WHITEPAPER

Rechenzentrum
Künstliche Intelligenz

# In 5 Schritten zum KI-Proof-of-Concept

## Ein Erfolgskonzept in fünf Schritten mit Proof of Concept (PoC) für Bilderkennung (Image Recognition), natürliche Sprachverarbeitung (Natural Language Processing, NLP) und vorausschauende Instandhaltung (Predictive Maintenance)

### Inhaltsverzeichnis



### Einleitung – die Rolle des PoC bei der KI

Ein Softwareprogramm mit künstlicher Intelligenz (KI) ist eines, das wahrnehmen, denken, handeln und sich anpassen kann. Hierfür „lernt" es zunächst von einem umfangreichen und vielseitigen Datenbestand, den es nutzt, um Modelle über die Daten zu trainieren. Sobald das Modell trainiert wurde, wird es eingesetzt, um Ergebnisse aus ähnlichen neuen oder bisher unbekannten Daten abzuleiten. Dazu gehört zum Beispiel das Umwandeln von gesprochener Sprache in Text, das Identifizieren von Anomalien in einer Reihe von Bildern oder das Berechnen des Zeitpunkts, wann ein Maschinenteil kaputtgehen wird. Dieser Ablauf wird in Abbildung 1 dargestellt.

KI-Algorithmen gibt es zwar schon seit vielen Jahren, aber erst in jüngerer Vergangenheit breiten sich KI-basierte Fähigkeiten rasant über alle Unternehmensbereiche aus. Das ist auf verschiedene Faktoren zurückzuführen. Zum einen liegt es daran, dass die Kosten für die Verarbeitung und Speicherung von Daten gleichermaßen schnell gesunken sind. Parallel dazu haben Informatiker das Algorithmen-Design für KI einschließlich neuronaler Netzwerke weiterentwickelt, was zu einer größeren Genauigkeit der Trainingsmodelle führte.

Da sich KI immer mehr durchsetzt, hat sie auch einen Innovationsschub bei der Infrastruktur ausgelöst. Intel hat KI-bezogene Funktionen direkt in seine Hardware eingebaut: Die neuesten skalierbaren Intel® Xeon® Prozessoren bieten skalierbare Performance für die unterschiedlichsten KI-Workloads sowie bahnbrechende Performance bei Deep-Learning-Modelltrainings und -Inferenzen. Der Intel® Nervana™ Prozessor enthält eine neue Architektur für neuronale Netze, die von Grund auf neu entwickelt wurde.

Solche Fortschritte beschleunigen die Verbreitung von KI noch weiter. Sie schaffen enorme Möglichkeiten für Unternehmen, die klügere Entscheidungen treffen und intelligentere Abläufe schaffen wollen - und bieten dadurch einen konkreten wirtschaftlichen Nutzen. Eine von Accenture im Jahr 2017 in verschiedenen Branchen und Ländern durchgeführte Umfrage ergab, dass künstliche Intelligenz die Rentabilität um 38 Prozent steigern kann. Das wird der Wirtschaft in den kommenden Jahrzehnten mehr als 14 Billionen US-Dollar einbringen.[1]

[1] https://www.accenture.com/t20171005T065828Z__w__/us-en/_acnmedia/Accenture/next-gen-5/insight-ai-industry-growth/pdf/Accenture-AI-Industry-Growth-Full-Report.pdf#la=en?la=en

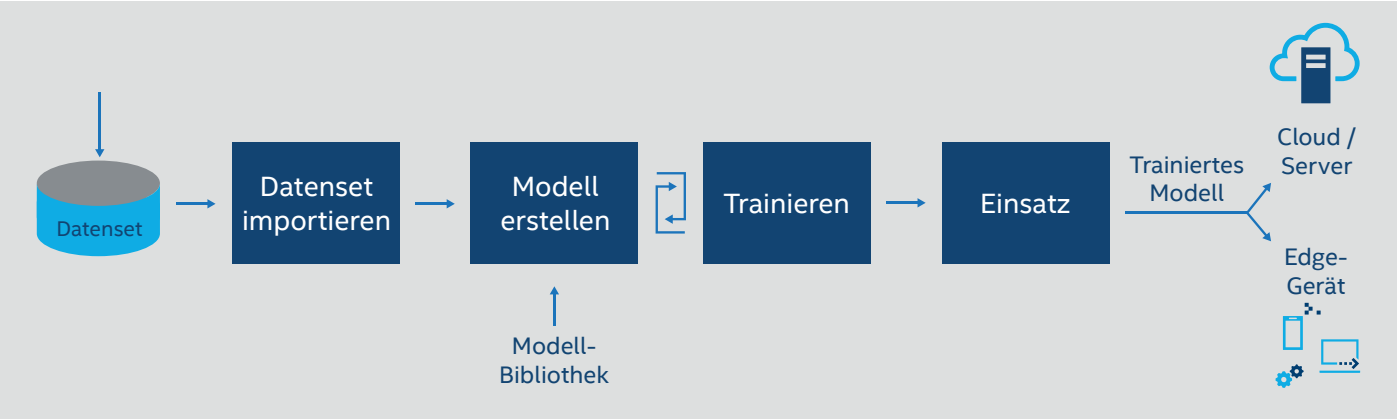


**Abbildung 1: KI-Systeme lernen und leiten dann von Daten Ergebnisse ab**

Trotz dieses eindeutigen Potenzials stehen viele Unternehmen noch vor dem Einstieg in das Thema KI. Deren Einführung geschieht nicht unbedingt so schnell wie Medien und Wissenschaft gerne glauben machen würden.[2] Wenn Unternehmen mit der Einführung von KI beginnen, finden sie die gängigsten Anwendungsfälle in den Bereichen natürliche Sprachverarbeitung (Natural Language Processing; NLP), maschinelles Sehen (Computer Vision) und vorausschauende Instandhaltung (Predictive Maintenance). Kundenbeispiele finden Sie in der Tabelle auf der folgenden Seite.

Zu den Anwendungsfällen von Deep Learning und maschinellem Lernen gehören:

- **Branchenbezogen**: Viele Unternehmen versuchen, branchentypische Probleme zu lösen. Beispiele hierfür sind das Ersatzteilmanagement in der Fertigung, die Lagerbestandsverwaltung im Einzelhandel und die Behandlungsergebnisse von Patienten im Gesundheitswesen.
- **Geschäftsbereiche**: Branchenübergreifend haben alle Unternehmen ähnliche Anforderungen, die jeweils von den einzelnen Geschäftsbereichen abhängen. Natürliche Sprachverarbeitung kommt beispielsweise in Kundendienstabteilungen zum Einsatz. Bilderkennung sowie vorausschauende Instandhaltung sind für Supply-Chain-Anwendungen von Relevanz.
- **Technologiearchitektur**: Viele Beispiele für KI, die uns begegnen, haben eine ähnliche Architektur, selbst wenn sie verschiedene Datenbestände nutzen und unterschiedliche Resultate liefern. So können zum Beispiel Bildverarbeitung und Anomalie-Erkennung, die von einem Kunden genutzt werden, um Defekte bei Solarmodulen zu erkennen, auf einer ähnlichen Plattform basieren wie das System, das Naturschützer verwenden, um Verhaltensänderungen von Fledermäusen zu „belauschen".
- **IT-bezogen**: Einige Anwendungsmöglichkeiten der KI lassen sich anwendungs- und serviceübergreifend nutzen, da sie sich mit der Verwaltung von Datenströmen, dem Vorbeugen von Engpässen, der Voraussage von Fehlern und der schnellen Reaktion auf Ausfälle und Sicherheitsverletzungen beschäftigen.

Diese Vielzahl potenzieller Möglichkeiten führt zu einigen Herausforderungen: Welche Möglichkeiten werden die besten Resultate erzielen und wie kann ein erfolgreiches Ergebnis sichergestellt werden? Die Rolle des PoC besteht darin, es den Entscheidungsträgern zu ermöglichen, diese Fragen zu beantworten und dabei den Nutzen zu maximieren sowie das Risiko zu minimieren.

### Was ist ein Proof of Concept?

Eine Proof of Concept (PoC) ist eine „geschlossene", aber funktionierende Lösung, die nach klaren Kriterien evaluiert und getestet werden kann – vom Erstellen eines Anforderungsprofils bis zur erfolgreichen Umsetzung. PoCs ermöglichen Managern für jegliches KI-Projekt oder -Programm:

- Mehr unmittelbaren Nutzen zu bieten
- Kompetenzen und Erfahrung zu erwerben
- Hardware, Software und Service-Optionen zu testen
- Potenzielle Datenengpässe zu identifizieren und zu beseitigen
- Die Auswirkungen auf die IT-Infrastruktur und das gesamte Unternehmen aufzuzeigen
- Das positive Bild von KI und das Vertrauen der Nutzer zu stärken

## Schritt 1: Bestimmen der Möglichkeiten

Es ist unerlässlich, sich von Anfang an im Klaren darüber zu sein, was mit der KI erreicht werden soll, warum sie für das Unternehmen wichtig ist und wie man sicher sein kann, dass sie funktionieren wird. Wenn Sie noch nicht ermittelt haben, wie Sie am besten von KI profitieren können, sollten Sie feststellen, wo KI am unmittelbarsten etwas bewirken kann:

- Finden Sie heraus, wie andere in Ihrer Branche KI nutzen.
- Finden Sie Bereiche in Ihrem Unternehmen, denen KI bei der Lösung eines klar definierten Problems helfen oder einen zusätzlichen Mehrwert bieten kann.
- Nutzen Sie bestehendes Fachwissen, indem Sie auf die firmenintern, bereits vorhandenen Kenntnisse und Erfahrungen zurückgreifen.

Wenn Sie eine Auswahlliste der Bereiche erstellt haben, in denen Ihr Unternehmen von KI profitieren kann, können Sie jede Möglichkeit in Hinblick auf mehrere Kriterien bewerten. Eine solche Bewertung muss

[2] https://www.gartner.com/newsroom/id/3856163

| Natürliche Sprachverarbeitung | Maschinelles Sehen | Vorausschauende Instandhaltung |
|---|---|---|
| Als sich das FinTech-Start-Up Clinc* für die Entwicklung von Finie entschied - einem persönlichen KI-Assistenten, der dazu dient, Menschen mittels natürlicher Sprache bei ihren persönlichen Finanzen zu helfen -, wurde ihm klar, dass bestehende Algorithmen zur natürlichen Sprachverarbeitung nicht ausreichten, um die gewünschten Kundenerlebnisse zu bieten. In Zusammenarbeit mit Intel hat Clinc* die neuesten Technologien in den Bereichen maschinelles Lernen und Deep Learning eingesetzt, um eine kundenorientierte KI-Lösung zu entwickeln. | Der Gourmet-Süßwarenhändler Lolli & Pops nutzt maschinelles Sehen und KI, um personalisierte Kundenerlebnisse anzubieten. Durch maschinelles Sehen erkennt „Magic Makers*“ von Lolli & Pops die Mitglieder des Treueprogramms, wenn sie einen Laden betreten. Durch den Einsatz von KI-gestützter Datenanalyse hat der Händler Zugriff auf die Vorlieben der Mitglieder und gibt personalisierte Empfehlungen ab. So erhalten die Kunden eine VIP-Behandlung und es wird sichergestellt, dass sie den Laden auch in Zukunft immer wieder besuchen. | Die Deutsche Telekom nutzt SAP*-Lösungen auf Cloud-Servern mit Intel® Xeon® E7 Prozessoren, um Daten von Leistungs-, Temperatur-, Schwingungs- oder Drehzahlsensoren zu erfassen und vorausschauende Analysen durchzuführen. Das fördert die vorausschauende Instandhaltung des Unternehmens und verringert proaktiv Stillstandszeiten und Wartungskosten durch Identifizierung defekter oder abgenutzter Teile, bevor ein größerer Schaden auftreten kann. |

nicht lange dauern, aber die folgenden Fragen können Lücken in der Planung aufzeigen und Sie davor bewahren, sich unvermittelt in ein KI-Projekt zu stürzen:

- Besteht Klarheit darüber, welches Problem gelöst werden soll, was seine besonderen Anforderungen sind und wie sich der Erfolg messen lässt? Haben Sie bereits andere Lösungen in Erwägung gezogen oder eingesetzt, die sich mit diesem Problem beschäftigen, und diese zugunsten von KI ausgeschlossen?
- Ist der Rahmen der Möglichkeit klar abgegrenzt? Können Sie zum Bespiel eine einfache Übersicht der genutzten Datenbestände, zentralen Elementen, betroffenen Personen und anderer Faktoren erstellen? Wird sie Teil einer größeren Lösung sein?
- Verfügen Sie über die technologischen Ressourcen und notwendigen finanziellen Mittel, um das zu erreichen? Können Sie ohne technische, vertragliche oder andere Hindernisse auf die Datenquellen zugreifen?
- Sind die Auswirkungen auf das Geschäft signifikant genug, um den Aufwand zu rechtfertigen? Ein starker Zugewinn an Sichtbarkeit ist wichtig, um das Vertrauen der Nutzer in KI und das Engagement der Stakeholder zu stärken.
- Sind Motivation und Akzeptanz ausreichend, beispielsweise in Form von Unterstützung durch Führungskräfte? Ist der betroffene Geschäftsbereich vollständig in die Lösung dieses Problems involviert?
- Ist der zeitliche Rahmen angemessen? Wurde das Bereitstellungsteam klar definiert und verfügt es über ausreichend Zeit, Fachkenntnisse und Motivation für die Umsetzung?
- Besitzt das Unternehmen eine umfassendere Data-Science- und/oder KI-Strategie und ist diese an seinen Zielen ausgerichtet? Welche Infrastruktur und Kompetenz im Bereich Data Science besitzt das Unternehmen bereits?
- Wie geht es nach einem erfolgreichen PoC weiter? Gibt es finanzielle Mittel zur Wartung oder Skalierung der Lösung? Wurde Ihre operative IT-Abteilung informiert und ist sie bereit, die Umsetzung zu unterstützen?

Letztlich dienen diese Fragen der Überprüfung von Nutzen, Kosten und Risiken der Lösung, was als Business Case beschrieben werden kann – obwohl ein formelles Dokument zu viel für einen einfachen PoC sein kann.

Lesen Sie das Whitepaper „Das KI-Bereitschaftsmodell“, um eine umfassendere Sichtweise auf die Bereitschaft für KI zu gewinnen.

## Schritt 2: Beschreiben des Problems und Erstellen eines Datenprofils

Nachdem Sie Ihre Möglichkeit identifiziert und getestet haben, können Sie sich darauf konzentrieren, das zu lösende Problem detaillierter zu verstehen und zu beschreiben, um es breiteren Kategorien wie Denken, Wahrnehmung und maschinelles Sehen zuzuordnen.

Insbesondere zu Beginn der KI-Einführung besteht eine Herausforderung in der Frage, ob innerhalb des Unternehmens ausreichend Fachwissen vorhanden ist. Intel hilft Unternehmen mit seinen technischen Experten und Beratungspartnern, und bietet zudem auch Schulungen an. Dazu gehört ein 12-wöchiger selbstgesteuerter Grundkurs über maschinelles Lernen und Deep Learning. Er bietet Entwicklern die Möglichkeit zu lernen, wie sie die Problemstellungen im Unternehmen den passenden KI-Technologien von Intel zuordnen können.

Innerhalb Ihres KI-Workflows ist das auch ein guter Zeitpunkt, um eine Reihe technischer Fragen zu stellen, die auf die Lösung Einfluss haben könnten. Zum Beispiel:

- Bevorzugen Sie einen Hard-/Software-Hersteller und warum (Benchmarkdaten, Gesamtbetriebskosten, bevorzugter Lieferant)?
- Bevorzugen Anforderungen in Hinblick auf Sicherheit/gesetzliche Vorgaben/Daten/andere Aspekte On-Premises-Systeme gegenüber der Cloud?
- Gibt es für Ihre Lösung lokale Selbsthilfe-Angebote oder wird sie im Rechenzentrum bereitgestellt?
- Wie hoch ist die aktuelle prozentuale Auslastung des Rechenzentrums und wie wichtig ist die Performance pro Watt?
- In welchem Intervall und Umfang werden Sie neue Daten für Training/Inferenz erhalten?
- Wie werden rohe Daten und daraus resultierende Erkenntnisse während der Speicherung und Übertragung geschützt?

## Schritt 3: Entwickeln und Einsetzen der Lösung

Die nächste Frage ist, wie man die im PoC getestete Lösung entwickelt und einsetzt. Wie in Abbildung 3 gezeigt, besteht diese aus einem Technologie-Stack. Dazu gehören:

- Zugrundeliegende Produkte und Systeminfrastrukturen
- KI-spezifische Software, um die Infrastruktur voranzutreiben
- Geeignete KI-Frameworks, um die geplante Lösung zu unterstützen
- Visualisierung und Front-End-Software und/oder -Hardware

In dieser Phase werden Sie sich vielleicht fragen, ob Sie Hard- und Software entwickeln, kaufen oder wiederverwenden und/oder Cloud-Services nutzen sollten. Das begleitende White Paper Wählen Sie die beste Infrastruktur-Strategie zur Unterstützung Ihrer KI-Lösung zeigt Ihnen unter anderem die Optionen für Entwickeln vs. Kaufen auf. Wenn Kunden in marktführende Intel® Xeon® Prozessoren investieren, bedeutet das in vielen Fällen, dass erste Tests unter Einsatz der bestehenden Infrastruktur stattfinden können.

Selbst Infrastruktur und Software, die in Übereinstimmung mit bewährten Methoden entwickelt und getestet wurden, erfordern immer noch die Berücksichtigung der KI-Anforderungen. Im Besonderen gehört dazu die Notwendigkeit eines konstanten, hochwertigen Datenfeeds. Datenwissenschaftler können zusammen mit IT-Systemarchitekten die Bereitstellungsarchitektur vom Rechenzentrum bis zum Netzwerkrand entwickeln, wobei Software-Integration, Netzwerkverbindung, physische Merkmale und andere Aspekte berücksichtigt werden. Möglicherweise müssen mehrere Optionen getestet werden. Das sollte durch den Einsatz eines „Test and Learn"-Ansatzes unterstützt werden, damit so viele Erfahrungen wie möglich gewonnen werden.

Nach dessen Abschluss können Sie andere KI-spezifische Elemente der Lösung durcharbeiten und so die Modelle erstellen, trainieren und anpassen.

**Erstellen der Modelle**

Das Erstellen von Modellen ist die KI-Kernaufgabe. Das erfordert Data Scientists, die Trainingsdaten verwenden und Parameter verwalten, um iterative Testläufe durchzuführen. Auf diese Weise können sie Modelle auf Konvergenzgenauigkeit überprüfen, bevor sie umfassender trainiert und angepasst werden.

**Trainieren und Anpassen**

Das Trainieren und Anpassen ist der rechenintensivste Teil des KI-Workflows. Hier bestimmen Data Scientists, mit welchen Parametern ihre Modelle mit den verfügbaren Trainingsdaten am effizientesten konvergieren, während sie gleichzeitig auch die traditionellen IT-Aufgaben Job-Scheduling und Infrastrukturmanagement erledigen.

Das ist äußerst arbeitsintensiv. Die Data Scientists verbringen ihre Zeit damit, sich manuell Daten zu beschaffen und hunderte Experimente durchzuführen. Das kann durch das Intel® Nervana™ Deep Learning Studio – eine umfassende Software-Suite – ebenfalls erleichtert werden. Diese Lösung ermöglicht Gruppen von Data Scientists die Reduktion dieser Testzyklen und die Entwicklung und den Einsatz unternehmenstauglicher Deep-Learning-Lösungen in Rekordzeit.

## Schritt 4: Evaluieren des geschäftlichen Nutzens

Im Zuge der Lösungsentwicklung werden Sie Evaluationskriterien für den PoC definiert haben: Techniker können diese in Evaluationskriterien umwandeln, die ausgearbeitet, gemessen und – am besten automatisch – laufend getestet werden.

Die folgenden Evaluationskriterien können in Hinblick auf den geschäftlichen Nutzen angewendet werden:

### Kann ich eine Standard-CPU für KI nutzen?

Grafikprozessoren (GPUs) haben zwar eine Rolle dabei gespielt, die algorithmische Verarbeitung weiterzuentwickeln, die bei KI zum Einsatz kommt. Aber Deep Learning (DL) funktioniert mittlerweile auf allgemein gebräuchlichen, CPU-basierten Architekturen.

In der Vergangenheit benötigte das DL-Training auf einer CPU unangemessen viel Zeit, da es Prozessoren an Hardware- und insbesondere Software-Optimierungen mangelte. Das ist nicht mehr der Fall. Die neueste Generation von skalierbaren Intel® Xeon® Prozessoren hat den Leistungsabstand drastisch verringert. Die skalierbaren Intel® Xeon® Prozessoren bieten bei der Ausführung von Deep-Learning-Aufgaben in Hinblick auf den Trainingsdurchsatz eine um bis zu 127-mal höhere Performance im Vergleich zur Vorgängergeneration ohne optimierte Software.[3]

Außerdem lassen sich die skalierbaren Intel® Xeon® Prozessoren sehr effizient horizontal skalieren (Scale-Out) und können dadurch beinahe jedes beliebige Deep-Learning-Durchsatzprofil erreichen. Die Möglichkeit zur Nutzung von CPUs löst eine Reihe von Problemen, denen Unternehmen mit ausschließlich GPU-basierter KI gegenüberstehen:

- Die GPU-Architektur erfordert, dass die Datenpipeline temporär in einen GPU-Datenspeicher hinein- und dann zurückkopiert wird. Das durchbricht den typischen Datenfluss und die Verarbeitungs-Toolkette.
- Im Vergleich zu CPU-basierten Nodes kann es schwierig sein, die Rechenleistung auf einer großen Anzahl von GPU-basierten Nodes im „Non-Cluster"-Modus zu skalieren und zu verwalten. Dadurch reduziert sich die potenzielle Zeitersparnis beim Training.
- Es kann Speicherbeschränkungen für Unternehmen geben, die mit dem kleinen Arbeitsspeicher einer GPU (16 oder 32 GB) sehr große Bilder verarbeiten wollen, z. B. im Gesundheitswesen und bei Geoinformationssystemen (GIS).
- Bei jeglicher Domain-spezifischen Architektur kann es zu Unterauslastung kommen. Bei einer Allzweck-CPU können unbenutzte Nodes für andere Workloads verwendet und/oder als IaaS vermietet werden.

Immer mehr Unternehmen erkennen die Vorteile von CPUs für das Deep Learning. Intel arbeitet mit Kunden wie Facebook, deepsense.ai, OpenAI, AWS, EMR, Databricks, Alibaba, Microsoft und Cloudera zusammen. Diese Liste wird noch länger werden, wenn sich die KI-Performance-Lücke zwischen CPU und GPU schließt.

Weitere Informationen darüber, wie Technologien von Intel die Grundlage für Ihren KI-PoC bilden, finden Sie in unserer Infografik über die Anatomie eines Proof of Concept für KI.

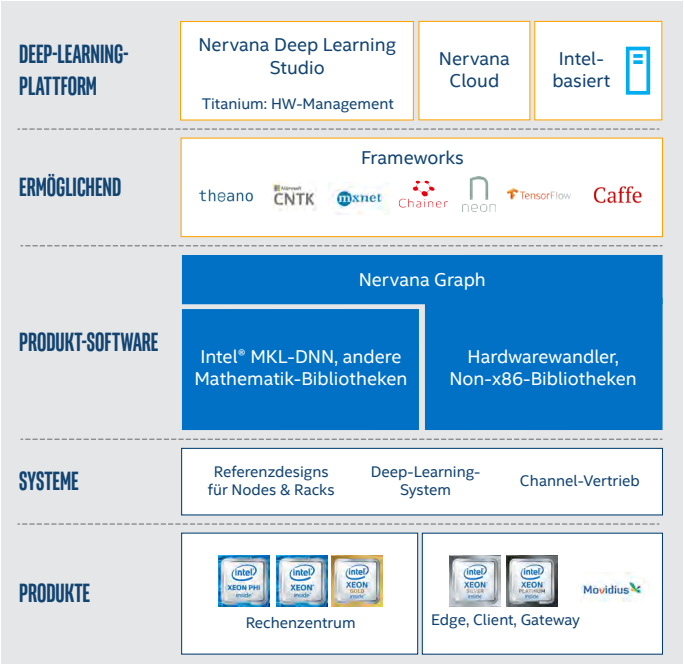


**Abbildung 3: Die Architektur der KI-Lösung kann als Stack dargestellt werden.**

- **Genauigkeit**: Liefert die Lösung korrekte Ergebnisse und Erkenntnisse und sind diese wiederholbar?
- **Vollständigkeit**: Nutzt die Lösung alle Datenquellen korrekt?
- **Rechtzeitigkeit**: Werden die Erkenntnisse geliefert, wo und wann sie benötigt werden?

Weitere Kriterien gelten für die Lösung und dafür, ob sie erwartungsgemäß funktioniert:

- **Skalierbarkeit**: Wird die Lösung immer noch funktionieren, wenn Datenvolumen oder Anwenderzahlen im Laufe der Zeit oder sprunghaft steigen?
- **Kompatibilität**: Ist die Lösung durch die Verwendung von Standardprotokollen offen für die Integration von Datenquellen und -diensten Dritter?
- **Flexibilität**: Kann sich die Lösung an veränderte Gegebenheiten anpassen, falls sich die Daten oder Modelle ändern?
- **Konstruktion**: Wie unkompliziert lassen sich falsche Ergebnisse aus einem trainierten Modell beseitigen?

Zuletzt muss die Lösung auf Grundlage dessen evaluiert werden, was KI-Spezialisten mit dem Begriff „Erklärbarkeit" („Explainability") bezeichnen - die Qualität von Entscheidungen. Die Kriterien im Bereich der Erklärbarkeit umfassen:

- Voreingenommenheit: Wie kann sichergestellt werden, dass das KI-System nicht ein voreingenommenes Bild der Welt hat (oder vielleicht ein unvoreingenommenes Bild einer voreingenommenen Welt), das auf Mängeln des Trainingsmodells, der Daten oder der Zielfunktion basiert? Was, wenn seine Entwickler bewusst oder unbewusst voreingenommen sind?
- **Fairness**: Wenn Entscheidungen basierend auf einem KI-System getroffen werden, wie kann verifiziert werden, dass diese fair getroffen wurden? Und was bedeutet „fair" in diesem Zusammenhang – fair für wen?
- **Kausalität**: Kann das Modell nicht nur korrekte Schlussfolgerungen ziehen, sondern auch eine Erklärung für die zugrundeliegenden Phänomene bieten?
- **Transparenz**: Werden KI-basierte Erkenntnisse derart erklärt, dass der Nutzer sie versteht? Und auf welcher Grundlage kann eine Erkenntnis infrage gestellt werden?
- **Sicherheit**: Wie können Nutzer Vertrauen in die Zuverlässigkeit des KI-Systems gewinnen, mit oder ohne Transparenz darüber, wie es zu seinen Schlussfolgerungen gelangt?

## Schritt 5: Skalieren des PoC

Das Problem wurde definiert, die Lösung entwickelt und für die Daten wurden ein Profil und ein Modell erstellt. Der PoC wurde erstellt, getestet und eingesetzt. Was geschieht nun als nächstes?

Positive Erfahrungen der Nutzer können zu einer größeren Nachfrage und daher zu noch mehr Erfolg führen. Es besteht aber auch das Risiko, dass der PoC überhöhtem Interesse zum Opfer fällt. Sie können einiges dafür tun, dass Ihr PoC weiterhin ein Erfolg bleibt, damit er die Grundlage für eine umfassendere KI-Strategie bilden kann:

- **Skalieren der Inferenzfähigkeiten.** KI wird nicht linear skaliert: Wenn zum Beispiel von einer Single- auf eine Multi-Node-Konfiguration umgestellt wird, werden 50 Prozessoren nicht unbedingt die 50-fache Leistung liefern. Sie müssen eine Multi-Node-Konfiguration auf sehr ähnliche Weise testen und konfigurieren, wie Sie es bei Ihrer Single-Node-Konfiguration getan haben.
- **Skalieren einer breiteren Infrastruktur.** Eine erfolgreiche KI erfordert, dass jedes Glied in der Schlussfolgerungskette untersucht wird. Prüfen Sie bestehende Technologieplattformen, Netzwerke und Speichersysteme mit dem Ziel, die verfügbare Datenmenge zu steigern und ihre Aktualität sowie Latenz zu verbessern. Das minimiert die Gefahr künftiger Engpässe und maximiert den Nutzen, den Sie aus Ihren Datenquellen ziehen können.
- **Abstimmen und Optimieren der PoC-Lösung.** Im Laufe der Zeit werden Sie Ihre Fertigkeiten hinsichtlich der Verbesserung und Erweiterung Ihrer KI-Lösung ausbauen. Sie können Software in Bereichen wie Datenpflege und -kennzeichnung optimieren und mit neuen Modellen - die möglicherweise bessere Ergebnisse bringen - experimentieren, sie trainieren und einsetzen.
- **Ausweiten auf andere Geschäftsszenarien.** Ihr PoC kann vielleicht in anderen Teilen Ihres Unternehmens zum Einsatz kommen. So kann zum Beispiel eine Predictive-Maintenance-Lösung, die für einen Bereich Ihrer Fertigung verwendet wurde, nun ausgebaut werden. Sie können einen Portfolio-Ansatz anwenden, um zu bestimmen, wie Sie den PoC auf eine größere Nutzerbasis erweitern.
- **Planen für Verwaltung und Abläufe.** Aufgrund ihrer Art verlangen viele KI-Anwendungsfälle von den Systemen, dass sie in Echtzeit schlussfolgern, anstatt das offline oder im Batch-Modus zu tun. Außerdem müssen Modelle möglicherweise im Laufe der Zeit neu

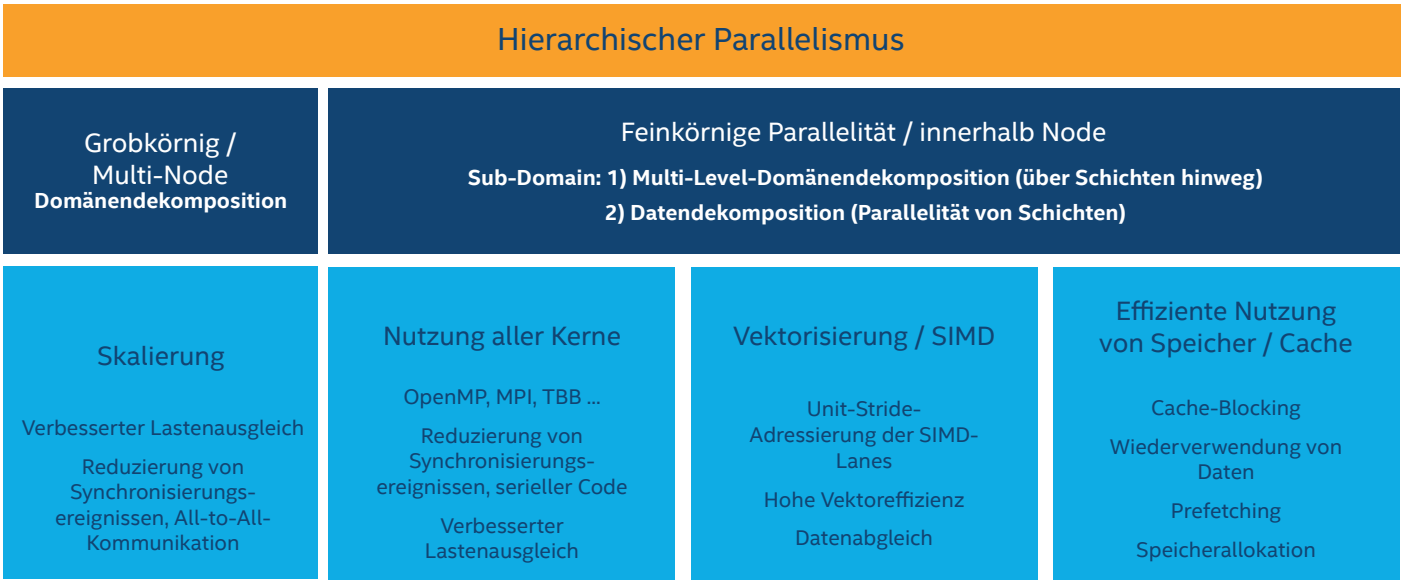


**Abbildung 4: Performance-Optimierung auf modernen Plattformen**

trainiert und aktualisiert werden. Diese Faktoren werden zusätzliche Anforderungen an die Servicebereitstellung stellen. Stellen Sie sicher, dass vorab ausreichend Zeit und qualifizierte Mitarbeiter zugewiesen werden, damit der PoC weiterhin genutzt werden kann.

Abbildung 4 zeigt eine Reihe von Bereichen, die Sie sich näher ansehen können, um Ihre KI-Lösung weiter zu optimieren. Intel hat die populärsten KI-Frameworks - darunter Theano* und TensorFlow* - direkt für die Intel® Architektur optimiert und so wesentliche Performancesteigerungen erzielt. Intel plant, mit dem Intel® nGraph™ Compiler[4] in Zukunft noch weitere Frameworks zu optimieren.

Zusätzlich hat Intel BigDL entwickelt, um Deep Learning im Big-Data-Bereich einzusetzen. Dabei handelt es sich um eine Distributed-Deep-Learning-Bibliothek für Apache Spark*, die direkt auf bestehenden Spark*- oder Apache-Hadoop*-Clustern eingesetzt werden kann und es Ihren Entwicklerteams ermöglicht, Deep-Learning-Anwendungen wie Scala*- oder Python*-Programme zu schreiben.

## Klein beginnen, überschaubar bleiben

Um die Erfolgschancen zu maximieren und schnell Nutzen zu generieren, empfehlen wir, klein und überschaubar zu beginnen. Dadurch wird sichergestellt, dass die Ziele klar und von Anfang an geschäftsorientiert sind.

Intel engagiert sich mit ganzer Kraft dafür, seine Kunden dabei zu unterstützen, das Potenzial von KI zu nutzen. Das geschieht durch:

- **Lösungen** – Data Scientists, technische Dienste und Referenzlösungs-Teams von Intel entwickeln, nutzen und teilen KI-Lösungen, damit Sie aus Daten schneller Erkenntnisse gewinnen.
- **Plattformen** – Intel bietet mehrere einsatzbereite, vollständige Stacks und benutzerfreundliche Systeme, die schnell bereitgestellt werden können, um den KI-Innovationszyklus zu beschleunigen.
- **Tools** – Die KI-Software-Suite von Intel beinhaltet Produktivitätstools für Data Scientists und Entwickler, die den Deep-Learning-Innovationszyklus verkürzen.
- **Frameworks** – Intel optimiert die populärsten Deep-Learning-Frameworks der Open-Source-Gemeinschaft, um auf einer Reihe von Prozessor-Plattformen Höchstleistungen zu bieten.
- **Bibliotheken** – Intel beschleunigt KI-Anwendungen durch die Optimierung von Primitives und schafft mit dem Intel® nGraph™ Compiler Frameworks, die jede Zielhardware mit Spitzenleistung nutzen können.
- **Hardware** – Das umfassende Produktportfolio von Intel reicht vom Rechenzentrum bis zum Netzwerkrand und adressiert alle aktuellen KI-Ansätze.

Intel ermöglicht es seinem starken Ökosystem und Partnernetzwerk, den Fortschritt von KI durch umfangreiche Zusammenarbeit in der Branche zu beschleunigen. Ebenso wie die von ihm betreuten Unternehmen, durchläuft Intel einen Prozess und verschiebt durch modernste F&E die Grenzen des KI-Computing in Bereiche wie Neuromorphic- und Quanten-Computing.

Dies ist jedoch erst der Anfang.

### Weitere Informationen

- **Nähere Informationen unter:** ai.intel.com
- **Erkunden Sie –** Nutzen Sie die Performance von Intel: Optimierte Bibliotheken & Frameworks
- **Treten Sie in Kontakt –** Wenden Sie sich für Hilfe an Ihren Ansprechpartner bei Intel und informieren Sie sich über die PoC-Möglichkeiten

[4] Bitte beachten Sie, dass die Frameworks über ein unterschiedliches Niveau von Optimierung und Konfigurationsprotokollen verfügen. Details finden Sie unter ai.intel.com/framework-optimizations/

## Referenzen und Quellen

Intel veröffentlicht auf ai.intel.com Fallstudien, Referenzlösungen und -architekturen, die Kunden dafür nutzen können, ihre eigenen ähnlichen KI-Lösungen zu entwickeln und zu realisieren.

Die Herausforderungen und Chancen erklärbarer KI, https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/

Die Zukunft des Einzelhandels wird ganz von künstlicher Intelligenz bestimmt, https://ai.intel.com/future-retail-artificial-intelligence

KI-Akademie von Intel – Lernen Sie die Grundlagen, https://software.intel.com/en-us/ai-acadmy/basics

Loihi – Intels neuer selbstlernender Chip verspricht Beschleunigung von künstlicher Intelligenz, https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/

Wie man als Entwickler von KI beginnt, von Niven Singh, https://software.intel.com/en-us/articles/how-to-get-started-as-a-developer-in-ai

Ihr persönlicher KI-Assistent fürs Finanz- und Bankenwesen, https://www.intel.de/content/www/de/de/analytics/artificial-intelligence/ai-personal-assistant.html

Vorausschauende Analyse hilft, die Instandhaltungskosten zu reduzieren, https://www.intel.de/content/www/de/de/big-data/intel-sap-telekom-predictive-analytics-paper.html

[3] INFERENZ mit FP32 Batchgröße Caffe GoogleNet v1 256 AlexNet 256.

Die Leistungsschätzungen wurden vor der Implementierung der neuesten Software-Patches und Firmware-Updates ermittelt, die als Gegenmaßnahmen für die als „Spectre“ und „Meltdown“ bezeichneten Exploits eingesetzt wurden. Die Implementierung dieser Updates kann dazu führen, dass diese Ergebnisse auf Ihr Gerät oder System nicht zutreffen. In Leistungstests verwendete Software und Workloads können speziell für die Leistungseigenschaften von Intel®-Mikroprozessoren optimiert worden sein. Leistungstests wie SYSmark* und MobileMark* werden mit spezifischen Computersystemen, Komponenten, Softwareprogrammen, Operationen und Funktionen durchgeführt. Jede Veränderung bei einem dieser Faktoren kann abweichende Ergebnisse zur Folge haben. Als Unterstützung für eine umfassende Bewertung Ihrer geplanten Anschaffung sollten Sie noch andere Informationen und Leistungstests heranziehen – auch im Hinblick auf die Leistung des betreffenden Produkts in Verbindung mit anderen Produkten. Weiterführende Informationen finden Sie unter http://www.intel.de/performance. Quelle: Intel Messergebnisse, Juni 2017. Anmerkung zur Optimierung: Unter Umständen können Intel®-Compiler bei Optimierungen, die nicht für Intel®-Mikroprozessoren spezifisch sind, auch bei Mikroprozessoren anderer Hersteller denselben Optimierungsgrad erzielen. Zu diesen Optimierungen gehören Befehlssätze für SSE2, SSE3 und SSSE3 sowie weitere Optimierungen. Intel übernimmt keine Garantie für die Verfügbarkeit, Funktionalität oder Wirksamkeit von Optimierungen für Mikroprozessoren, die nicht von Intel hergestellt wurden. Vom Mikroprozessor abhängige Optimierungen in diesem Produkt sind für die Anwendung in Verbindung mit Intel®-Mikroprozessoren bestimmt. Bestimmte, nicht für die Intel®-Mikroarchitektur spezifische Optimierungen sind für Intel®-Mikroprozessoren reserviert. Entnehmen Sie weitere Informationen zu den spezifischen Befehlssatzerweiterungen, auf die dieser Hinweis zutrifft, bitte den entsprechenden Benutzer- und Referenzhandbüchern.

**Konfigurationen für Inferenz-Durchsatz:**
Prozessor: Intel® Xeon® Platinum 8180 (2 Sockel, 2,5 GHz, 28 Kerne, HT aktiviert, Turbo aktiviert). Arbeitsspeicher: 376,46 GB (12 Slots, 32 GB, 2.666 MHz). CentOS Linux-7.3.1611-Core. SSD sda RS3WC080 HDD 744,1 GB, sdb RS3WC080 HDD 1,5 TB, sdc RS3WC080 HDD 5,5 TB, Deep-Learning-Framework Caffe-Version: f6d01efbe93f70726ea3796a4b89c612365a6341. Topologie: googlenet_v1. BIOS: SE5C620.86B.00.01.0004.071220170215. MKL-DNN-Version: ae00102be506ed0fe2099c6557df2aa88ad57ec1 NoDataLayer. Gemessen: 1190 Bilder/s im Vergleich zu Plattform: Intel® Xeon® Prozessor E5-2699 v3 (2 Sockel, 2,3 GHz, 18 Kerne, HT aktiviert, Turbo deaktiviert, CPU-Skalierung mittels intel_pstate Treiber auf „Performance“ gesetzt), Arbeitsspeicher: 256 GB, DDR4-2133 ECC RAM. CentOS Linux, Release 7.3.1611 (Core), Linux-Kernel 3.10.0-514.10.2.el7.x86_64. Betriebssystemlaufwerk: Interne Festplatte Seagate* Enterprise ST2000NX0253 (2 TB, 2,5”) Leistung gemessen mit: Umgebungsvariablen: KMP_AFFINITY=granularity=fine, compact, 1,0, OMP_NUM_THREADS=36, CPUFreq festgelegt mit: cpupower frequency-set -d 2.2G -u 2.2G -g performance. Deep-Learning-Frameworks: Intel® Caffe: (http://github.com/intel/caffe/), Revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inferenz gemessen mit „caffe time --forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein Test-Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. Topologie-Spezifikation von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet, AlexNet und ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19) und https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet-Benchmarks; die Dateien wurden aktualisiert und verwenden das neuere prototxt-Format, sind aber funktional gesehen äquivalent). GCC 4.8.5, MKLML Version 2017.0.2.20170110. BVLC-Caffe: https://github.com/BVLC/caffe, Inferenz & Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein Test-Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. BVLC/Caffe (http://github.com/BVLC/caffe), Revision 1b09280f5233cafc62954c98ce8bc4c204e7475 (Commit-Datum 14.5.2017). BLAS: ATLAS Ver. 3.10.1.

**Konfigurationen für Trainings-Durchsatz:**
Prozessor: Intel® Xeon® Platinum 8180 (2 Sockel, 2,5 GHz, 28 Kerne, HT aktiviert, Turbo aktiviert). Arbeitsspeicher: 376,46 GB (12 Slots, 32 GB, 2.666 MHz). CentOS Linux-7.3.1611-Core. SSD sda RS3WC080 HDD 744,1 GB, sdb RS3WC080 HDD 1,5 TB, sdc RS3WC080 HDD 5,5 TB, Deep-Learning-Framework Caffe-Version: f6d01efbe93f70726ea3796a4b89c612365a6341. Topologie: alexnet. BIOS: SE5C620.86B.00.01.0009.101920170742. MKL-DNN-Version: ae00102be506ed0fe2099c6557df2aa88ad57ec1 NoDataLayer. Gemessen: 1023 Bilder/s im Vergleich zu Plattform: Intel® Xeon® Prozessor E5-2699 v3 (2 Sockel, 2,3 GHz, 18 Kerne, HT aktiviert, Turbo deaktiviert, CPU-Skalierung mittels intel_pstate Treiber auf „Performance“ gesetzt), Arbeitsspeicher: 256 GB, DDR4-2133 ECC RAM. CentOS Linux, Release 7.3.1611 (Kern), Linux-Kernel 3.10.0-514.el7.x86_64. Betriebssystemlaufwerk: Interne Festplatte Seagate* Enterprise ST2000NX0253 (2 TB, 2,5”) Leistung gemessen mit: Umgebungsvariablen: KMP_AFFINITY=granularity=fine, compact, 1,0, OMP_NUM_THREADS=36, CPUFreq festgelegt mit: cpupower frequency-set -d 2.2G -u 2.2G -g performance. Deep-Learning-Frameworks: Intel® Caffe: (http://github.com/intel/caffe/), Revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inferenz gemessen mit „caffe time --forward_only“-Befehl, Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein Test-Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. Topologie-Spezifikation von https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet, AlexNet und ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19) und https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet-Benchmarks; die Dateien wurden aktualisiert und verwenden das neuere prototxt-Format, sind aber funktional gesehen äquivalent). GCC 4.8.5, MKLML Version 2017.0.2.20170110. BVLC-Caffe: https://github.com/BVLC/caffe, Inferenz & Training gemessen mit „caffe time“-Befehl. Für die „ConvNet“-Topologien wurde ein Test-Datensatz verwendet. Für andere Topologien wurden Daten im lokalen Datenspeicher gespeichert und vor dem Training im Systemspeicher zwischengespeichert. BVLC/Caffe (http://github.com/BVLC/caffe), Revision 1b09280f5233cafc62954c98ce8bc4c204e7475 (Commit-Datum 14.5.2017). BLAS: ATLAS Ver. 3.10.1.

Durch Technologien von Intel ermöglichte Funktionsmerkmale und Vorteile hängen von der Systemkonfiguration ab und können entsprechend geeignete Hardware, Software oder die Aktivierung von Diensten erfordern. Die Leistungsmerkmale variieren je nach Systemkonfiguration. Kein Computersystem bietet absolute Sicherheit. Informieren Sie sich beim Systemhersteller oder Fachhändler oder auf intel.de.

In Leistungstests verwendete Software und Workloads können speziell für die Leistungseigenschaften von Intel-Mikroprozessoren optimiert worden sein. Leistungstests wie SYSmark* und MobileMark* werden mit spezifischen Computersystemen, Komponenten, Softwareprogrammen, Operationen und Funktionen durchgeführt. Jede Veränderung bei einem dieser Faktoren kann abweichende Ergebnisse zur Folge haben. Als Unterstützung für eine umfassende Bewertung Ihrer geplanten Anschaffung sollten Sie noch andere Informationen und Leistungstests heranziehen – auch im Hinblick auf die Leistung des betreffenden Produkts in Verbindung mit anderen Produkten. Ausführlichere Informationen finden Sie unter http://www.intel.de/benchmarks.

Die geschätzten Ergebnisse wurden vor der Implementierung der neuesten Software-Patches und Firmware-Updates ermittelt, die als Gegenmaßnahmen für die als „Spectre“ und „Meltdown“ bezeichneten Exploits eingesetzt wurden. Die Implementierung dieser Updates kann dazu führen, dass diese Ergebnisse auf Ihr Gerät oder System nicht zutreffen.

Alle hierin gemachten Angaben können sich jederzeit ohne besondere Mitteilung ändern. Wenden Sie sich an Ihren Ansprechpartner bei Intel, um die neuesten Produktspezifikationen und Roadmaps zu erhalten.



0318/RD/CAT/PDF 337357-001DE