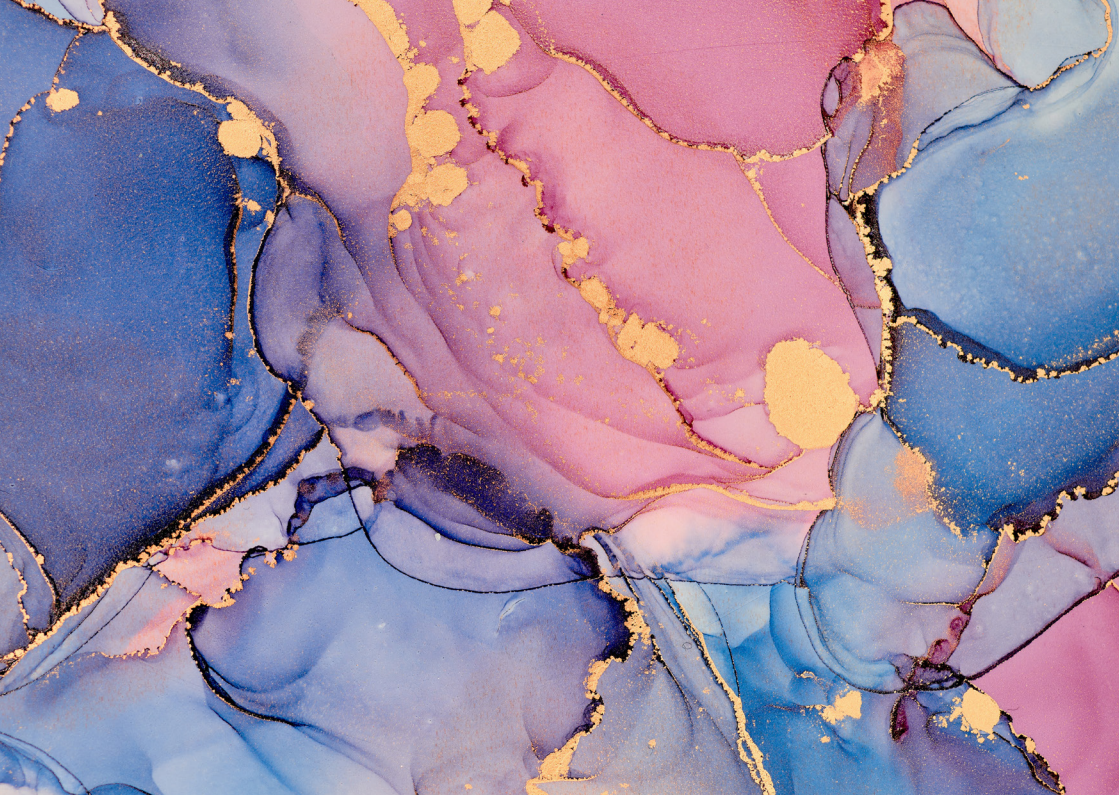


MLOps erfolgreich umsetzen

So meistern Sie die Komplexität beim Aufbau und Einsatz von maschinellem Lernen in Ihrem Unternehmen.

thoughtworks | aws

Bei MLOps gibt es kein Patentrezept, aber ein großes Fehlerpotenzial

Zu diesem E-Book: Dieses E-Book erläutert das Konzept von MLOps sowie dessen Chancen und Herausforderungen. Es geht auch auf Continuous Delivery for Machine Learning (CD4ML) ein – den von Thoughtworks entwickelten Ansatz für eine erfolgreiche Umsetzung von MLOps. Anhand verschiedener Tools und Anbieterlösungen kann Amazon Web Services (AWS) zu Ihrem MLOps-Erfolg beitragen.

Bis zu produktionsreifen ML-Modellen ist es ein langer Weg

Data-Science-Teams in Unternehmen haben gezeigt, dass maschinelles Lernen (ML) zahlreiche Möglichkeiten bietet, die Effizienz zu verbessern, Prozesse zu automatisieren, Kosten zu senken und das Kundenerlebnis zu verbessern. Weitaus schwieriger ist es jedoch, ML-Modelle in IT-Infrastrukturen zu deployen und zu integrieren, damit diese Vorteile in

87%

der Data-Science-Projekte schaffen es nicht bis zur Produktionsreife.

—VentureBeat[1]

[1]VentureBeat, Why do 87% of data science projects never make it into production?

der Alltagspraxis tatsächlich genutzt werden können. Die Echtzeitdaten, mit denen ML-Modelle trainiert werden, ändern sich ständig. Laut Harvard Business Review brauchen die meisten Organisationen zwischen vier Monaten und einem Jahr, bis sie ihr erstes minimal funktionsfähiges ML-Produkt (Minimum Viable Product, MVP) auf den Markt bringen[2]. ML-Modelle in Produktion zu bringen, ist gegenüber dem Deployment herkömmlicher Software mit neuen Hürden ver-bunden. Die Tools und Prozesse der professionellen Softwareentwicklung können hierbei von Nutzen sein.

„Der springende Punkt im Bereich Infrastruktur und Operations wird sein, zu lernen, wie ML-Produkte zur Produktionsreife geführt werden. ML und KI stellen Ops-Teams vor nicht alltägliche Herausforderungen."

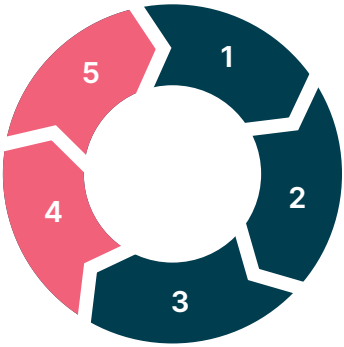
—Mike Loukides, O'Reilly[3]

[2]Harvard Business Review, How to Choose Your First AI Project
[3]O'Reilly, Radar trends to watch: January 2020

Die Implementierung von ML-Modellen kann länger dauern als gedacht

ML und künstliche Intelligenz (KI) haben gewaltiges Potenzial – sie erleichtern den Menschen nicht nur Entscheidungen oder optimieren Prozesse, sondern transformieren Organisationen. Diese Möglichkeiten lassen sich aber nur ausschöpfen, wenn ML-Modelle deployed werden. Damit die Modelle Produktionsreife erlangen, gilt es zunächst, Daten zu erfassen, zu speichern, zu bereinigen und zu kuratieren. Um so Erkenntnisse zu extrahieren und ML-Modelle zu trainieren. Einmal live lassen sich die Modelle dann mit neuen Daten weiter trainieren und verbessern. Diesen Vorgang bezeichnet man auch als „Cycle of Intelligence“.

Cycle of Intelligence bei ML-Modellen



Data Science

1. **Daten:** Beschaffen
   Parametern zugeordnete Werte.
2. **Informationen:** Speichern, bereinigen, kuratieren, Merkmale extrahieren
   Aussagekräftige Daten, die zur Nutzung und Analyse geeignet sind.
3. **Erkenntnisse:** Modellieren
   Verstehen, vorhersagen, klassifizieren und erkennen.

Data Engineering

4. **Entscheidung:** Produktiv nutzen
   Maßnahmen planen und priorisieren. Hypothesen prüfen.
5. **Aktion:** Ausführen
   Realität verändern!

Einer Studie von Algorithmia aus dem Jahr 2020 zufolge dauert es immer länger, bis ein Modell einsatzbereit ist. 2020 brauchten 64 Prozent der Organisationen einen Monat oder länger bis zur Implementierung[4]. Für viele Organisationen kann ein Iterationszyklus im Cycle of Intelligence also sehr viel Zeit in Anspruch nehmen.

Mit MLOps können Sie den Feedbackzyklus verkürzen. Ob Sie ein problembehaftetes Modell zurücknehmen oder ein neues Modell binnen weniger Tage oder gar Stunden freigeben müssen: MLOps vereinfacht dies. Da hiermit die technischen Beschränkungen des Prozesses beseitigt werden, treffen Sie eine reine Geschäftsentscheidung.

[4]Algorithmia, **2021 enterprise trends in machine learning**

Die Trennung zwischen Data Scientists und Engineering-Teams

Data Scientists arbeiten häufig isoliert in lokalen Umgebungen auf ihren persönlichen Notebooks. Für ihre an Forschung & Entwicklung (F&E)-orientierten Aufgaben ist das ideal, spiegelt aber nicht unbedingt die Realität der Produktivumgebung wider, in der ihre ML-Modelle laufen sollen.

Reale Produktivsysteme beinhalten mehrere Umgebungen mit verschiedenen Datenquellen sowie Interaktionen mit anderen Produktivsystemen. Produktivsysteme müssen von hoher Qualität, zuverlässig, skalierbar, sorgfältig getestet, wartungsfähig und überprüfbar sein. Infolgedessen sind Data Scientists häufig gezwungen, ihre Arbeit an das Engineering-Team zu übergeben, damit Letzteres die Modelle integriert – oder manchmal auch komplett umschreibt.

Oftmals sind die beiden Bereiche in komplett unterschiedlichen Teilen der Organisation angesiedelt, was die Übergabe zusätzlich beeinträchtigt. Getrennte Teams mit unterschiedlicher Zielsetzung erschweren es, ML-Modelle zur Produktionsreife zu führen, und machen das Vorhaben fehleranfällig.

Das Beste aus DevOps für Ihre ML-Modelle

MLOps weitet DevOps auf den ML-Bereich aus und verbessert die Zusammenarbeit und Integration, damit Data Scientists mehr bewirken können.

In Anlehnung an die DevOps-Definition von Ken Mugrage, Principal Technologist bei Thoughtworks, definieren wir MLOps als **eine Kultur, in der Menschen unabhängig von ihrem Titel oder Hintergrund zusammenarbeiten, um Machine-Learning-Systeme zu konzipieren, zu entwickeln, zu implementieren, zu betreiben, zu überwachen und kontinuierlich zu verbessern**.

Die einzelnen Bestandteile dieser Definition im Überblick:

Kultur: MLOps geht über Tools oder Praktiken hinaus: Es handelt sich um eine Kultur, die durch die Art des Zusammenarbeitens sowie gemeinsame Werte geprägt ist. MLOps-Prinzipien setzen einen kulturellen Wandel voraus.

Menschen arbeiten unabhängig von Titel oder Hintergrund zusammen: Um ML-Modelle zur Produktionsreife zu bringen, bedarf es diverser Kompetenzen: Data Science, Data Engineering, ML-Engineering, Software-Engineering, Builds und Releases, Infrastruktur, Operations. Wichtig dabei ist, dass die Beteiligten zusammenarbeiten und keine Silos erzeugen.

Konzipieren: ML-Design ist eine Mischung aus Kreativität und wissenschaftlicher Präzision. Damit das Modell den Vorstellungen entspricht, müssen die Daten – die Basis jedes ML-Modells – frei zugänglich sein. Daten in Silos, proprietären Datenbanken oder im Besitz anderer Abteilungen sind nicht auffindbar und verhindern die Konzeption künftiger Modelle.

Entwickeln, deploy, betreiben: Hier liegt der Kern der ML-Entwicklung. Damit die Ergebnisse reproduzierbar und zuverlässig sind, müssen Code, Modell und Daten stets synchron sein.

Überwachen: Da sich Daten laufend ändern können, sollte das Modell im Produktivbetrieb überwacht werden. Abweichungen und Verzerrungen in den genutzten Daten können die Modellperformance beeinträchtigen – mit potenziell katastrophalen Folgen.

Kontinuierlich verbessern: Die im Produktivmodell verwendeten realen Daten ändern sich ständig. Somit ist MLOps kein temporärer Prozess, sondern ein ewiger Kreislauf. Zur Verbesserung der Modellperformance muss das Modell also angepasst oder neu trainiert werden und den MLOps-Zyklus erneut durchlaufen. Je schneller und reibungsloser dies geschieht, desto rascher kann eine Organisation ML-basierte Prozesse auf die veränderte Realität abstimmen.

> **„DevOps ist eine Kultur, in der Menschen unabhängig von ihrem Titel oder Hintergrund zusammenarbeiten, um ein System zu konzipieren, zu entwickeln, zu implementieren und zu betreiben."**
>
> **—Ken Mugrage, Thoughtworks**

Prozessoptimierung und zuverlässiger Produktivbetrieb mit MLOps

MLOps verbindet den kreativ-wissenschaftlichen Prozess von Data Scientists mit dem modernen Software-Engineering-Prozess, um Software sicher, schnell und nachhaltig in Produktion zu bringen.

- Vermeidung fehleranfälliger, manueller Schritte durch nahezu **vollständige Automatisierung**
- **Qualität ist fester Prozessbestandteil** und hängt nicht mehr nur von menschlichen Tests ab
- Zur Risikominimierung und kontinuierlichen Verbesserung wird häufig und in kleinen Iterationen deployed.
- Durchgehende Versionierung: von der Idee zum **wiederholbaren und überprüfbaren Prozess**

Vorteile von MLOps

- Schnellere und zuverlässigere Implementierung und Verbesserung produktiv genutzter Modelle
- Mehr Einfluss und höhere Produktivität von Data Scientists
- Implementierung neuer Modelle: Geschäftsentscheidung statt technischer Entscheidung

Funktionsweise von Continuous Delivery for ML (CD4ML)

Mit dem von Thoughtworks entwickelten Ansatz Continuous Delivery for Machine Learning (CD4ML)[5] lässt sich MLOps mit den Prinzipien, Praktiken und Tools von Continuous Delivery implementieren. In einem seiner ersten ML-Projekte erstellte Thoughtworks für AutoScout24, den führenden Online-Marktplatz für Fahrzeuge in Europa, mit CD4ML eine Preisempfehlungs-Engine auf AWS[6]. Heute ist CD4ML bei Thoughtworks der Standard für ML-Projekte.

Continuous Delivery sorgt für Automatisierung, Qualität und Disziplin im Prozess für Software Releases und Deployments. Doch ML- Systeme bringen neue Hürden mit sich, die über diese hinausgehen:

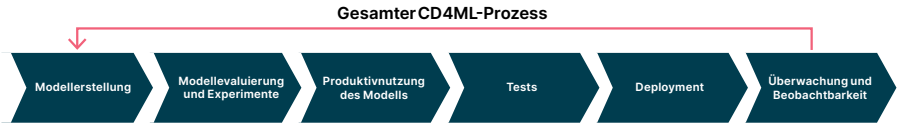
- **Dynamik der Daten:** Wenn Sie ein ML-Modell mit veralteten Daten trainieren, erhalten Sie sub-optimale Ergebnisse.
- **Stetiger Wandel der Modelle:** Data Scientists experimentieren und untersuchen, wie sich die Modellergebnisse verbessern lassen.
- **Überführung in den Produktivbetrieb:** Haben sich die neuen Modelle bei Trainingsdaten bewährt, müssen sie in den Produktivbetrieb überführt werden. Dies erfordert ein gewisses Maß an Governance, um vor ihrem Einsatz mögliche Befangenheiten, Fairness, Datenschutz, ethische Aspekte und andere relevante Qualitätsaspekte zu berücksichtigen.

[5]Danilo Sato, Arif Wider, Christoph Windheuser, Continuous Delivery for Machine Learning
[6]Thoughtworks, Getting Smart: Applying Continuous Delivery to Data Science to Drive Car Sales

- **Automatisches Deployen:** Der Deploymentprozess selbst muss robust sowie automatisiert sein und im Fall etwaiger Probleme einen schnellen Rollback zulassen.
- **Überwachung im Produktivbetrieb:** Wird das Modell dann wirklich eingesetzt, müssen Sie überwachen, ob es sich mit realen Daten bewährt, um Leistungsabweichungen zu vermeiden.

CD4ML adressiert jeden dieser Punkte, da es den Prozess in seiner Gesamtheit berücksichtigt.

CD4ML: bewährter Prozess für MLOps bei Thoughtworks



Modellerstellung: Data Scientists beginnen ihre Forschungen, indem sie die verfügbaren Daten untersuchen, das Problem nachvollziehen und erste Modelle trainieren.

Modellevaluierung und Experimente: Es werden mehrere Modelle trainiert, um mit verschiedenen Ansätzen zu experimentieren. Anschließend werden sie anhand eines Testdatenbestands evaluiert.

Produktivnutzung des Modells: Das ausgewählte Modell wird im Produktivbetrieb eingesetzt. Nicht unbedingt ein eigener Schritt, aber gelegentlich sind Transformationen nötig, damit sich geeignete Technologien produktiv nutzen lassen.

Tests: Auf dem Weg zur Produktionsreife muss das ausgewählte Modell auf verschiedene Aspekte hin getestet werden. Bei identischen Testdaten sollten auch transformierte Modelle die gleichen Ergebnisse erbringen. Die Modelle können zudem nach möglichen Befangenheiten, Fairness oder ethischen Bedenken sowie nach anderen nicht funktionalen Anforderungen wie Sicherheit oder Skalierbarkeit bewertet werden.

Deployment: Bezüglich der Freigabe des neuen Modells gibt es mehrere Strategien. Das Modell kann parallel zu bestehenden Modellen ausgeführt werden oder diese nach und nach ablösen, um Risiken zu mindern und das Vertrauen zu steigern. Falls im Produktivbetrieb Fehler auftreten, ist auch ein Rollback möglich.

Überwachung und Beobachtbarkeit: Zur Erkennung möglicher Abweichungen wird überwacht, wie die Modelle mit realen Produktivdaten performen. Dabei werden neue Daten generiert, die für den nächsten Zyklus genutzt werden können.

CD4ML richtig implementieren

Die Implementierung und vollständige Automatisierung des gesamten CD4ML-Prozesses kann sich schwierig gestalten und erfordert diverse Tools, Technologien und Architekturentscheidungen. Zwar gibt es hierfür etliche Produkte und Lösungen, doch sollte man deren Stärken und Schwächen kennen. Die ideale Technologie behebt 80 Prozent der Problemstellung, gibt Ihrer Organisation in den übrigen 20 Prozent aber ausreichend Anpassungs- und Erweiterungsspielraum.

Bei Ihrer CD4ML-Implementierung sollten Sie folgende technische Aspekte berücksichtigen:

Auffindbare und zugängliche Daten: Ermöglichen Sie Data Scientists, die benötigten Daten zu finden, zu verwenden und bei Bedarf zu erweitern.

Reproduzierbares Modelltraining: Automatisieren Sie und verlassen Sie sich nicht auf implizites Wissen oder auf lokal und manuell erstellte Umgebungen.

Experimente festhalten: Dokumentieren Sie zu Prüf- und Vergleichszwecken die jeweiligen Versuche und ihre Ergebnisse.

Elastische Infrastruktur: Nutzen Sie Cloud- und andere Umgebungen, um die Infrastruktur für ML-Trainings schnell und nach Bedarf bereitzustellen.

Versionskontrolle und Artefakt-Repositories: Prüfen und reproduzieren Sie, wer was wann und weshalb geändert hat.

Tests und Qualität: Bewerten Sie verschiedene Qualitätsaspekte des ML-Systems und automatisieren Sie weitestgehend.

Modell-Serving: Stimmen Sie Hosting und Serving bei produktionsreifen Modellen auf die gewünschten nicht funktionalen Anforderungen ab.

Model-Deployment: Automatisieren Sie Freigabe und Rollback neuer Modelle sowie deren Ausführung (parallel zu/anstelle von bisherigen Modellen).

Überwachung und Beobachtbarkeit: Verfolgen Sie, wie sich die Modelle im Produktivbetrieb verhalten.

Continuous-Delivery-Orchestrierung: Automatisieren Sie den gesamten Prozess vom Code über die Daten bis zum Produktivbetrieb, einschließlich der aus Governance-Gründen erforderlichen manuellen Genehmigung.

Produktionsreife ML-Modelle sind erst der Anfang

Im Rahmen der Feedbackschleife müssen laufend neue Produktivdaten erfasst und überwacht werden. Diese werden dann kuratiert, in neue Trainingsdaten gegliedert und zur Verbesserung künftiger ML-Modelle verwendet. Neben angepassten Modellen erhalten Sie so einen kontinuierlichen Verbesserungsprozess.

CD4ML ist einer der Wege, der Sie bei MLOps ans Ziel bringt. In diesem **technischen Whitepaper** erfahren Sie mehr über die verschiedenen Möglichkeiten, wie Sie Ihr MLOps-Vorhaben erfolgreich umsetzen können.

Erfolgreich mit MLOps?

Über die Autoren

Christoph Windheuser
Global Head of Artificial Intelligence, Thoughtworks

Bevor er zu Thoughtworks wechselte, war Christoph über 20 Jahre lang in verschiedenen Positionen (u. a. bei SAP und Capgemini) tätig. Davor studierte er an der Universität Bonn, an der Carnegie Mellon University in Pittsburgh (USA), an der Waseda-Universität in Tokio (Japan) sowie an der E.N.S.T. in Paris (Frankreich) und promovierte über das Thema Spracherkennung mit neuronalen Netzen.

Danilo Sato
Head of Data and AI Services UK, Thoughtworks

In seiner 20-jährigen Karriere hat Danilo seine Erfahrung als Kunden- und Teammanager mit umfangreichem technischem Fachwissen in den Bereichen Architektur, Engineering, Software, Daten, Infrastruktur und maschinellem Lernen kombiniert. Danilo ist der Verfasser von „DevOps in Practice: Reliable and Automated Software Delivery“, gehört dem Technology Advisory Board von Thoughtworks und dem Office of the CTO an und tritt regelmäßig auf internationalen Konferenzen als Redner auf.