



UNIVERSIDADE  
FEDERAL DO CEARÁ



LogIA

# Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2025

# Agenda

## ① Projeto e aplicação de algoritmos de aprendizagem

- Passos gerais

- Reprodutibilidade

- Crítica e melhoria

## ② Desafios em ambientes reais

## ③ Tópicos e referências adicionais

# Passos gerais

- ① Entender o seu problema.
  - Qual a pergunta a ser respondida? Ela faz sentido?
  - Qual o conhecimento especialista disponível?
  - Qual a literatura mais relacionada ao problema?
  - Quais as variáveis independentes observadas?
  - Quais as variáveis dependentes?
  - Aprendizagem de máquina é realmente necessária?

# Passos gerais

## ① Entender o seu problema.

- Qual a pergunta a ser respondida? Ela faz sentido?
- Qual o conhecimento especialista disponível?
- Qual a literatura mais relacionada ao problema?
- Quais as variáveis independentes observadas?
- Quais as variáveis dependentes?
- Aprendizagem de máquina é realmente necessária?

## ② Construir um conjunto de dados de fácil manipulação.

- Quais dados disponíveis são úteis ao problema?
- Abordagens de mensuração.
- Uso de técnicas de mineração de dados.
- Uso de técnicas de rotulação de dados.
- Armazenamento, versionamento e disponibilização de dados.

# Passos gerais

- ① Entender o seu problema.
  - Qual a pergunta a ser respondida? Ela faz sentido?
  - Qual o conhecimento especialista disponível?
  - Qual a literatura mais relacionada ao problema?
  - Quais as variáveis independentes observadas?
  - Quais as variáveis dependentes?
  - Aprendizagem de máquina é realmente necessária?
- ② Construir um conjunto de dados de fácil manipulação.
  - Quais dados disponíveis são úteis ao problema?
  - Abordagens de mensuração.
  - Uso de técnicas de mineração de dados.
  - Uso de técnicas de rotulação de dados.
  - Armazenamento, versionamento e disponibilização de dados.
- ③ Escolher métricas de avaliação adequadas.
  - Coerentes com a pergunta original.
  - Equilíbrio entre análises quantitativas e qualitativas.

# Passos gerais

- ④ Analisar e pré-processar os dados coletados.
  - Estatísticas básicas.
  - Visualização exploratória.
  - Representação adequada dos dados.
  - Normalização dos dados.
  - Dados discrepantes (outliers)?
  - Dados faltantes?
  - Seleção/combinção de atributos?
  - Dados desbalanceados? Subamostragem? Sobreamostragem?

# Passos gerais

- ④ Analisar e pré-processar os dados coletados.
  - Estatísticas básicas.
  - Visualização exploratória.
  - Representação adequada dos dados.
  - Normalização dos dados.
  - Dados discrepantes (outliers)?
  - Dados faltantes?
  - Seleção/combinção de atributos?
  - Dados desbalanceados? Subamostragem? Sobreamostragem?
- ⑤ Escolher uma família de modelos e um algoritmo de aprendizagem.
  - Comece por abordagens simples.
  - Coerência com a aplicação final.
  - Conheça as suposições e limitações impostas por cada modelo.
  - Teste múltiplas estratégias, mas de forma sistemática.

# Passos gerais

- ⑥ Realizar o treinamento e a avaliação.
  - Ajuste dos hiperparâmetros (grid-search, random search, otimização Bayesiana).
  - Conjunto de teste separado da validação ou validação cruzada aninhada.
  - Dados de validação/teste normalizados a partir das estatísticas dos dados de treinamento.
  - Automatize todo o procedimento.
  - Armazene e visualize as informações geradas.



# Passos gerais

- ⑥ Realizar o treinamento e a avaliação.
  - Ajuste dos hiperparâmetros (grid-search, random search, otimização Bayesiana).
  - Conjunto de teste separado da validação ou validação cruzada aninhada.
  - Dados de validação/teste normalizados a partir das estatísticas dos dados de treinamento.
  - Automatize todo o procedimento.
  - Armazene e visualize as informações geradas.
- ⑦ Reportar os resultados obtidos.
  - Tabelas com valores médios e desvios padrões (e.g.  $\mu \pm \sigma$ ).
  - Matriz de confusão (em problemas de classificação).
  - Análise de resíduos (em problemas de regressão).
  - Figuras: curvas, histogramas, boxplots, ilustração de casos particulares, etc.

Consulte: LONES, Michael, **How to avoid machine learning pitfalls: a guide for academic researchers**, 2023.

# Passos gerais

## Modelos de aprendizagem probabilística

- Verifique quais partes do problema e do modelo apresentam incerteza e podem ser representadas por variáveis aleatórias.

# Passos gerais

## Modelos de aprendizagem probabilística

- Verifique quais partes do problema e do modelo apresentam incerteza e podem ser representadas por variáveis aleatórias.
- Seja pragmático, mas sempre que possível explore modelagens alternativas.

# Passos gerais

## Modelos de aprendizagem probabilística

- Verifique quais partes do problema e do modelo apresentam incerteza e podem ser representadas por variáveis aleatórias.
- Seja pragmático, mas sempre que possível explore modelagens alternativas.
- Escolha um método de inferência compatível com a aplicação.

# Passos gerais

## Modelos de aprendizagem probabilística

- Verifique quais partes do problema e do modelo apresentam incerteza e podem ser representadas por variáveis aleatórias.
- Seja pragmático, mas sempre que possível explore modelagens alternativas.
- Escolha um método de inferência compatível com a aplicação.
- Antes de codificar a solução, escreva todas as expressões que regem o modelo e o procedimento de inferência.

# Passos gerais

## Modelos de aprendizagem probabilística

- Verifique quais partes do problema e do modelo apresentam incerteza e podem ser representadas por variáveis aleatórias.
- Seja pragmático, mas sempre que possível explore modelagens alternativas.
- Escolha um método de inferência compatível com a aplicação.
- Antes de codificar a solução, escreva todas as expressões que regem o modelo e o procedimento de inferência.
- Escolha métricas que avaliem a incerteza nas previsões do modelo.

# Passos gerais

## Modelos de aprendizagem probabilística

- Verifique quais partes do problema e do modelo apresentam incerteza e podem ser representadas por variáveis aleatórias.
- Seja pragmático, mas sempre que possível explore modelagens alternativas.
- Escolha um método de inferência compatível com a aplicação.
- Antes de codificar a solução, escreva todas as expressões que regem o modelo e o procedimento de inferência.
- Escolha métricas que avaliem a incerteza nas previsões do modelo.
- Procure pacotes de software bem estabelecidos: (py)Stan, PyMC, Pyro, GPflow, Tensorflow Probability, etc.

# Passos gerais

- Não custa lembrar:
  - Escolha da família de modelos e do algoritmo de aprendizagem devem levar em consideração as características do problema e dos dados.



# Passos gerais

- Não custa lembrar:
  - Escolha da família de modelos e do algoritmo de aprendizagem devem levar em consideração as características do problema e dos dados.
  - Seleção de hiperparâmetros faz parte do procedimento de aprendizagem.

# Passos gerais

- Não custa lembrar:
  - Escolha da família de modelos e do algoritmo de aprendizagem devem levar em consideração as características do problema e dos dados.
  - Seleção de hiperparâmetros faz parte do procedimento de aprendizagem.
  - Dados de teste nunca devem estar envolvidos nas etapas de treinamento ou ajuste de hiperparâmetros.

# Passos gerais

- Não custa lembrar:
  - Escolha da família de modelos e do algoritmo de aprendizagem devem levar em consideração as características do problema e dos dados.
  - Seleção de hiperparâmetros faz parte do procedimento de aprendizagem.
  - Dados de teste nunca devem estar envolvidos nas etapas de treinamento ou ajuste de hiperparâmetros.
  - As métricas de avaliação devem ser coerentes com a aplicação e com o algoritmo usado.

# Reprodutibilidade

- Automatizar tudo.
  - Coleta e pré-processamento dos dados.
  - Treinamento e avaliação.
  - Geração de tabelas e figuras.

# Reprodutibilidade

- Automatizar tudo.
  - Coleta e pré-processamento dos dados.
  - Treinamento e avaliação.
  - Geração de tabelas e figuras.
- Disponibilizar dados e códigos usados no treinamento e para reportar os resultados, informando as versões dos softwares usados.

# Reprodutibilidade

- Automatizar tudo.
  - Coleta e pré-processamento dos dados.
  - Treinamento e avaliação.
  - Geração de tabelas e figuras.
- Disponibilizar dados e códigos usados no treinamento e para reportar os resultados, informando as versões dos softwares usados.
- No caso de algoritmos estocásticos, fixar e informar o random seed usado.

# Os resultados não estão adequados?

- Colete (gere?) mais dados.
- Reveja/recombine/selecione os atributos usados.
- Verifique a possibilidade de overfitting/underfitting.
- Analise (ou peça para alguém analisar) todos os passos do treinamento.
- Analise os erros encontrados e suas possíveis causas.
- Escolha outra função custo/objetivo.
- Escolha outra família de modelos de aprendizagem e/ou algoritmos de aprendizagem.

# Agenda

## ① Projeto e aplicação de algoritmos de aprendizagem

Passos gerais

Reprodutibilidade

Crítica e melhoria

## ② Desafios em ambientes reais

## ③ Tópicos e referências adicionais



---

# Hidden Technical Debt in Machine Learning Systems

---

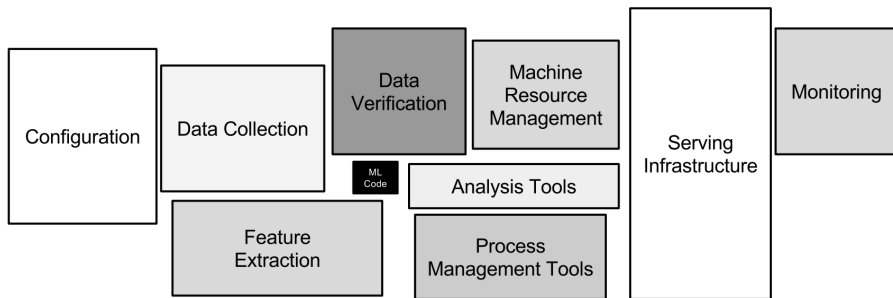
**D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips**  
{dsculley, gholt, dgg, edavydov, toddphillips}@google.com  
Google, Inc.

**Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison**  
{ebner, vchaudhary, mwyong, jfcrespo, dennison}@google.com  
Google, Inc.

## Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

SCULLEY, David *et al.* **Hidden technical debt in machine learning systems.** Advances in neural information processing systems, v. 28, 2015.



# Débito técnico em Aprendizagem de Máquina

- Erosão de fronteiras
  - Acoplamento (*Changing Anything Changes Everything*).
  - Consumidores não declarados.
- Dependências relacionadas aos dados
  - Fontes de dados instáveis.
  - Dependências subutilizadas.
- Loops de realimentação
- Glue code e pipeline jungles
- “Dívidas” de configuração
- Mudanças no mundo externo

---

# Challenges in Deploying Machine Learning: a Survey of Case Studies

---

**Andrei Paleyes**

Department of Computer Science  
University of Cambridge  
ap2169@cam.ac.uk

**Raoul-Gabriel Urma**

Cambridge Spark  
raoul@cambridgespark.com

**Neil D. Lawrence**

Department of Computer Science  
University of Cambridge  
nd121@cam.ac.uk

## Abstract

In recent years, machine learning has received increased interest both as an academic research field and as a solution for real-world business problems. However, the deployment of machine learning models in production systems can present a number of issues and concerns. This survey reviews published reports of deploying machine learning solutions in a variety of use cases, industries and applications and extracts practical considerations corresponding to stages of the machine learning deployment workflow. Our survey shows that practitioners face challenges at each stage of the deployment. The goal of this paper is to layout a research agenda to explore approaches addressing these challenges.

PALEYES, Andrei; URMA, Raoul-Gabriel; LAWRENCE, Neil D. **Challenges in deploying machine learning: a survey of case studies**. NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses.

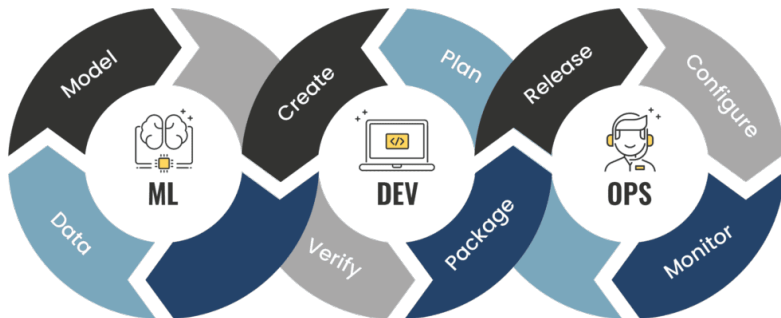
| Deployment Stage   | Deployment Step           | Considerations, Issues and Concerns                                                    |
|--------------------|---------------------------|----------------------------------------------------------------------------------------|
| Data management    | Data collection           | Data discovery                                                                         |
|                    | Data preprocessing        | Data dispersion<br>Data cleaning                                                       |
|                    | Data augmentation         | Labeling of large volumes of data<br>Access to experts<br>Lack of high-variance data   |
|                    | Data analysis             | Data profiling                                                                         |
| Model learning     | Model selection           | Model complexity<br>Resource-constrained environments<br>Interpretability of the model |
|                    | Training                  | Computational cost<br>Environmental impact                                             |
|                    | Hyper-parameter selection | Resource-heavy techniques<br>Hardware-aware optimization                               |
| Model verification | Requirement encoding      | Performance metrics<br>Business driven metrics                                         |
|                    | Formal verification       | Regulatory frameworks                                                                  |
|                    | Test-based verification   | Simulation-based testing                                                               |

PALEYES, Andrei; URMA, Raoul-Gabriel; LAWRENCE, Neil D. **Challenges in deploying machine learning: a survey of case studies**. NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses.

|                       |                  |                                                                                                                         |
|-----------------------|------------------|-------------------------------------------------------------------------------------------------------------------------|
| Model deployment      | Integration      | Operational support<br>Reuse of code and models<br>Software engineering anti-patterns<br>Mixed team dynamics            |
|                       | Monitoring       | Feedback loops<br>Outlier detection<br>Custom design tooling                                                            |
|                       | Updating         | Concept drift<br>Continuous delivery                                                                                    |
| Cross-cutting aspects | Ethics           | Country-level regulations<br>Focus on technical solution only<br>Aggravation of biases<br>Authorship<br>Decision making |
|                       | End users' trust | Involvement of end users<br>User experience<br>Explainability score                                                     |
|                       | Security         | Data poisoning<br>Model stealing<br>Model inversion                                                                     |

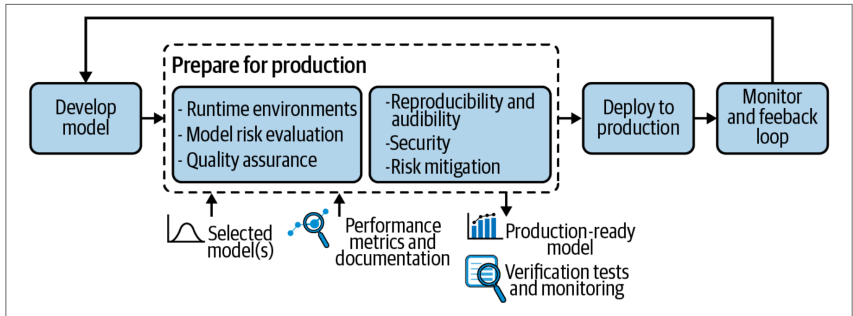
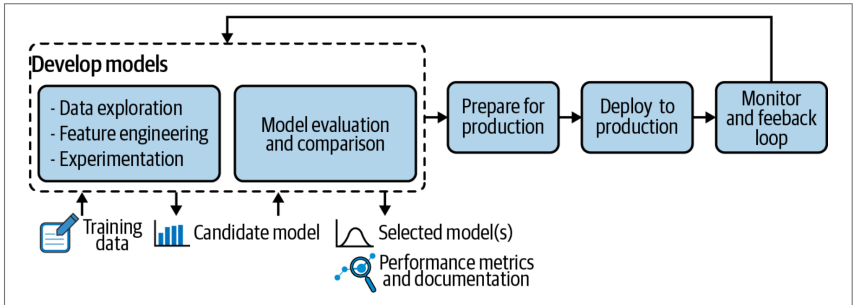
# MLOps

- Conjunto de práticas que visam disponibilizar e manter modelos de ML de maneira confiável e eficiente em ambientes reais.



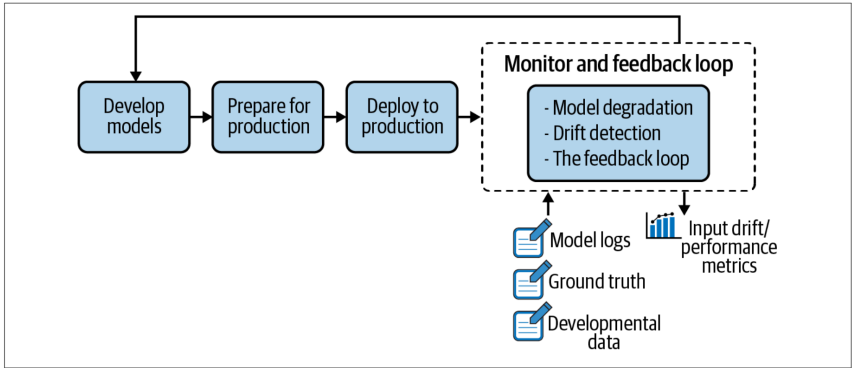
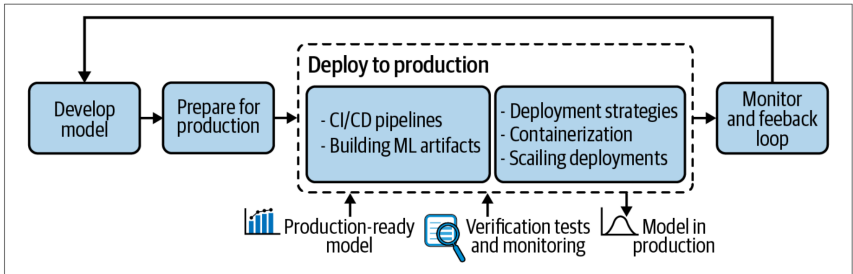
Consulte: SHANKAR, Shreya *et al.*, **Operationalizing Machine Learning: An Interview Study**, 2022.

# MLOps





# MLOps



# Agenda

## ① Projeto e aplicação de algoritmos de aprendizagem

Passos gerais

Reprodutibilidade

Crítica e melhoria

## ② Desafios em ambientes reais

## ③ Tópicos e referências adicionais

# Tópicos e referências adicionais

- Lawrence, **Data Science and Digital Systems: The 3Ds of Machine Learning**, 2019.
  - Discute 3 aspectos fundamentais de sistemas inteligentes: dados, design e desenvolvimento, reforçando a necessidade de uma abordagem *data first*.

# Tópicos e referências adicionais

- Lawrence, **Data Science and Digital Systems: The 3Ds of Machine Learning**, 2019.
  - Discute 3 aspectos fundamentais de sistemas inteligentes: dados, design e desenvolvimento, reforçando a necessidade de uma abordagem *data first*.
- Lawrence, **Living Together: Mind and Machine Intelligence**, 2017.
  - Discute as diferenças entre as inteligências humana e de máquina, focando em fatores de “encarnação” (*embodiment*) e propondo uma visão baseada em processos duais de cognição.

# Tópicos e referências adicionais

- Lawrence, **Data Science and Digital Systems: The 3Ds of Machine Learning**, 2019.
  - Discute 3 aspectos fundamentais de sistemas inteligentes: dados, design e desenvolvimento, reforçando a necessidade de uma abordagem *data first*.
- Lawrence, **Living Together: Mind and Machine Intelligence**, 2017.
  - Discute as diferenças entre as inteligências humana e de máquina, focando em fatores de “encarnação” (*embodiment*) e propondo uma visão baseada em processos duais de cognição.
- Malik, **A Hierarchy of Limitations in Machine Learning**, 2020.
  - Estrutura as limitações conceituais, procedurais e estatísticas de modelos de aprendizagem de máquina, especialmente em aplicações reais.

# Tópicos e referências adicionais

- Lawrence, **Data Science and Digital Systems: The 3Ds of Machine Learning**, 2019.
  - Discute 3 aspectos fundamentais de sistemas inteligentes: dados, design e desenvolvimento, reforçando a necessidade de uma abordagem *data first*.
- Lawrence, **Living Together: Mind and Machine Intelligence**, 2017.
  - Discute as diferenças entre as inteligências humana e de máquina, focando em fatores de “encarnação” (*embodiment*) e propondo uma visão baseada em processos duais de cognição.
- Malik, **A Hierarchy of Limitations in Machine Learning**, 2020.
  - Estrutura as limitações conceituais, procedurais e estatísticas de modelos de aprendizagem de máquina, especialmente em aplicações reais.
- Fazelpour e Lipton, **Algorithmic Fairness from a Non-ideal Perspective**, 2020.
  - Discute questões de justiça e equidade em sistemas inteligentes contrastando visões filosóficas de mundo ideal versus não-ideal.

# Tópicos e referências adicionais

- Lipton, **The Mythos of Model Interpretability**, 2018.
  - Aborda a dualidade entre a importância da interpretabilidade como propriedade desejável e a dificuldade de defini-la objetivamente.

# Tópicos e referências adicionais

- Lipton, **The Mythos of Model Interpretability**, 2018.
  - Aborda a dualidade entre a importância da interpretabilidade como propriedade desejável e a dificuldade de defini-la objetivamente.
- Cheeseman, **In Defense of Probability**, 1985.
  - Defende a teoria da probabilidade como suficiente para a tarefa de racionalizar na presença de incerteza, enfatizando sua função como representação de crenças, em oposição a interpretações frequentistas.



# Tópicos e referências adicionais

- Lipton, **The Mythos of Model Interpretability**, 2018.
  - Aborda a dualidade entre a importância da interpretabilidade como propriedade desejável e a dificuldade de defini-la objetivamente.
- Cheeseman, **In Defense of Probability**, 1985.
  - Defende a teoria da probabilidade como suficiente para a tarefa de racionalizar na presença de incerteza, enfatizando sua função como representação de crenças, em oposição a interpretações frequentistas.
- Ghahramani, **Probabilistic machine learning and artificial intelligence**, 2015.
  - Um survey que motiva o uso da linguagem probabilística em tarefas de aprendizagem, sobretudo na representação da incerteza.

# Tópicos e referências adicionais

- Lipton, **The Mythos of Model Interpretability**, 2018.
  - Aborda a dualidade entre a importância da interpretabilidade como propriedade desejável e a dificuldade de defini-la objetivamente.
- Cheeseman, **In Defense of Probability**, 1985.
  - Defende a teoria da probabilidade como suficiente para a tarefa de racionalizar na presença de incerteza, enfatizando sua função como representação de crenças, em oposição a interpretações frequentistas.
- Ghahramani, **Probabilistic machine learning and artificial intelligence**, 2015.
  - Um survey que motiva o uso da linguagem probabilística em tarefas de aprendizagem, sobretudo na representação da incerteza.
- Wilson, **The Case for Bayesian Deep Learning**, 2020.
  - Argumenta sobre a importância de incluir o conceito Bayesiano de marginalização de valores incertos em frameworks de deep learning.