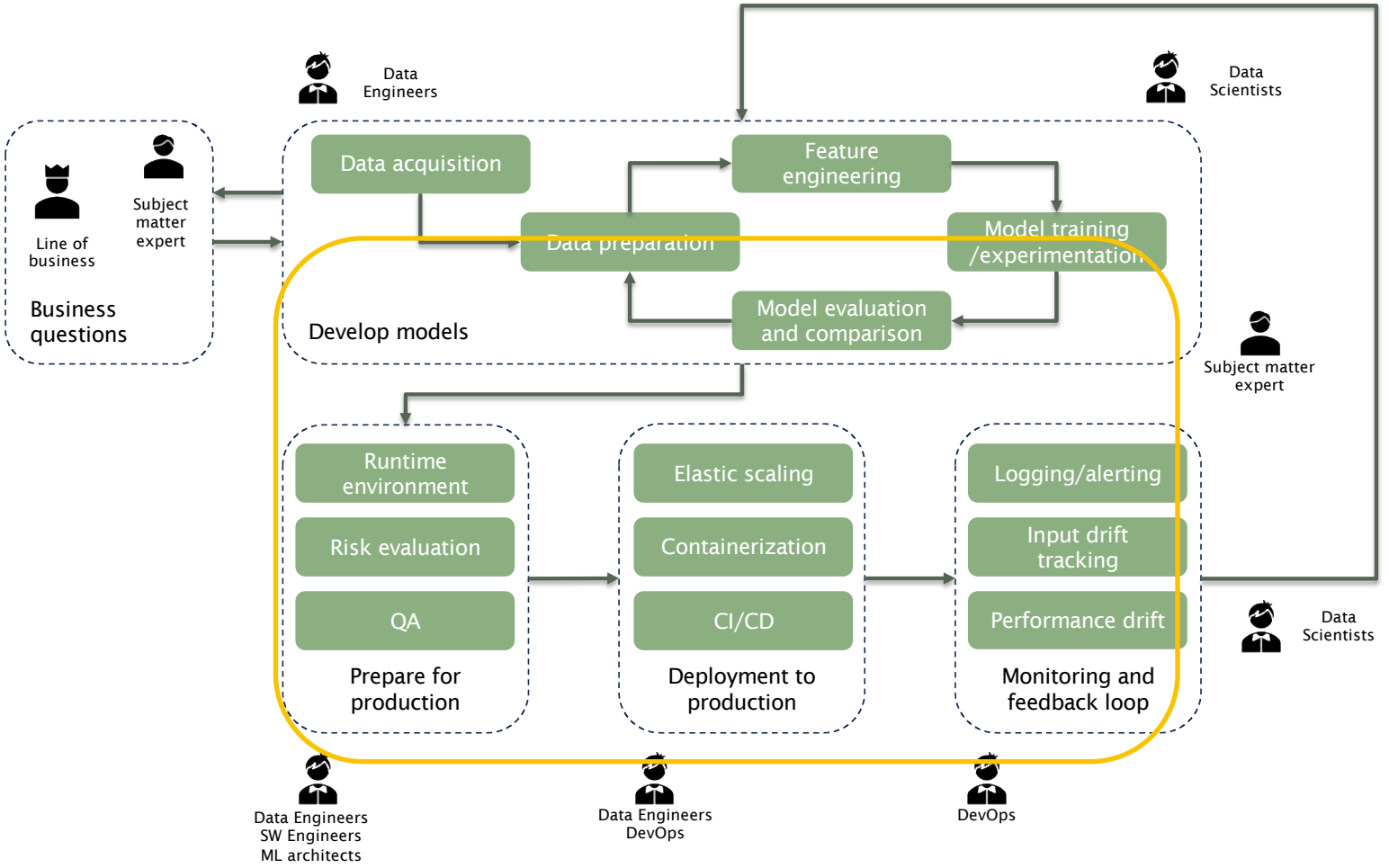
Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

MLOps PML

Untertitel

▸ Violeta Vogel, TI BFH

Das realistische Bild eines ML-Lebenszyklus innerhalb einer durchschnittlichen Organisation



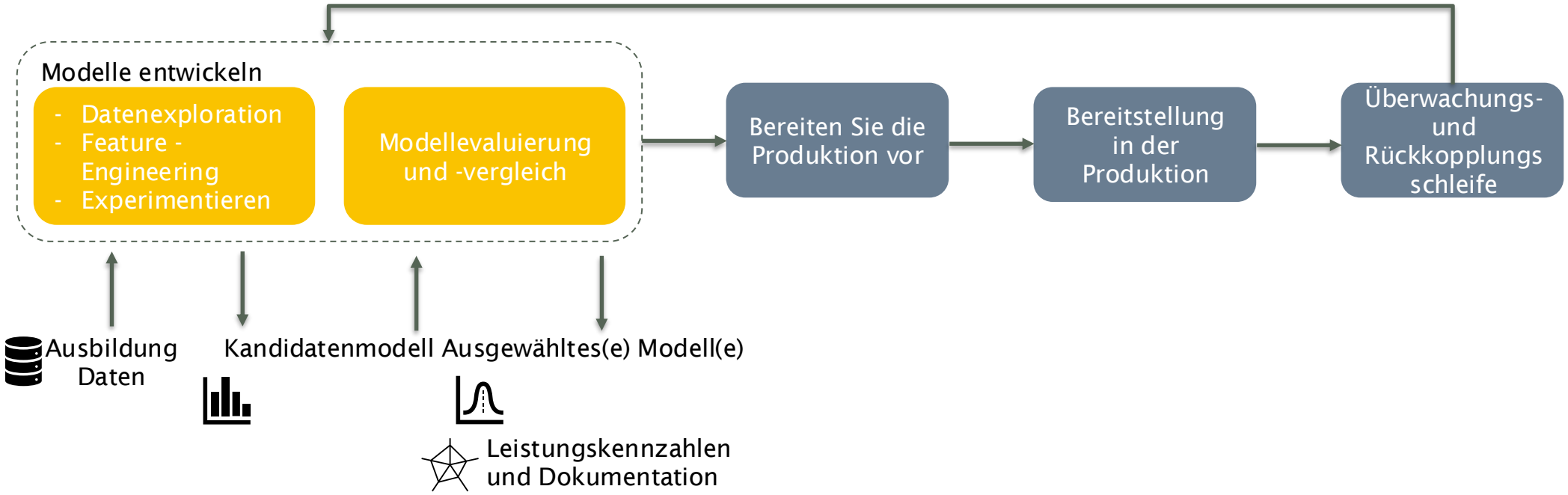
MLOps wie?

Hauptaufgaben

1. Entwicklung der Modelle
2. Vorbereitungen für die Produktion
3. Bereitstellung in der Produktion
4. Überwachungs- und Rückkopplungsschleife
5. Modellsteuerung

MLOps wie?

Modelle entwickeln



Modelle entwickeln

Was ist ein Modell des maschinellen Lernens?

- Ein ML-Modell ist eine Projektion der Realität
- Es handelt sich um eine partielle und annähernde Darstellung eines oder mehrerer Aspekte einer realen Sache oder eines realen Prozesses.
- Ein ML- Modell reduziert sich nach dem Training auf eine mathematische Formel, die bei Eingabe bestimmter Parameter ein Ergebnis liefert.
- ML-Modell ist
  - Die Menge der Parameter, die zum Wiederaufbau und zur Anwendung der Formel erforderlich sind. Sie ist üblicherweise zustandslos und deterministisch.
  - einfach eine berechenbare mathematische Funktion, die den Eingangsdaten

Modelle entwickeln

Erforderliche Komponenten

ML-Komponente	Beschreibung
Trainingsdaten	Der für das Modelltraining verwendete Datensatz: Er ist üblicherweise für den Vorhersagefall mit Beispielen dessen beschriftet, was modelliert wird (überwachtes Lernen).
Eine Leistungskennzahl	Was versucht das Modell zu optimieren?
ML-Algorithmus	Manche Algorithmen eignen sich besser für bestimmte Aufgaben als andere, aber ihre Auswahl hängt auch davon ab, was Priorität haben muss: Leistung, Stabilität, Interpretierbarkeit, Rechenkosten usw.
Hyperparameter	Dies sind Konfigurationen für ML-Algorithmen. Der Algorithmus enthält die Grundformel, die Parameter, die er lernt, sind die Operanden, aus denen diese Formel für die jeweilige Vorhersageaufgabe besteht, und die Hyperparameter sind die Wege, die der Algorithmus beschreiten kann, um diese Parameter zu finden. Fe in einem Entscheidungsbaum ist einer der Hyperparameter die Tiefe des Baums.
Auswertungsdaten satz	Es unterscheidet sich vom Trainingsdatensatz und wird verwendet, um zu bewerten, wie das Modell bei unbekannten Daten abschneidet.

Modelle entwickeln

MLOps- Überlegungen nach Algorithmustyp

Algorithmustyp	Name	MLOps-Überlegungen
Linear	Regression (linear und logistisch)	Es besteht eine Tendenz zur Überanpassung.
Baumbasiert	Entscheidungsbaum	Kann instabil sein – kleine Datenänderungen können zu großen Veränderungen in der Struktur des optimalen Entscheidungsbaums führen.
	Random Forest	Vorhersagen können schwer verständlich sein, was aus Sicht verantwortungsvoller KI eine Herausforderung darstellt. Random-Forest-Modelle liefern zudem oft relativ langsam Vorhersagen, was für Anwendungen problematisch sein kann.
	Gradientenboosting	Vorhersagen können schwer verständlich sein. Eine kleine Änderung der Merkmale oder des Trainingsdatensatzes kann radikale Änderungen im Modell bewirken.
Tiefgründiges Lernen	Neutrale Netzwerke	Diese Modelle sind nahezu unmöglich zu verstehen. Ihr Training ist extrem langsam und sie benötigen viel Rechenleistung und Daten. Lohnt sich der Ressourcenaufwand, oder würde ein einfacheres Modell genauso gut funktionieren?

Modelle entwickeln

Wie sich die Merkmalsauswahl auf die MLOps- Strategie auswirkt

- Weitere Auswirkungen auf die Daten:
  - Genaueres Modell
  - Um Fairness zu gewährleisten, sollte die Aufteilung in präzisere Gruppen erfolgen.
  - Ergänzen Sie einige nützliche fehlende Informationen.
  - Die Berechnung des Modells kann zunehmend teurer werden.
  - Mehr Funktionen erfordern mehr Eingaben und mehr Wartung.
  - Mehr Funktionen bedeuten einen gewissen Stabilitätsverlust.
  - Die schiere Anzahl der Funktionen kann Bedenken hinsichtlich des Datenschutzes aufwerfen.

Modelle entwickeln

Experimentieren

- ***Experimente finden während des gesamten Modellentwicklungsprozesses statt, und in der Regel wird jede wichtige Entscheidung oder Annahme durch mindestens ein Experiment oder vorherige Forschung begründet.***
- Zu den Zielen der Experimente gehören:
  - Beurteilung, wie nützlich oder wie gut ein Modell mit den erforderlichen Komponenten erstellt werden kann.
  - Die besten Modellparameter finden (Algorithmen, Hyperparameter, Merkmalsvorverarbeitung usw.)
  - Optimierung des Bias/Varianz-Kompromisses für gegebene Trainingskosten, um dieser Definition von „bestmöglich“ zu entsprechen
  - Das richtige Gleichgewicht zwischen Modellverbesserung und reduzierten Rechenkosten finden.

„Alle Modelle sind falsch, aber manche sind nützlich.“
George EP Box

Modelle entwickeln

Modelle bewerten und vergleichen

- Es ist wichtig, ein Modell im Kontext zu bewerten und es mit dem Zustand vor der Einführung des Modells vergleichen zu können, um eine Vorstellung davon zu bekommen, wie das Ergebnis aussehen würde, wenn das aktuelle Modell oder der aktuelle Entscheidungsprozess durch das neue ersetzt würde.

Modelle entwickeln

Versionsverwaltung

- Es werden zwei unterschiedliche Bedürfnisse berücksichtigt:
  - Die Möglichkeit, zu verschiedenen „Zweigen“ der Experimente zurückzukehren. Zum Beispiel, um einen früheren Projektstand wiederherzustellen, wenn der Experimentierprozess in einer Sackgasse geendet hat.
  - Die Berechnungen, die zur Modellimplementierung geführt haben, sollen einem Audit-Team mehrere Jahre nach der ersten Durchführung der Experimente erneut zur Verfügung gestellt werden können.

Modelle entwickeln

Modellreproduzierbarkeit

- Die Operationalisierung erfordert die Reproduktion des Modells nicht nur in einer anderen Umgebung, sondern auch von einem anderen Ausgangspunkt aus, sowie das Debuggen usw.
- Deshalb müssen alle Aspekte des Modells dokumentiert und wiederverwendbar sein:
  - *Annahmen* über das Problem, seinen Umfang, die Daten usw. sollten explizit formuliert und protokolliert werden, damit sie anhand neuer Informationen überprüft werden können.
  - *Zufälligkeit* : Es ist notwendig, die Pseudozufälligkeit, wie z. B. Stichproben, zu kontrollieren.
  - *Daten* : Es müssen einige Daten verfügbar sein.
  - *Einstellungen* : Alle durchgeführten Verarbeitungsschritte müssen mit denselben Einstellungen reproduzierbar sein.

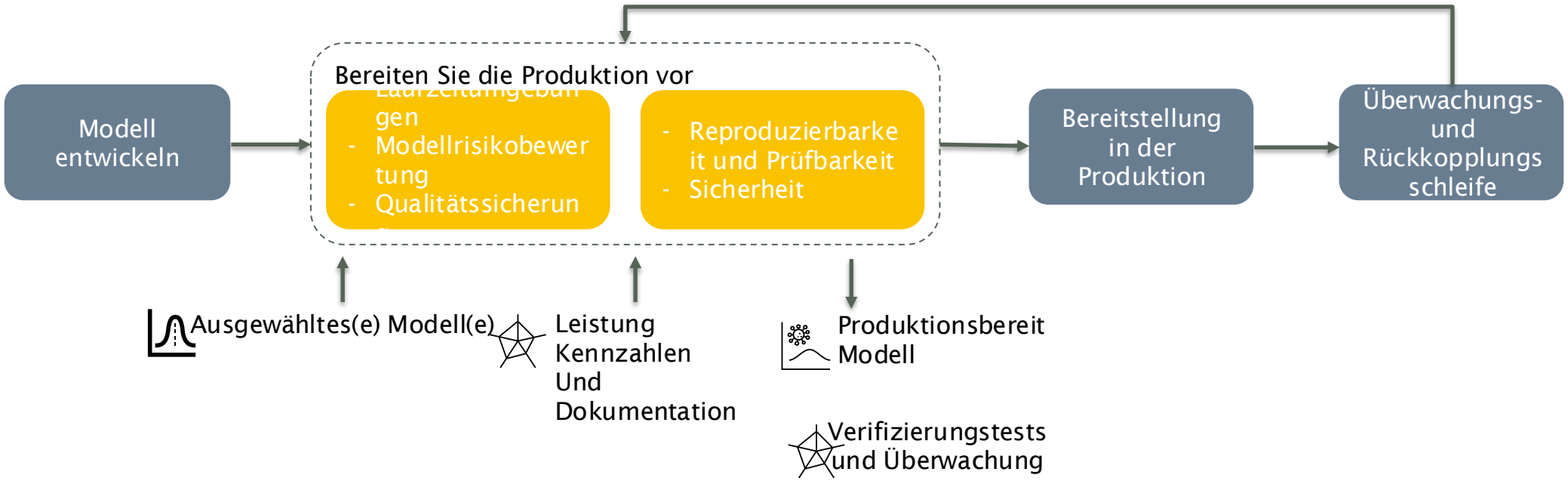
Modelle entwickeln

Modellreproduzierbarkeit

- *Ergebnisse* : Es ist notwendig, Modelle eingehend zu vergleichen und zu analysieren, um Modelle zu erhalten, die die Anforderungen erfüllen.
- *Implementierung* : Schon geringfügig unterschiedliche Implementierungen desselben Modells können zu unterschiedlichen Modellen führen, was ausreicht, um die Vorhersagen bei einigen knappen Entscheidungen zu verändern.
- *Umgebung* : Beispielsweise kann eine leicht abweichende Version eines in einem Schritt verwendeten Python-Pakets die Ergebnisse auf schwer vorhersehbare Weise verändern.

MLOps wie?

Vorbereitungen für die Produktion



Vorbereitungen für die Produktion

Laufzeitumgebung

- Modell-Pipeline: Die Werkzeuge müssen während der Modellentwicklung eingerichtet werden, idealerweise bevor die erste Version des Modells fertiggestellt oder überhaupt begonnen wird.
- Idealerweise würden Modelle, die in der Entwicklungsumgebung laufen, validiert und unverändert in die Produktion übernommen.
  - Dadurch wird der Anpassungsaufwand minimiert und
  - verbessert die Wahrscheinlichkeit, dass sich das Modell in der Produktion so verhält wie in der Entwicklungsphase.
- Die Produktionsumgebung sollte auf eine Datenbank zugreifen können, über die entsprechende Netzwerkverbindung verfügen, die für die Kommunikation mit dem Datenspeicher erforderlichen Bibliotheken oder Treiber installiert haben und die Authentifizierungsdaten in Form einer Produktionskonfiguration gespeichert sein.

Vorbereitungen für die Produktion

Modellrisikobewertung

- Bevor Sie ein Modell in Produktion nehmen, sollten Sie die unbequemen Fragen stellen:
  - Was passiert, wenn sich das Modell auf die denkbar schlechteste Weise verhält?
  - Was passiert, wenn es einem Benutzer gelingt, die Trainingsdaten oder die interne Logik des Modells zu extrahieren?
  - Welche finanziellen, geschäftlichen, rechtlichen, sicherheitsrelevanten und reputationsbezogenen Risiken bestehen?

Vorbereitungen für die Produktion

Modellrisikobewertung

- Wodurch entsteht im Wesentlichen das Risiko von ML-Modellen?

Vorbereitungen für die Produktion

Modellrisikobewertung

- Das Risiko von ML-Modellen entsteht im Wesentlichen durch Folgendes:
  - Fehler, Irrtümer beim Entwurf, Training oder der Evaluierung des Modells (einschließlich Datenaufbereitung)
  - Fehler im Laufzeit-Framework
  - Geringe Qualität der Trainingsdaten
  - Großer Unterschied zwischen Produktions- und Trainingsdaten
  - Missbrauch des Modells oder Fehlinterpretation seiner Ergebnisse
  - Gegnerische Angriffe
  - Rechtliches Risiko
  - Reputationsrisiko aufgrund von Voreingenommenheit und unethischem Einsatz von maschinellem Lernen

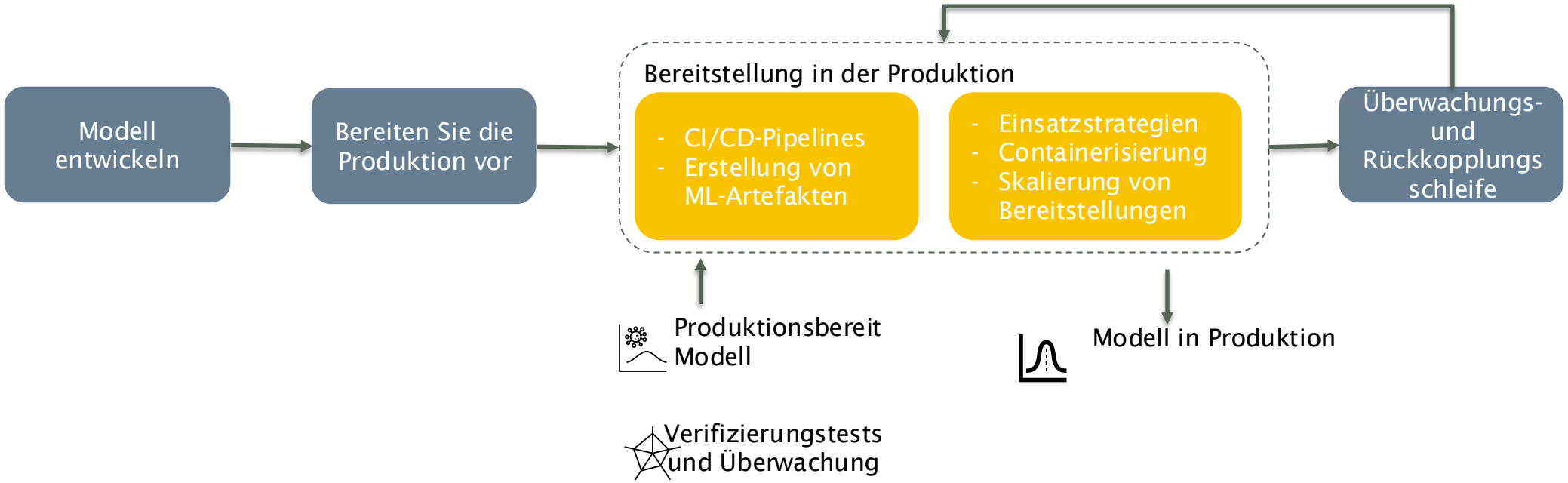
Vorbereitungen für die Produktion

Qualitätssicherung für ML

- Ziel der Qualitätskontrolle ist die Sicherstellung der Reproduzierbarkeit und Prüfbarkeit.
- Die Qualitätsprüfung besteht üblicherweise aus:
  - Vollständige Dokumentation
  - Für das Training und die Prüfung verwendete Daten
  - Ein Artefakt, das die Implementierung des Modells sowie die Spezifikation der Umgebung, in der es ausgeführt wurde, bündelt.
  - Validierung des Modells anhand organisatorischer Richtlinien
  - Genehmigung der Datengovernance
  - Testergebnisse, einschließlich Modellerläuterungen und Fairnessberichten
  - Detailliertes Modellprotokoll und Überwachungsmetadaten

MLOps wie?

Bereitstellung in der Produktion



Bereitstellung in der Produktion

CI/CD-Pipelines (Modell-Pipelines)

- Nach erfolgreicher Entwicklung eines Modells sollte der Dateningenieur oder Datenwissenschaftler den Code, die Metadaten und die Dokumentation in das zentrale Repository hochladen und die CI/CD-Pipeline auslösen.

Bereitstellung in der Produktion

CI/CD-Pipelines (Modell-Pipelines) Beispiel

1. Erstelle ein Modell:
    1. Erstellen Sie die Modellartefakte
    2. Die Artefakte in ein Langzeitlager einlagern.
    3. Führen Sie grundlegende Überprüfungen durch (Rauchtests / Plausibilitätsprüfungen)
    4. Fairness- und Erklärbarkeitsberichte erstellen
2. Bereitstellung in der Testumgebung
    1. Führen Sie Tests durch, um die ML-Leistung und die Rechenleistung zu validieren.
    2. Manuelle Validierung
3. Bereitstellung in der Produktionsumgebung
    1. Das Modell als Canary bereitstellen
    2. Das Modell vollständig implementieren

Bereitstellung in der Produktion

Erstellung von ML-Artefakten

- Sobald sich Code und Daten in einem zentralen Repository befinden, muss ein testbares und bereitstellbares Projektpaket erstellt werden. Diese Pakete werden als Artefakte bezeichnet. Folgende Elemente müssen in ein Artefakt aufgenommen werden:
  - Code für das Modell und seine Vorverarbeitung
  - Hyperparameter und Konfiguration
  - Trainings- und Validierungsdaten
  - Trainiertes Modell in seiner ausführbaren Form
  - Eine Umgebung einschließlich Bibliotheken
  - Dokumentation
  - Kader und Daten für Testszenarien

Bereitstellung in der Produktion

Einsatzstrategien

- *Integration* : der Prozess des Zusammenführens eines Beitrags mit einem zentralen Repository (typischerweise das Zusammenführen eines Git-Feature-Branches mit dem Haupt-Branch)
- *Lieferung* : der Prozess der Erstellung einer vollständig verpackten und validierten Version des Modells, die bereit für den Einsatz in der Produktion ist.
- *Bereitstellung* : der Prozess der Ausführung einer neuen Modellversion auf einer Zielinfrastruktur
- *Veröffentlichung* : Der Prozess der Umleitung der Produktionslast auf die neue Version. (Mehrere Versionen eines Modells können gleichzeitig in der Produktion ausgeführt werden, aber nur eine wird veröffentlicht.)

Bereitstellung in der Produktion

Instandhaltung in der Produktion

- *Ressourcenüberwachung* : Das Erfassen von IT-Kennzahlen wie CPU-, Speicher-, Festplatten- oder Netzwerknutzung kann hilfreich sein, um Probleme zu erkennen und zu beheben.
- *Statusprüfung* : Prüfen Sie, ob das Modell online ist und analysieren Sie seine Latenz. Typischerweise jede Minute.
- Überwachung von ML-Metriken: Analyse der Modellgenauigkeit und Vergleich mit einer anderen Version sowie Erkennung von Veralterung. Typischerweise einmal wöchentlich oder einmal monatlich.

Bereitstellung in der Produktion

Instandhaltung in der Produktion

- *Ressourcenüberwachung* : Das Erfassen von IT-Kennzahlen wie CPU-, Speicher-, Festplatten- oder Netzwerknutzung kann hilfreich sein, um Probleme zu erkennen und zu beheben.
- *Statusprüfung* : Prüfen Sie, ob das Modell online ist und analysieren Sie seine Latenz. Typischerweise jede Minute.
- *Überwachung von ML-Metriken* : Analyse der Modellgenauigkeit und Vergleich mit einer anderen Version sowie Erkennung von Veralterung. Typischerweise einmal wöchentlich oder einmal monatlich.

Bereitstellung in der Produktion

Skalierung von Bereitstellungen

- Welche Herausforderungen ergeben sich für Organisationen, die skalieren müssen?
- Wie würdest du sie lösen?

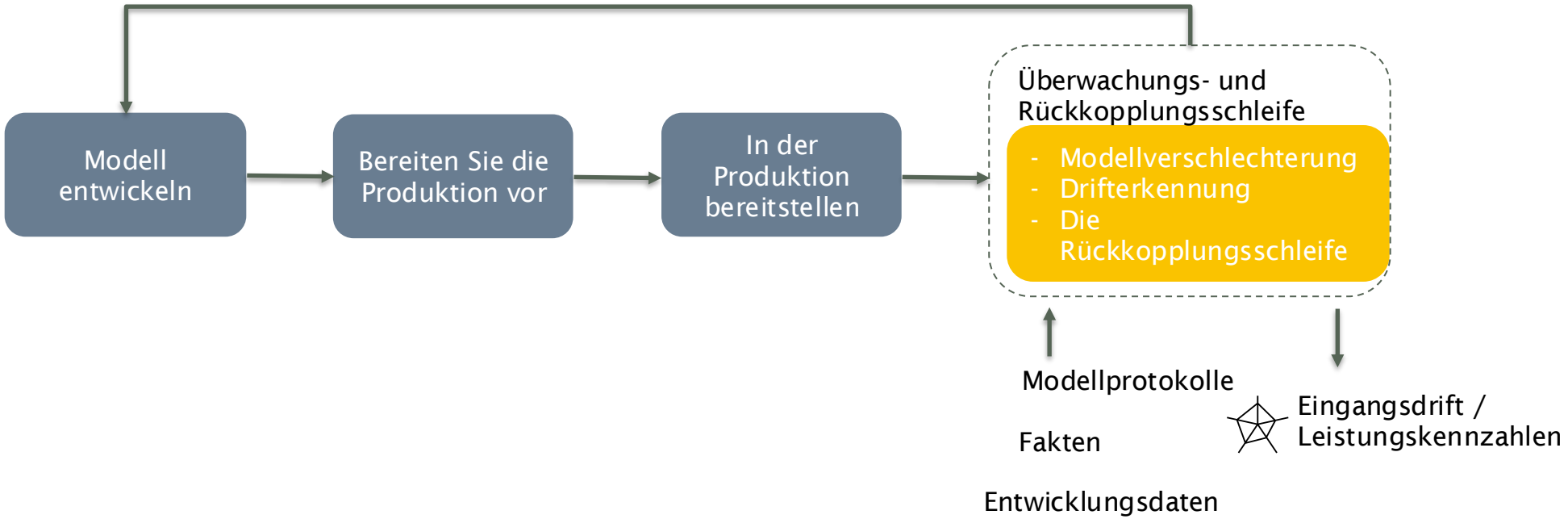
Bereitstellung in der Produktion

Skalierung von Bereitstellungen

- Organisationen stehen vor zwei Herausforderungen:
  - Die Fähigkeit, ein Modell in der Produktion mit umfangreichen Daten zu verwenden
  - Die Fähigkeit, immer größere Anzahlen von Modellen zu trainieren.
- Grundsätzlich gibt es zwei Arten von Strategien zur Verteilung der Berechnungen:
  - Erhöhen Sie die Rechenleistung und stellen Sie sicher, dass die Überwachungsinfrastruktur die Arbeitslast bewältigen kann.
  - Partitionieren Sie die Daten

MLOps wie?

Überwachungs- und Rückkopplungsschleife



Überwachungs- und Rückkopplungsschleife

Überwachung der ML-Modelle

- ML-Modelle müssen auf zwei Ebenen überwacht werden:
  - Ressourcenebene. Zu den Schlüsselfragen gehören:
    - Ist das System am Leben?
    - Entsprechen CPU-, RAM-, Netzwerk- und Festplattenauslastung den Erwartungen?
    - Werden die Anfragen im erwarteten Tempo bearbeitet?
  - Auf der Leistungsebene. Zu den Kernfragen gehören:
    - Stellt das Modell das Muster der neu eingehenden Daten noch zutreffend dar?
    - Funktioniert es genauso gut wie in der Entwurfsphase?

Überwachungs- und Rückkopplungsschleife

Wie oft sollten Modelle neu trainiert werden?

Überwachungs- und Rückkopplungsschleife

Wie oft sollten Modelle neu trainiert werden?

- Es hängt davon ab:
  - *Der Bereich* : In der Cybersicherheit und Echtzeitverarbeitung müssen Systeme regelmäßig aktualisiert werden. Andere Systeme wie die Spracherkennung sind stabiler.
  - *Die Kosten* : Organisationen müssen abwägen, ob sich die Umschulung lohnt.
  - *Die Leistungsfähigkeit des Modells* : In manchen Situationen wird die Leistungsfähigkeit des Modells durch die begrenzte Anzahl an Trainingsbeispielen eingeschränkt, und daher hängt die Entscheidung für ein erneutes Training von der Sammlung ausreichend neuer Daten ab.

Überwachungs- und Rückkopplungsschleife

Modellverschlechterung

- Sobald das Modell in Produktion ist, gibt es zwei Ansätze, um seine Leistungsverschlechterung zu überwachen:
  - Bewertung der Realität
  - Eingangsdrifterkennung

Überwachungs- und Rückkopplungsschleife

Modellverschlechterung

- Bewertung der tatsächlichen Gegebenheiten:
  - Warten Sie auf das Label-Ereignis. Zum Beispiel Betrugserkennung.
  - Berechnen Sie die Leistung des Modells auf Basis der tatsächlichen Werte.
  - Vergleichen Sie es mit den in der Trainingsphase erfassten Metriken.
  - Wenn die Differenz den Schwellenwert überschreitet, muss das Modell neu trainiert werden.
- Überwachungsmetriken:
  - Statistische Größen wie Genauigkeit, Verlust usw.
  - Unternehmensbezogene Kennzahlen wie die Kosten-Nutzen-Analyse

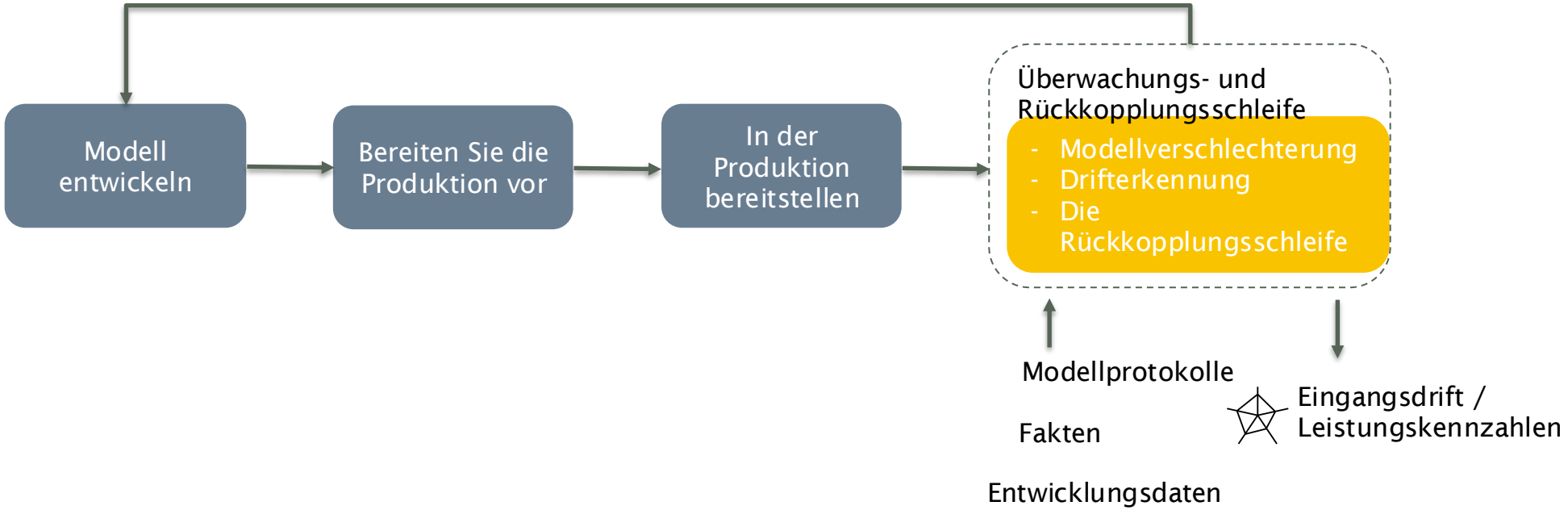
Überwachungs- und Rückkopplungsschleife

Modellverschlechterung

- Eingangsdrifterkennung
  - Die Logik dahinter ist, dass eine Abweichung der Datenverteilung (Mittelwert, Standardabweichung, Korrelationen zwischen Merkmalen usw.) zwischen der Trainings- und Testphase einerseits und der Entwicklungsphase andererseits ein starkes Indiz dafür ist, dass die Leistung des Modells nicht gleich sein wird.
- Hauptursachen der Datenabweichung:
  - Verzerrung durch Stichprobenauswahl, bei der die Trainingsstichprobe nicht repräsentativ für die Grundgesamtheit ist
  - Nicht-stationäre Umgebung, in der die aus der Quellpopulation erhobenen Trainingsdaten nicht die Zielpopulation repräsentieren, z. B. aufgrund von Saisonalität.
- Erkennung: Mit statistischen Tests

MLOps wie?

Überwachungs- und Rückkopplungsschleife



Überwachungs- und Rückkopplungsschleife

Rückkopplungsschleife

- Die im Überwachungs- und Feedback-Kreislauf gesammelten Daten werden an die Modellentwicklungsphase weitergeleitet.
- Das System analysiert, ob das Modell wie erwartet funktioniert.

Wokshop

- Literatur: https://yalebooks.yale.edu/book/9780300158564/the-illusions-of-entrepreneurship/