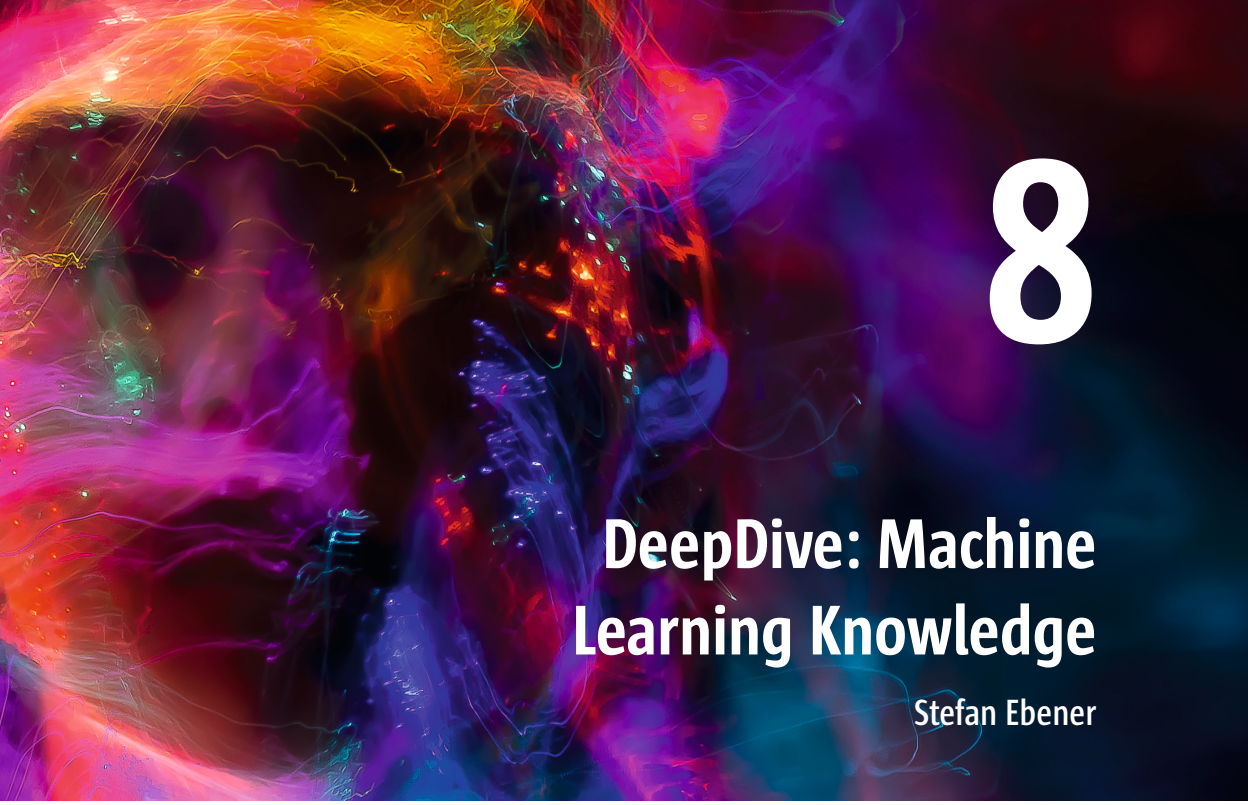

8 DeepDive: Machine Learning Knowledge

Stefan Ebener

8.1 Anbruch einer neuen Dekade: Die Dekade des maschinellen Lernens

Die ganze Welt redet über Künstliche Intelligenz (KI) und wie Daten und Algorithmen unser tägliches Leben verändern werden. Ohne jeden Zweifel können wir feststellen, dass eine Verschiebung beinahe in allen Branchen stattfindet. Sei sie noch so analog. Sei es Fertigung, der Finanzsektor, der Einzelhandel, Versicherungen oder das Gesundheitswesen. Nicht wenige sind davon überzeugt, dass insbesondere das Gesundheitswesen am stärksten davon profitieren kann. Hochpersonalisierte Behandlungen auf Basis von DNA Sequencing, das maschinelle Markieren von Pathologie-Objektträgern, Prävention durch KI-gestützte Apps, computergesteuerte Nanopartikel oder auch nur die Messung der Vitalwerte per Smartphone-Kamera sind nur einige Beispiele.

Interessanterweise denken die meisten Menschen, dies sei vor allem eine Frage der richtigen Technologie. Tatsächlich jedoch ist das wesentliche Unterscheidungsmerkmal zwischen den erfolgreichen und weniger erfolgreichen Institutionen, Krankenhäusern oder Forschungseinrichtungen nicht allein die Technologie, es ist vor allem das Mindset. Das Mindset nicht einiger weniger Vordenker, sondern das Mindset der Belegschaft hinsichtlich der Wichtigkeit, dem offenen Zugang und der Behandlung von Daten. Künstliche Intelligenz hat hier das Potenzial, grundlegende transformative Prozesse in Gang zu setzen. Dabei sollte man Künstliche Intelligenz mehr als die Vision, Machine Learning als Kerntechnologie und Deep Learning als spezielle Implementierungsart verstehen.

Auszug aus: Matusiewicz, Werner (Hrsg.) Future Skills in Medizin und Gesundheit.
ISBN: 9783954665945. Urheberrechtlich geschützt.
© MWV Medizinisch Wissenschaftliche Verlagsgesellschaft 2021

Machine Learning tritt an, um die großen Probleme in der Gesundheitsbranche zu lösen und fußt dabei auf drei grundlegenden Säulen:

- **der stetig steigenden Fülle an Signalen und Daten,**
- **der Konnektivität und den Beziehungen zwischen Menschen, Geräten und Daten sowie**
- **dem Einsatz neuer Technologien für eine radikale Erneuerung in kürzeren Abständen.**

Eine Transformation des Gesundheitswesens und die Erreichung echter Meilensteine bspw. in der Erkennung von Krankheiten, der Vorsorge oder auch der Behandlung, ergibt sich nicht länger aus dem Erfahrungsaufbau eines Gebietes, sondern aus der Kombination der drei Säulen. Dies bedingt eine ganze Reihe an Skills, die Institutionen zukünftig aufbauen müssen samt klarem Fokus auf interdisziplinäre und agile Teams.

8.2 Die Entwicklung eines ML-Modells ist weit mehr als nur der Code

Die Entwicklung, das Bereitstellen und Betreiben von auf Machine Learning basierenden Applikationen in der IT bedarf weit mehr als nur das Schreiben des ML-Codes. Es umfasst eine Vielzahl an Phasen, Prozessen und Rollen in einem Unternehmen. Zu berücksichtigen ist zudem, dass viele Modelle in Produktion einer stetigen Veränderung – getrieben durch sich verändernde Daten – unterliegen. Die Folge sind ein kontinuierlicher Verbesserungsprozess samt erneutem Training der Modelle, Benchmarking der Modellversionen, eine Etablierung von Teststrategien etc. Für eingehende Daten müssen eine Datenvorverarbeitung und -validierung eingeführt werden um Qualitätsprobleme auszuschließen. Gefolgt von Feature-Engineering, Hyperparameter-Tuning, Monitoring usw (s. Abb. 1).

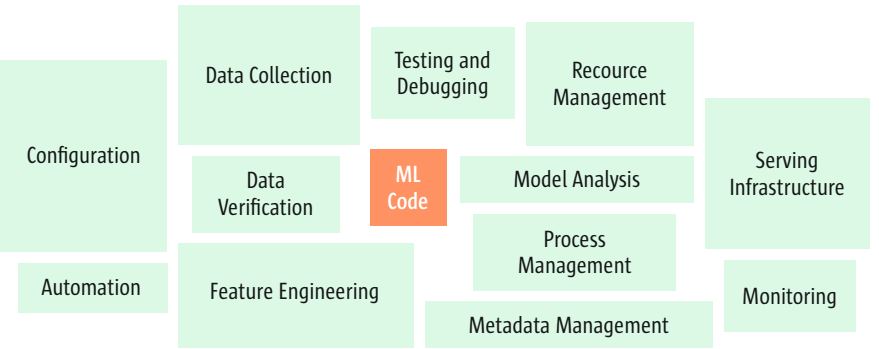

Abb. 1 Hidden Technical Debt in Machine Learning Systems (basierend auf Sculley et al. 2015)

Auszug aus: Matusiewicz, Werner (Hrsg.) Future Skills in Medizin und Gesundheit.
ISBN: 9783954665945. Urheberrechtlich geschützt.
© MWV Medizinisch Wissenschaftliche Verlagsgesellschaft 2021

Überträgt man all diese Anforderungen und Verantwortlichkeiten auf Rollen in einem Unternehmen, entstehen zugleich eine Vielzahl an Positionen mit unterschiedlichen Skills, die es zu besetzen gilt. Zu nennen sind u.A.:

- **Data Scientist:** sammelt, analysiert und interpretiert zumeist große Volumen an Daten
- **Data Analyst:** sammelt, verarbeitet und führt statistische Datenanalysen durch
- **Data Engineer:** entwickelt, konstruiert, testet und pflegt Architekturen
- **Data Architect:** erstellt Blaupausen für Datenmanagementsysteme zur Integration, Zentralisierung und Pflege von Datenquellen
- **Business Analyst:** verbessert Business-Prozesse als Vermittler zwischen IT und Fachabteilung

Nicht alle Positionen werden tatsächlich auch in einem Unternehmen benötigt oder müssen immer von einer einzelnen Person besetzt werden. Es zeigt jedoch eindrucksvoll, welche Bandbreite an Skills heute notwendig erscheinen.

Das Verständnis über diese Rollen einschließlich der Verantwortlichkeiten unterlag in den letzten Jahren einem Paradigmenwechsel – und hier lassen sich Parallelen zur modernen Softwareentwicklung ziehen. War man zu Beginn noch der Auffassung, dass viele dieser Rollen für den Aufbau der ML-Pipelines samt Modell sowohl in der Trainingsphase als auch im späteren Betrieb notwendig seien, geht man heute den Weg des kontinuierlichen Verbesserungsprozesses – Machine Learning Operations (kurz: MLOps) genannt (s. Abb. 2).

Wie bereits angemerkt, besteht die große Herausforderung nicht darin, ein ML-Modell zu entwickeln, sondern ein integriertes ML-System zu erstellen und es kontinuierlich in der Produktion zu betreiben. Daher kann MLOps als die Zusammenarbeit von Data Scientists und IT-Betriebsexperten zum Management des gesamten ML-Lebenszyklus in der Produktion verstanden werden. Es legt die Idee zugrunde, die Prinzipien von DevOps auf Machine Learning anzuwenden. Technisch beschrieben, umfasst MLOps Techniken zur Implementierung und Automatisierung von „Continuous Integration“ (CI), „Continuous Delivery“ (CD) sowie „Continuous Training“ (CT) für auf Machine Learning basierende Systeme.

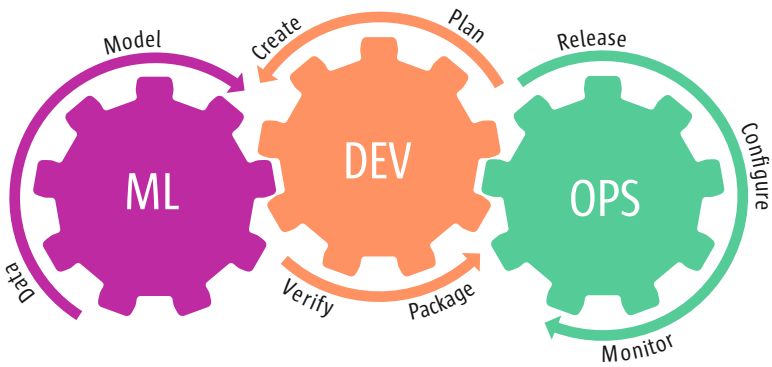

Abb. 2 MLOps (basierend auf NEAL ANALYTICS 2020)

Auszug aus: Matusiewicz, Werner (Hrsg.) Future Skills in Medizin und Gesundheit.
ISBN: 9783954665945. Urheberrechtlich geschützt.
© MWV Medizinisch Wissenschaftliche Verlagsgesellschaft 2021

In der Praxis bedeutet dies konkret, dass eine kontinuierliche Entwicklung und Integration für Änderungen an Modell- oder Pipelinekomponenten, Metadatenspeicher, Feature-Store, Deployment-, Quellcodeverwaltungs-, Test- und Buildservices erbracht werden müssen. Hinzu kommt ein umfassendes Monitoring samt Performanceüberwachung usw.

Was zunächst recht aufwendig und komplex klingt – und mitnichten für den initialen Aufbau unterschätzt werden sollte – bringt im Betrieb unschätzbare Vorteile mit sich. Wiederverwendbare Komponenten und gesamte Pipelines, eine Pipeline-Orchestrierung mit einer von der Code-Laufzeit entkoppelten Ausführungsumgebung, einen Metadaten-Speicher zur Rückverfolgbarkeit und Reproduzierbarkeit oder etwa einen Feature-Store. Das kontinuierliche Training kann auf Basis eines Triggers – einer Nachfrage, eines Zeitplans, der Verfügbarkeit neuer Trainingsdaten, der Verschlechterung der Modell-Performance oder etwa des Konzeptdrifts – angestoßen werden. Insgesamt spart dies nicht nur Kosten, sondern professionalisiert vor allem den gesamten ML-Lebenszyklus und kommt u.A. der Dokumentationspflicht genannter Branchen nach.

Basis von MLOps ist bei dem überwiegenden Teil der Unternehmen inzwischen „Kubeflow“ geworden. Das Open Source Kubeflow-Projekt hat zum Ziel, die Bereitstellung von ML-Workflows auf Kubernetes einfach, portabel und skalierbar zu gestalten. Von besonderer Bedeutung ist dabei „Kubeflow Pipelines“, ein Kubeflow-Dienst, mit dem sich ML-Systeme erstellen, orchestrieren und automatisieren lassen. Jede Komponente des Systems kann auf Kubeflow, Google Cloud oder anderen Cloud-Plattformen ausgeführt werden. Dies hat einen erheblichen Einfluss auf die genannten Rollen und Verantwortlichkeiten. Einige sind obsolet, andere kommen neu hinzu.

8.3 Schnellere Entwicklung dank neuer Ansätze im Machine Learning

Konterkariert werden viele der alten aber auch der neuen Rollen durch immer umfangreichere Werkzeuge, ausgefeiltere Services, respektive wiederverwendbaren ML-Komponenten. Im Laufe der Zeit hat dies zu drei Arten geführt, wie auf Machine Learning basierte Applikationen und Services entwickelt bzw. konsumiert werden können (s. Tab. 1).

Exemplarisch heruntergebrochen auf das Gesundheitssystem – im speziellen die Medizin – ergeben sich einige interessante Anwendungsbeispiele.

Genannt sei beispielsweise die automatisierte Transkription von Arztberichten. Generell ist Speech-to-Text (STT) ein Feld, das inzwischen durch Millionen von Menschen tagtäglich eingesetzt wird – sei es durch Google Assistant oder Siri. Übertragen auf die Transkription von Arztberichten scheitern diese Systeme aber fundamental, da sie die spezifische Terminologie, Abkürzungen oder bspw. eine Medikation nicht verstehen. Mit dem Einsatz von Transfer Learning, in diesem Fall STT-Modellen, können vortrainierte Modelle mit dem spezifischen Vokabular von Arztberichten erweitert werden.

Auszug aus: Matusiewicz, Werner (Hrsg.) Future Skills in Medizin und Gesundheit.
ISBN: 9783954665945. Urheberrechtlich geschützt.
© MWV Medizinisch Wissenschaftliche Verlagsgesellschaft 2021

Tab. 1 State-of-the-Art für eine Machine Learning Modellentwicklung

Ansatz	Implementierung	Implementierung
Black-Box-ML-Services	konsumierbar per API	■ Daten und Modell sind im Besitz des Anbieters ■ keine Anpassmöglichkeiten der Modelle
Vortrainierte ML-Modelle	konsumierbar per API oder als offline-Modell	■ Grundmodell kommt vom Anbieter und wird durch eigene Daten erweitert (Transfer Learning) ■ individuelle Modelle auf Basis der eigenen Daten ■ vergleichsweise schnell umsetzbar
Eigens entwickelte Modelle	universell konsumierbar je nach Einsatzzweck	■ Daten und Modell bleiben beim Eigentümer ■ aufwendigste Variante, die große Datenmengen benötigt

Ein weiteres Beispiel sind Modelle, die für die Diagnostik zur Erkennung von Lungenkrebs oder Schlaganfällen auf der Basis von CT-Scans, der Beurteilung des Risikos eines plötzlichen Herztodes oder anderer Herzerkrankungen anhand von Elektrokardiogrammen und Herz-MRT-Aufnahmen, die Klassifizierung von Hautläsionen in Hautbildern oder auch dem Finden von Indikatoren für diabetische Retinopathie in Augenbildern entwickelt wurden. Dies sind zumeist eigens zu diesem Zweck entwickelte Modelle. ML entfaltet sein Potenzial immer dann, wenn eine Kombination mehrerer Datenquellen bspw. CT, MRT, Genomik oder Proteomik gleichzeitig herangezogen werden kann, um Muster oder Auffälligkeiten im Krankheitsverlauf zu identifizieren.

Aus dem Bereich der Black-Box-Modelle existieren eine ganze Reihe an Services zu denen u.A. die automatischen Annotationen von Bildern gehören oder auch sehr komplexe Services, wie eine „Healthcare API“, die die Nutzung intelligenter Analyse- und ML-Funktionen in der Cloud mit vordefinierten Connectoren ermöglicht.

8.4 Die Zukunft ist da, sie ist einfach nur nicht gleichmäßig verteilt

Die Zukunft des Gesundheitswesens samt Durchbrüchen in Diagnose, Behandlung oder Vorsorge werden zu einem großen Teil auf dem Einsatz von Software bzw. Machine Learning basieren. Dies kann nur funktionieren, indem sich das Gesundheitssystem dem Patienten durch hochgradig personalisierte Services zuwendet. Dies wiederum lässt nur mit dem Einsatz moderner Technologien, datengetriebenen Ansätzen und neuen Verfahrensweisen skalieren, zu denen Machine Learning gezählt werden darf.

Akteure, die in einem Umfeld von exponentiellen Wandel agieren wollen, benötigen ein völlig neues Skillset. Je nach Level auf dem agiert wird, beinhaltet dies ein umfassendes Domänenwissen gepaart mit einem guten Technologieverständnis und Hintergrundwissen für den Aufbau datengetriebener Services.

Auszug aus: Matusiewicz, Werner (Hrsg.) Future Skills in Medizin und Gesundheit.
ISBN: 9783954665945. Urheberrechtlich geschützt.
© MWV Medizinisch Wissenschaftliche Verlagsgesellschaft 2021

Literatur

NEAL ANALYTICS (2020) Enable MLOps with a proven methodology. URL: https://nealanalytics.com/expertise/mlops (abgerufen am 04.12.2020)

Sculley et al. (2015) Hidden Technical Debt in Machine Learning Systems. URL: https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf (abgerufen am 04.12.2020)

Dr. Stefan Ebener

Stefan Ebener leitet für Google Cloud ein internationales Expertenteam. Seine Leidenschaft gilt den datengetriebenen Zukunftstechnologien und der Weiterentwicklung von Technologiekompetenzen in Unternehmen und Gesellschaft. Zudem ist er freiberuflicher Dozent der Wirtschaftsinformatik, Mitglied der d*Health Academy sowie Startup-Mentor, Autor und Keynote Speaker.

Auszug aus: Matusiewicz, Werner (Hrsg.) Future Skills in Medizin und Gesundheit.
ISBN: 9783954665945. Urheberrechtlich geschützt.
© MWV Medizinisch Wissenschaftliche Verlagsgesellschaft 2021