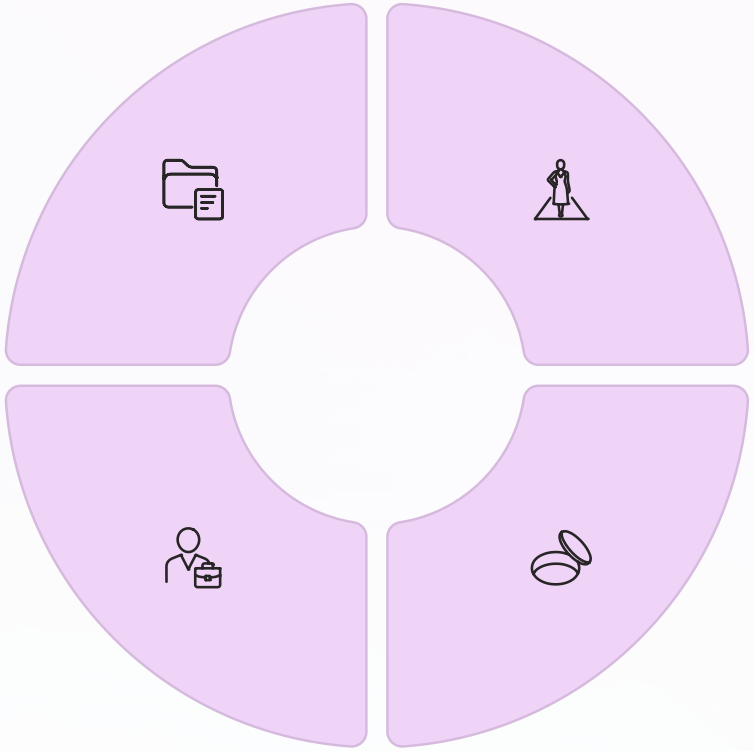
Desafio Real

O difícil é manter o modelo funcionando bem ao longo do tempo com dados em mudança

Mito #2: ML é Apenas Sobre Modelos

Dados

Coleta, limpeza, rotulagem



Modelos

Treinamento, avaliação

Negócios

Métricas, integração

Infraestrutura

Serviço, monitoramento

Mito #3: Quanto Maior o Modelo, Melhor

Velocidade

Modelos menores são mais rápidos para inferência

Custo

Modelos menores usam menos recursos computacionais

Mobilidade

Modelos menores podem rodar em dispositivos com recursos limitados

Próximas Aulas

Aula 2

Modelagem de Dados e Feature Engineering

Aula 3

Treinamento e Avaliação de Modelos

Aula 4

Implantação e Monitoramento

Visite **https://unb-sistemas-de-machine-learning.github.io/Disciplina/** para notas de aula e materiais adicionais