

KI mit InterSystems – Vector Search und GenAI

Dejan Lunginovic, Saba Nagervadze, Sylwester Boldt
Sales Engineering
InterSystems





Agenda



01 KI entmystifiziert

02 Machine Learning

03 Vector Search

04 Retrieval Augmented Generation

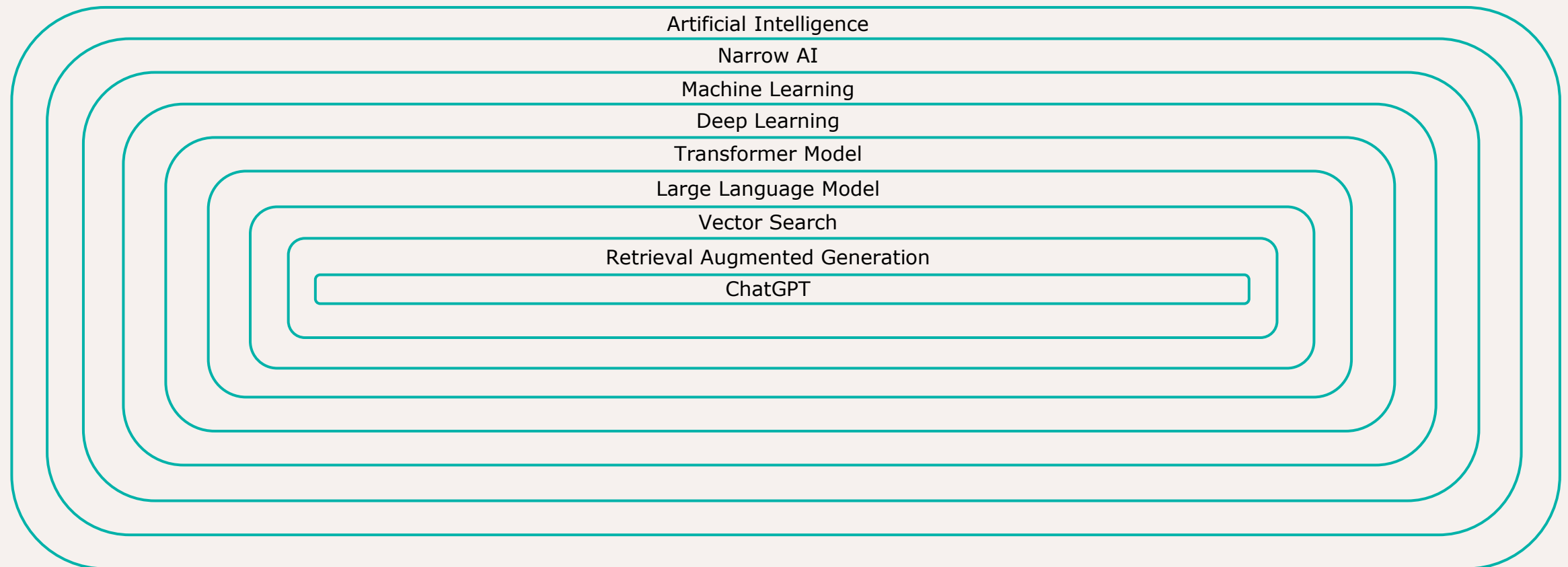
05 HealthShare AI Assistant

06 Fragen

KI entmystifiziert



Ein Versuch den gesamten KI-Stack als Diagramm einzuteilen – von der allgemeinen KI-Ebene bis hin zu LLMs, Vector Search und RAG-Anwendungen



KI entmystifiziert



Auf einen Punkt gebracht: KI ahmt menschliche Intelligenz nach

Artificial Intelligence

General AI: Allgemeine KI

Ziel ist die Lösung von Problemen, die das System noch nie zuvor gesehen hat (Forschung).

General AI = sehr gut in **allen Sachen** (theoretisch).

Beispiel: Eine hypothetische KI, die gleichzeitig Gedichte schreiben, Schach spielen, Diagnosen stellen und Roboter steuern kann, ohne speziell für jede Aufgabe programmiert zu sein.

Narrow AI: Spezialisierte KI

Ziel ist die Lösung eines spezifischen Problems, für das das System entwickelt und trainiert wurde.

Narrow AI = sehr gut in **einer Sache**.

Beispiel: Ein Sprachassistent wie Siri oder Google Translate kann Sprache verarbeiten, aber nicht selbstständig Mathematik betreiben oder ein Auto fahren.

KI entmystifiziert



Aus Daten lernen, um Muster zu erkennen und Entscheidungen zu treffen

Artificial Intelligence

Narrow AI

Machine Learning

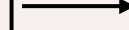
Maschinelles Lernen hat die neusten Durchbrüche in der Narrow AI ermöglicht

Kernidee: Muster in historischen Daten finden, um eine bestimmte Aufgabe zu lösen

Traditional

if/then/else
+ - : *

Input: Daten,
Rules



Der Computer kann die ihm zugewiesene Aufgabe ausführen
Output: Daten

Machine Learning

10011001
00101001

Input: Data,
Output-Data



Der Computer kann die Aufgabe ausführen, die er gelernt hat
Output: Regeln

KI entmystifiziert



Neuronale Netzwerke mit mehreren Schichten

Artificial Intelligence

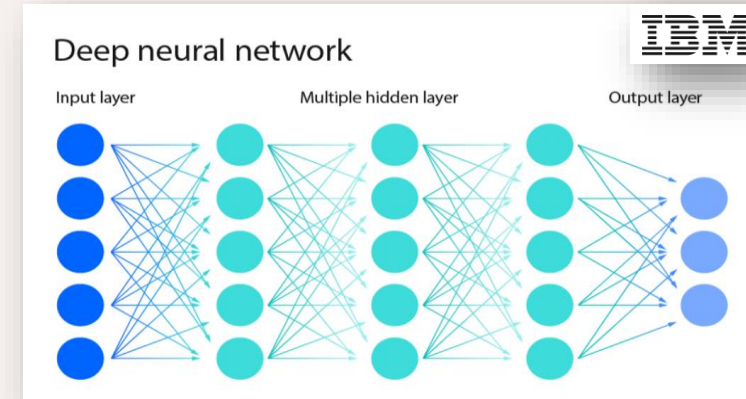
Narrow AI

Machine Learning

Deep Learning

Basiert auf neuronalen Netzwerken, die es erlauben, dass sich der Algorithmus eigenständig weiterentwickelt

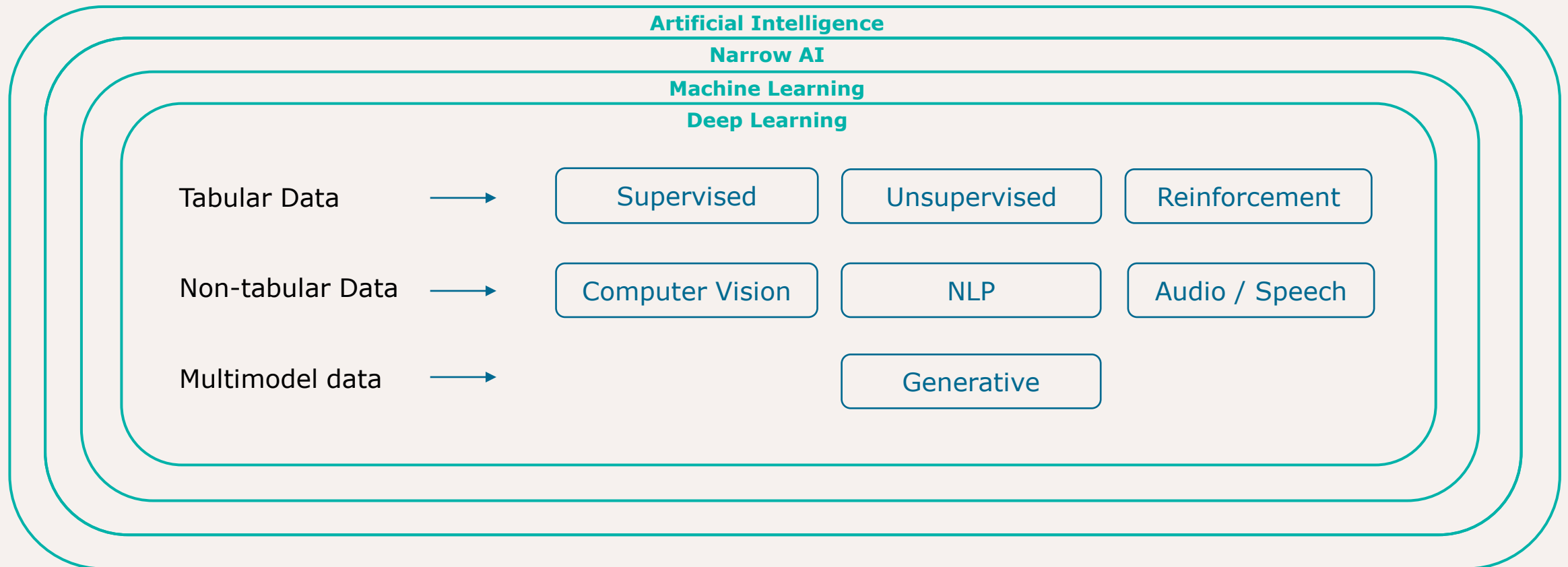
Ist eine besonders komplexe Form des maschinellen Lernens, die häufig auf große, unstrukturierte Datenmengen wie Text, Bilder oder Audiodateien angewendet wird



KI entmystifiziert



Verarbeitet große Mengen von unstrukturierten Daten



KI entmystifiziert



Analysiert alle relevanten Teile einer Daten-Eingabe

Artificial Intelligence

Narrow AI

Machine Learning

Deep Learning

Transformer Model

Wandeln Daten (Token) in Vektoren (Embeddings) um

NLP: Self-Attention – berechnet für jedes Token, wie stark es auf andere Token achtet oder mit denen in Beziehung ist

NLP: Positional Encoding – Zusatzinformationen im Satz

KI entmystifiziert



Ein LLM ist immer ein Transformer, aber nicht jeder Transformer ist ein LLM

Artificial Intelligence

Narrow AI

Machine Learning

Deep Learning

Transformer Model

Large Language Model

NLP: Ein LLM ist ein sehr grosses neuronales Netzwerk, das darauf trainiert wurde, Sprache zu verstehen, zu verarbeiten und zu generieren.

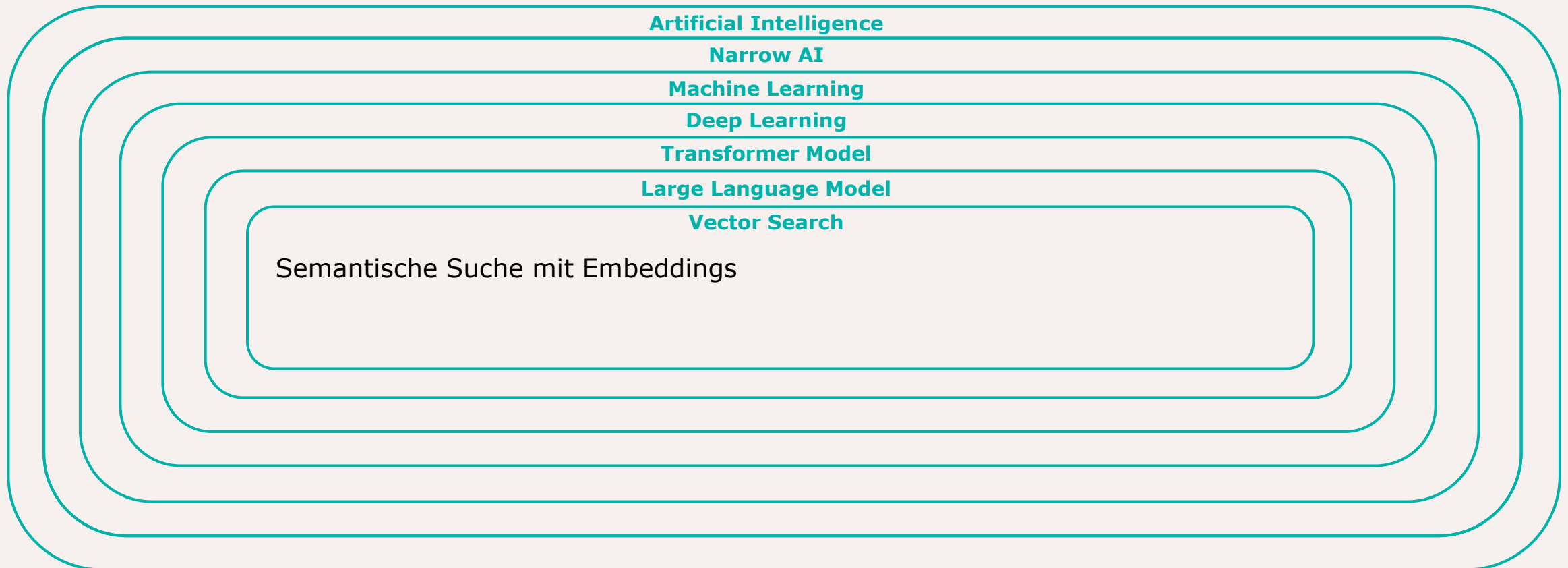
Es basiert fast immer auf der Transformer-Architektur und wird mit gewaltigen Datenmengen von Text trainiert.

Zum Beispiel: GPT, Claude, LLaMa, Gemini, Mixtral, etc.

KI entmystifiziert



Eine Technologie/Methodik aus dem Bereich Information Retrieval



KI entmystifiziert



Retrieval (Abruf) Augmented (Anreicherung) Generation (Antworterstellung)

Artificial Intelligence

Narrow AI

Machine Learning

Deep Learning

Transformer Model

Large Language Model

Vector Search

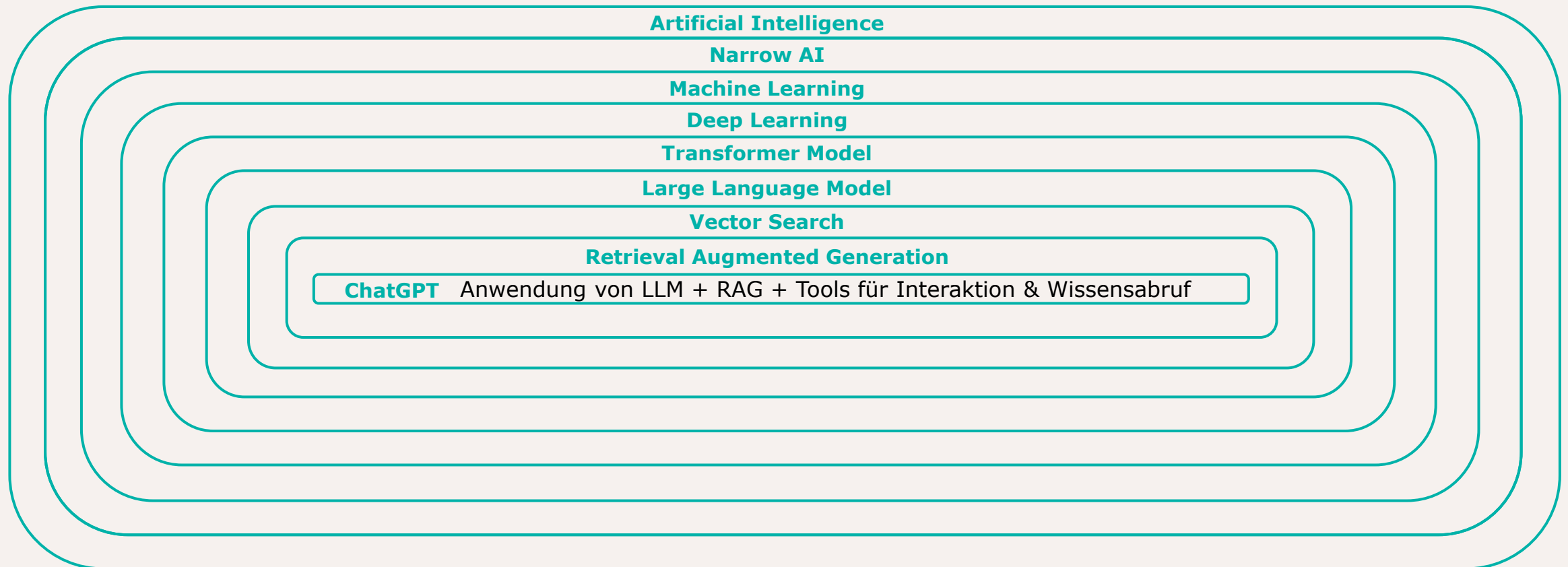
Retrieval Augmented Generation

Kombination aus LLM + Suche (Vector Search) + Daten-Eingabe (Context)

KI entmystifiziert



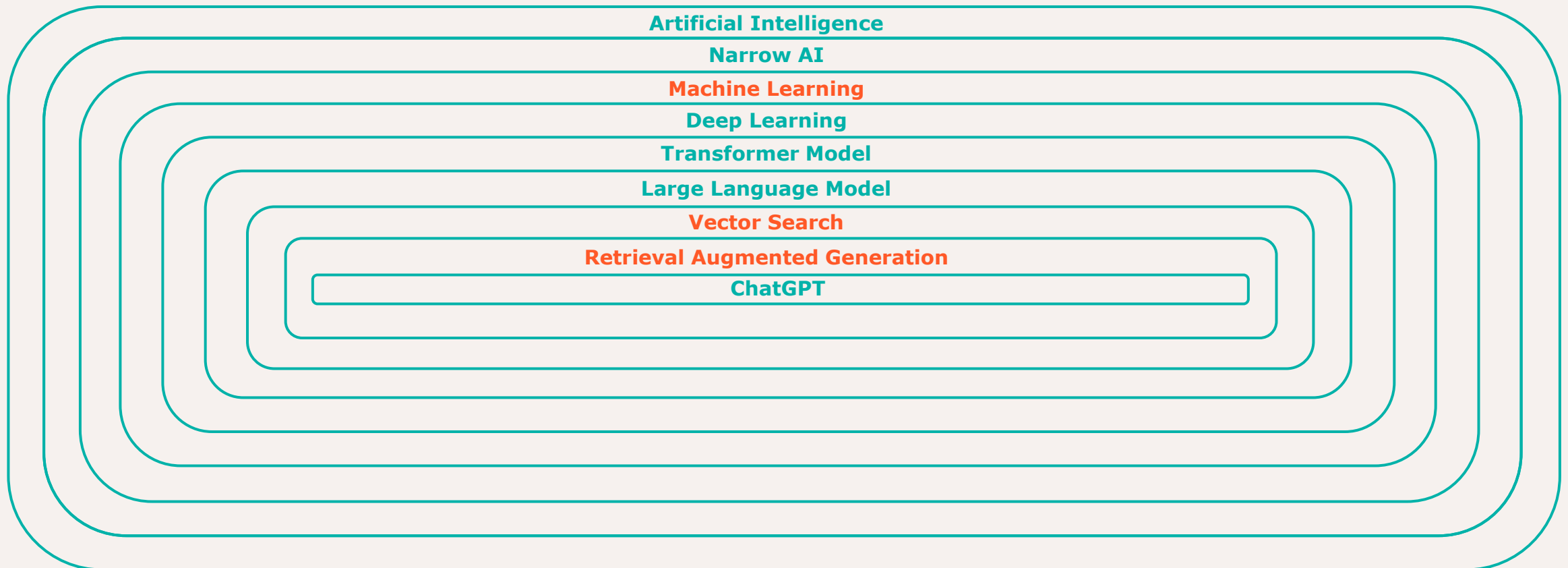
GPT ist das Modell – RAG ist eine Methode, um ein Modell wie GPT mit frischen, externen Daten anzureichern



KI entmystifiziert



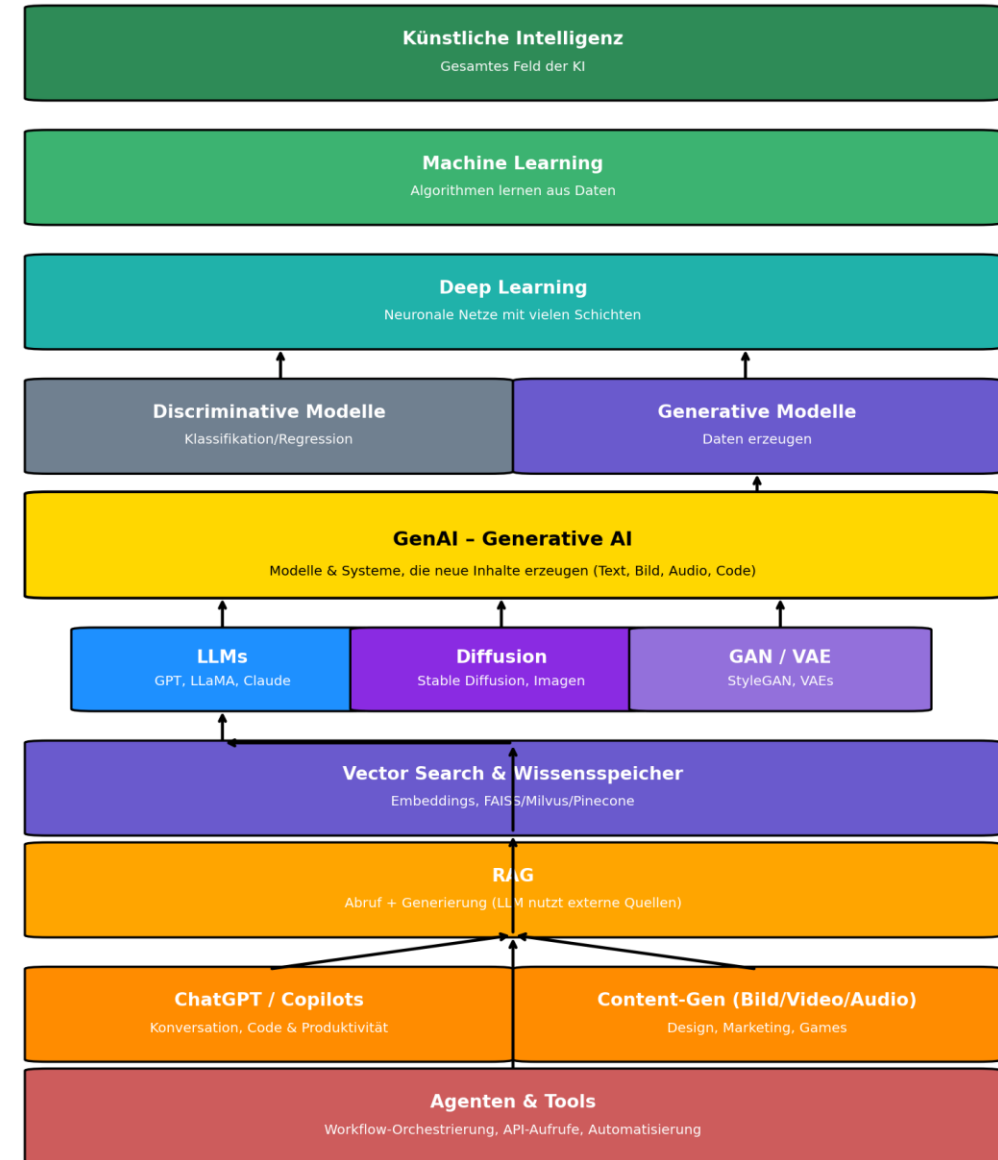
KI-Stack Diagramm ist nicht perfekt, aber brauchbar



KI entmystifiziert

KI-Stack inklusive GenAI mit ChatGPT 5 generiert

Erweiterter KI-Stack mit GenAI



Machine Learning

The Machine Learning Process



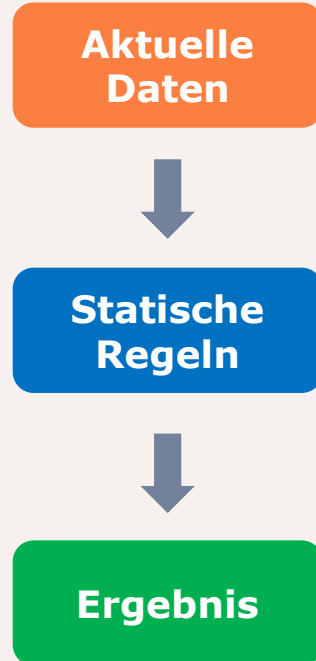
Automating the Machine Learning Process



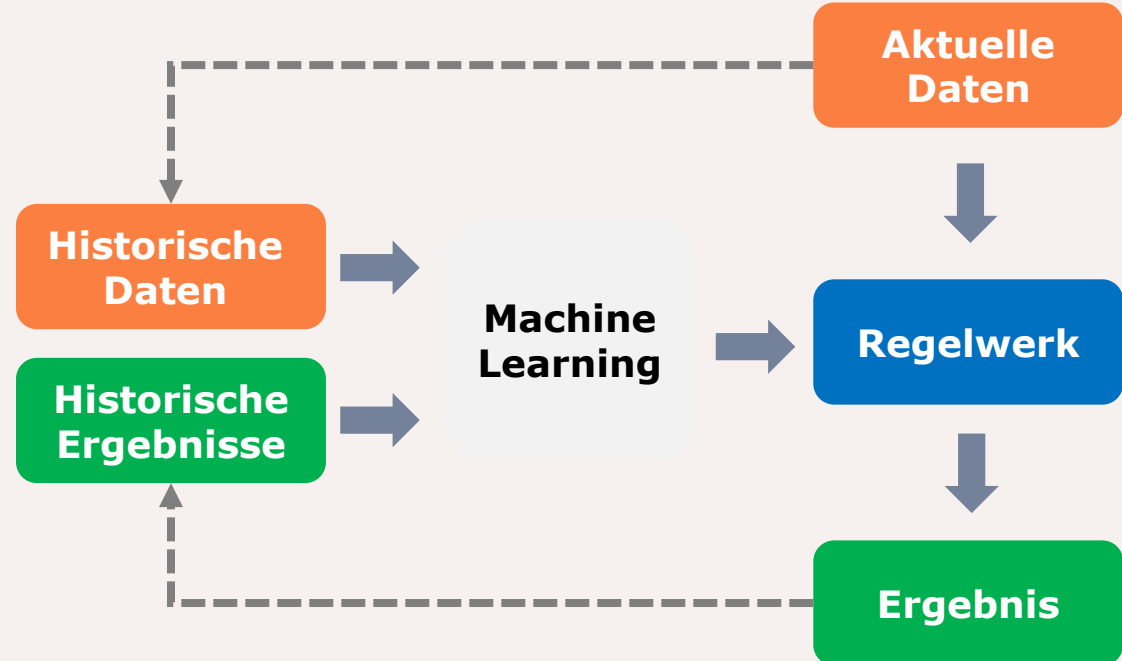
Machine Learning

Definition

Traditionelle Programmierung

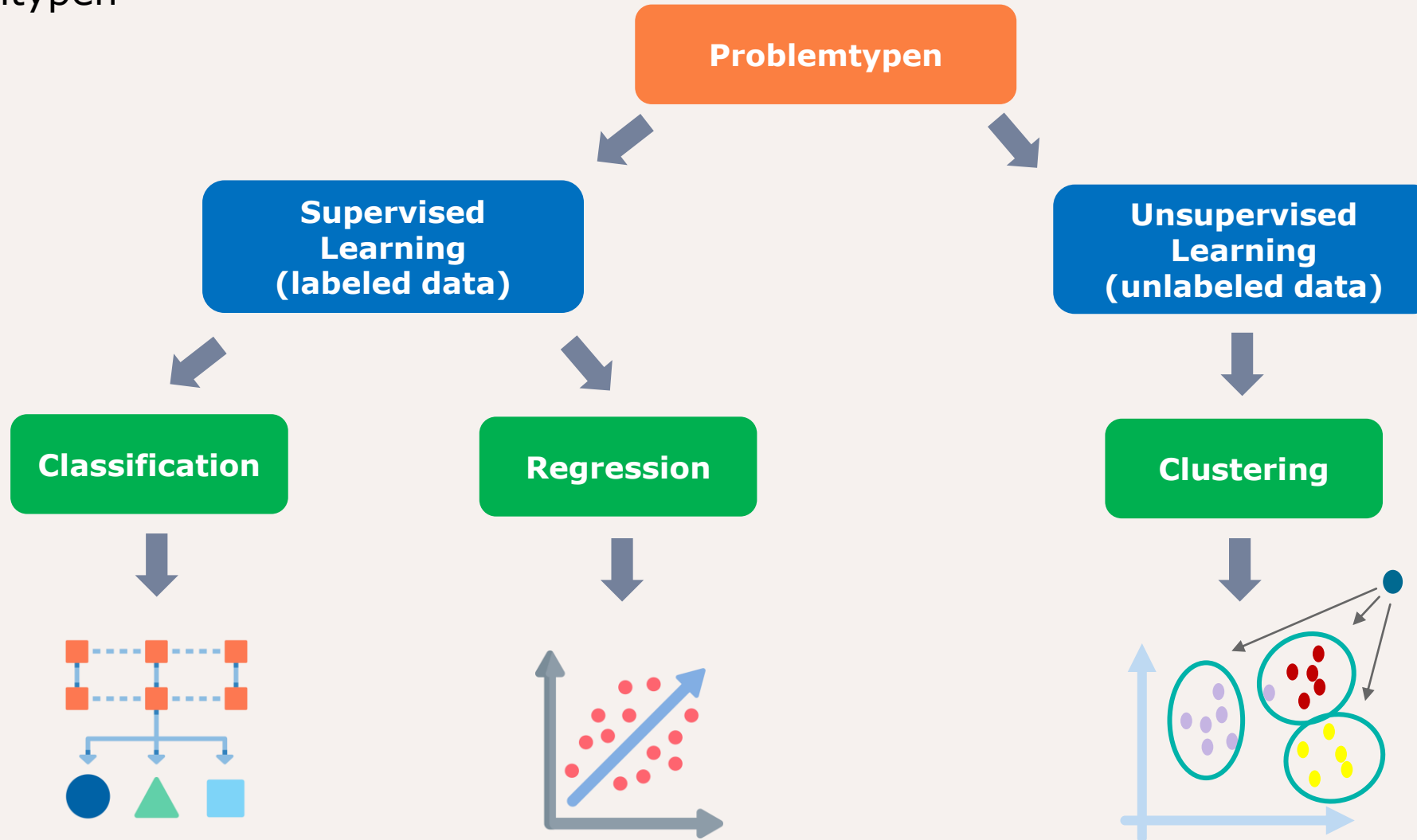


Machine Learning



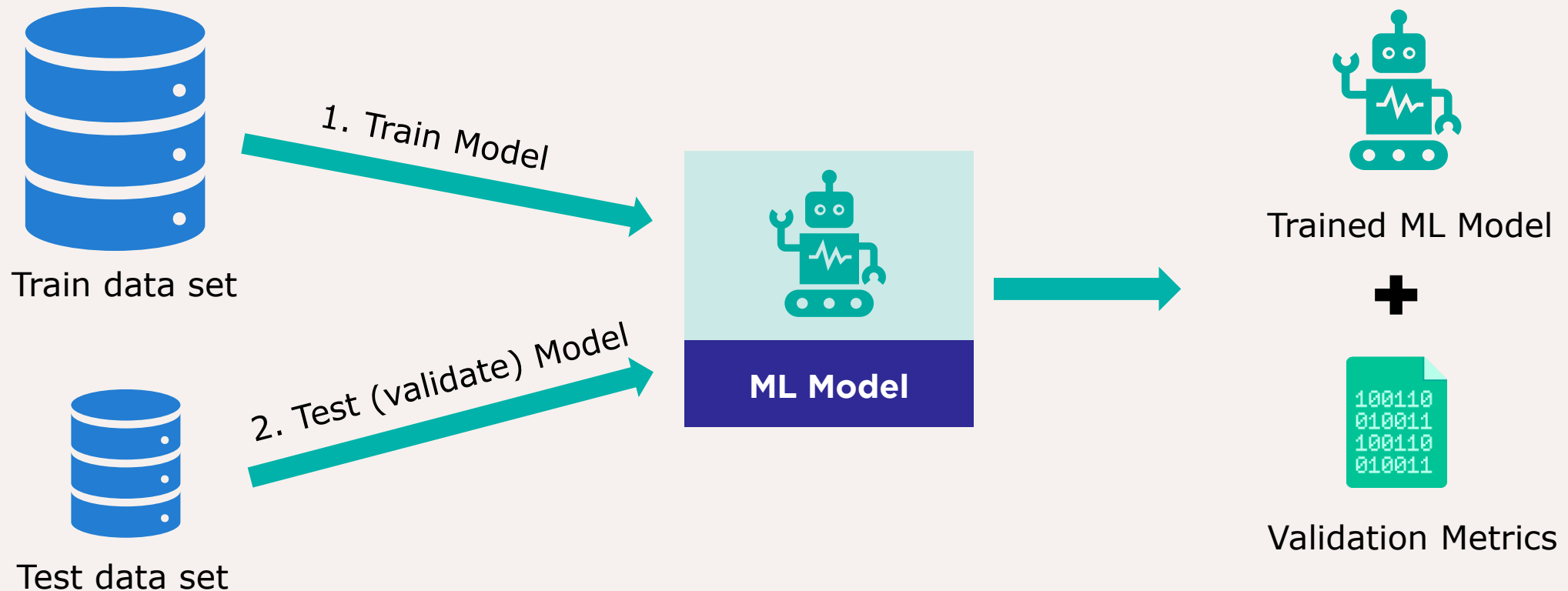
Machine Learning

Problemtypen



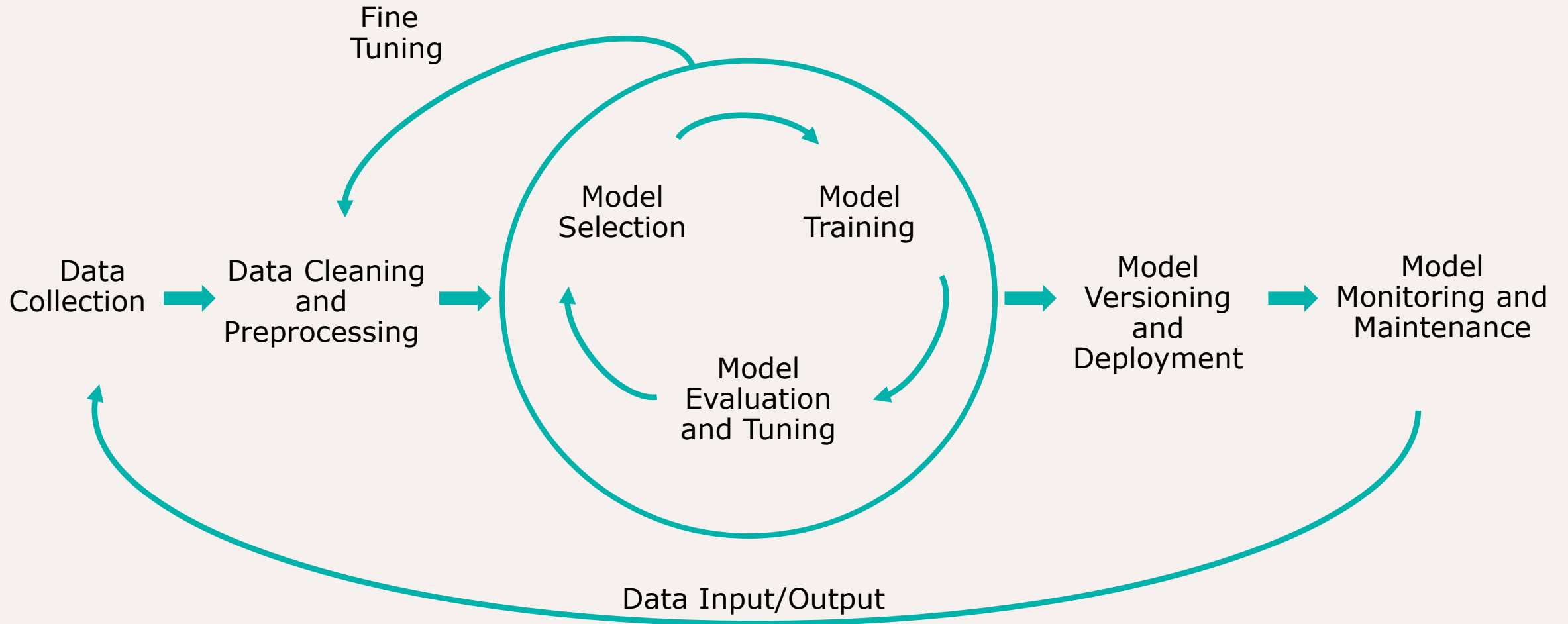
Machine Learning

Supervised Machine Learning Process



Machine Learning

Life-Cycle



Machine Learning



Definition

[Introduction to IntegratedML](#) | [Using IntegratedML](#) | [InterSystems IRIS Data Platform 2025.2](#)

Traditional Programming vs. Machine Learning

Traditional Programming



Machine Learning

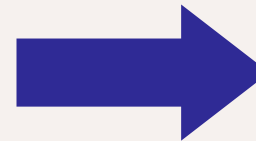


Machine Learning

InterSystems IntegratedML – the SQL Interface

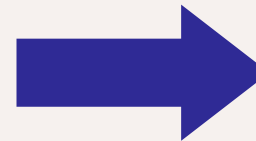


```
CREATE MODEL <model name>  
PREDICTING(<label>)  
FROM <table/view>
```



ML Model
Definition

```
TRAIN MODEL <model name>  
FROM <table/view>
```



Trained
ML Model

```
VALIDATE MODEL <model name>  
FROM <table/view>
```



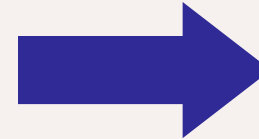
Validation
Metrics

Machine Learning

InterSystems IntegratedML – the SQL interface

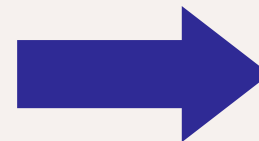


```
CREATE TIME SERIES MODEL <model name>  
PREDICTING(*) BY (date)  
FROM <table/view>
```



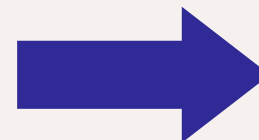
ML TimeSeries
Model Definition

```
SELECT PROBABILITY(<model>)  
FROM <table/view/record>
```



Calculate
Probabilities

```
SELECT PREDICT(<model>)  
FROM <table/view/record>
```



Predicted
Result

Machine Learning

Feature Engineering



Features					Label
Model	Temperature	Moisture	Load	Vibration	Thread Level
Alstom 200C	72.53	3.23	0.29	23.71	0
Mitsubishi 100B	73.01	0.002	0.25	247.01	0
Siemens 300D	69.00		0.02	314.88	0
Alstom 200C	85.86	1.78	0.51	15.92	1
Mitsubishi 100B	93.06	0.21	0.61	4.24	2
Siemens 300D	1072.72	1.04		124.85	1
Alstom 200C	71.01	1.96	0.18	13.88	0

Machine Learning

Klassischer Prozess

Feature Engineering

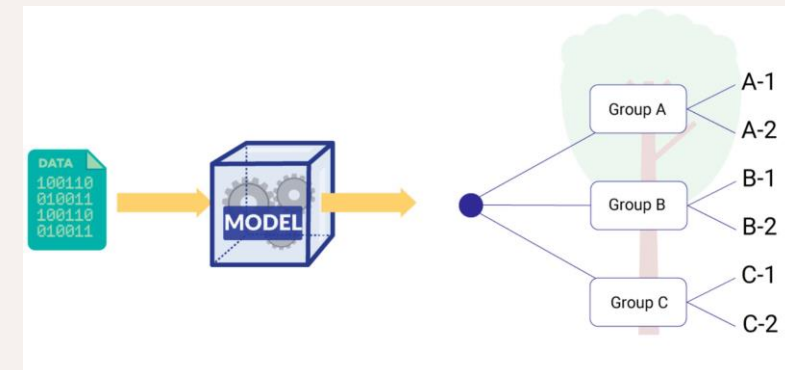
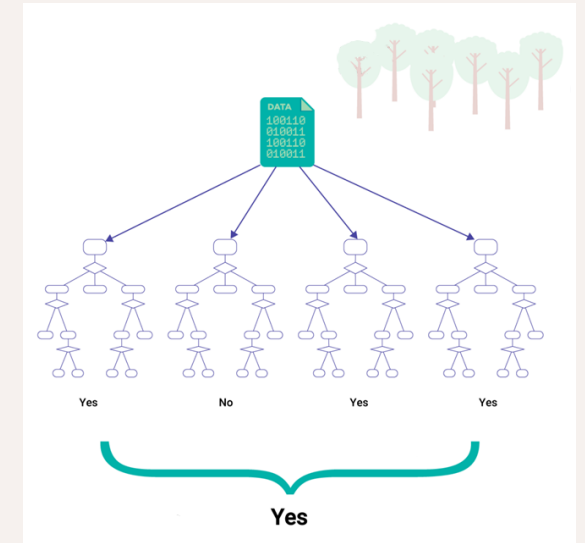
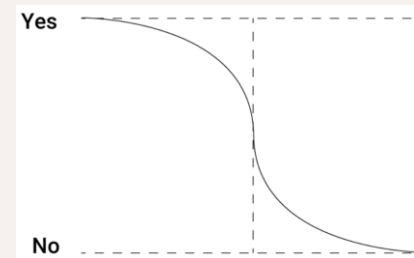
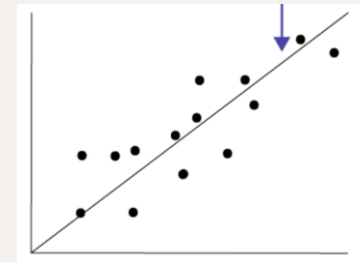
Features					Label
Model	Temperature	Moisture	Load	Vibration	Thread Level
Alstom 200C	72.53	3.23	0.29	23.71	0
Mitsubishi 100B	73.01	0.002	0.25	247.01	0
Siemens 300D	69.00		0.02	314.88	0
Alstom 200C	85.86	1.78	0.51	15.92	1
Mitsubishi 100B	93.06	0.21	0.61	4.24	2
Siemens 300D	1072.72	1.04		24.85	1
Alstom 200C	71.01	1.96	0.18	13.88	0

Problemtyp

Classification
Gruppenzugehörigkeit

Regression
Stetige Werte

ML Model



Machine Learning



Wie kann man klassischen Prozess automatisieren...?

Feature Engineering

- Automatisches Feature Engineering
- Unterstützung von Strings (NLP)

Problemtyp

- Automatische Anwendung von ML-Algorithmen

ML Model

- Automatisches Trainieren
- Automatische Validierung und Iteration
- Auswahl des optimalsten Algorithmus
- Keine Programmierung für das ML-Training notwendig
- Alternative ML-Provider
- PMML Import / Export

Machine Learning



Provider InterSystems AutoML

[Introduction to AutoML](#) | [AutoML Reference](#) | [InterSystems IRIS Data Platform 2025.2](#)

The Machine Learning Process



Automating the Machine Learning Process



Machine Learning



Supported Provider

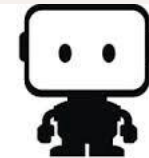


InterSystems AutoML

- Verwendung bekannter Python Bibliotheken (numpy, sklearn, XGBoost, TensorFlow ...)
- Aufruf über Embedded Python Support in IRIS
- Daten als Python Data Frames über Shared Memory (Push)



- Verwendung des OpenSource H2O Frameworks
- Aufruf über das IRIS Java Gateway
- Daten per JDBC über Shared Memory (Pull)



DataRobot

- Verwendung der ML Plattform über SaaS
- Aufruf über die DataRobot REST API
- Daten per JDBC (Pull)
- Verwendung trainierter Modelle im DataRobot GUI

Machine Learning

Live-Demo



IRIS ML Studio

Upload CSV to Split and Load

● Connected

Select CSV delimiter:

☒ Comma (,)

☐ Semicolon (;)

☐ Custom

Select split mode:

☒ Random

☐ Sequential (first/last)

Train/Test Split: 80% Train, 20% Test

10% Train

90% Train

Drag and drop file here

Limit 750 MB per file • CSV

Browse files

</> Model Commands

Create Model

Train Model

See Models

Validate Model

Delete Data

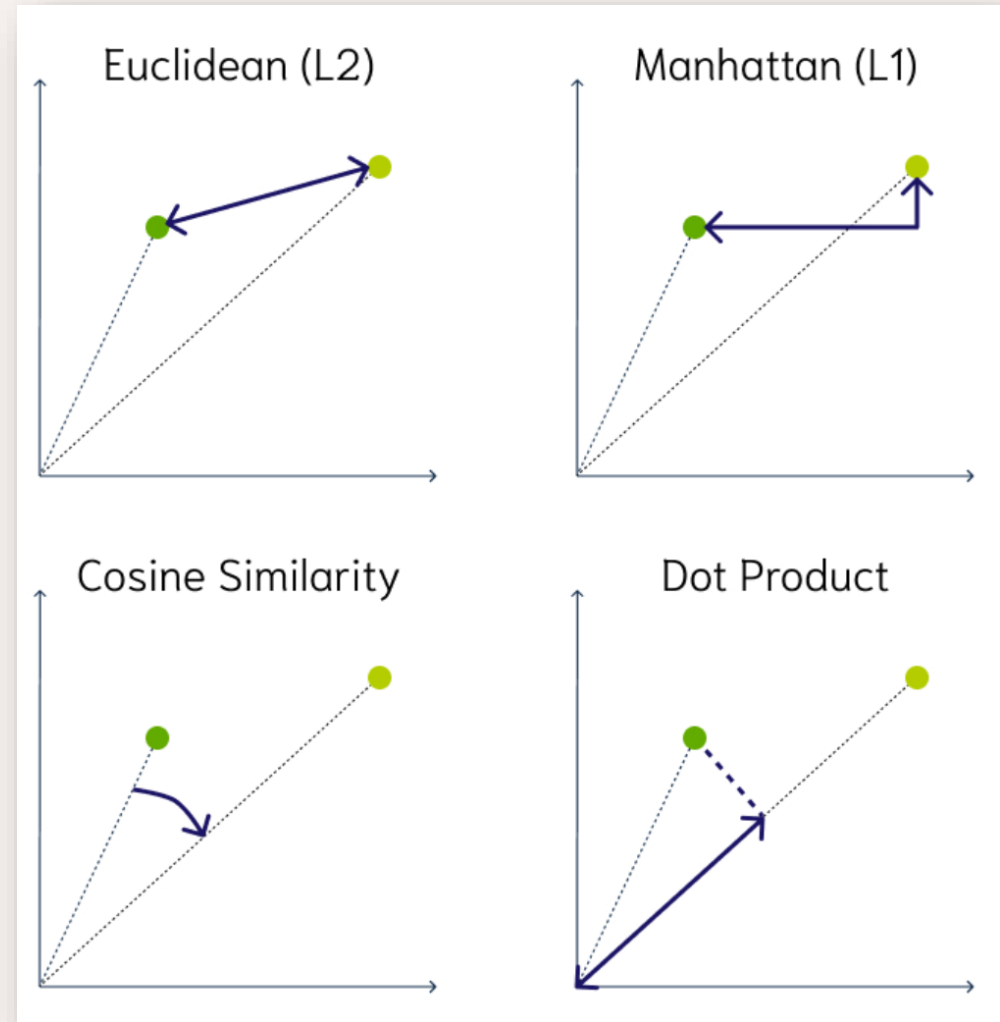
Delete Model

🔍 Predictions

Predict

Validation Metrics

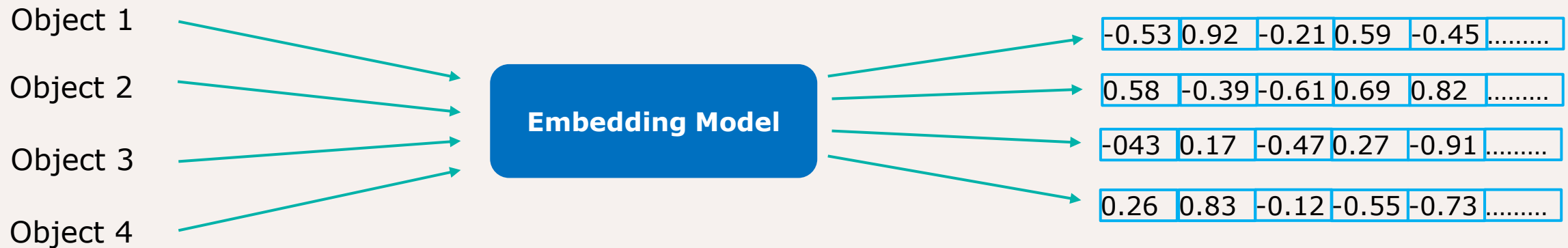
Vector Search



Vector Search



Grundvoraussetzung ist das Vorhandensein von Embeddings



Set an Objekten

Objekte als Vektoren

Vector Search



Grundvoraussetzung ist das Vorhandensein von Embeddings

Object 1

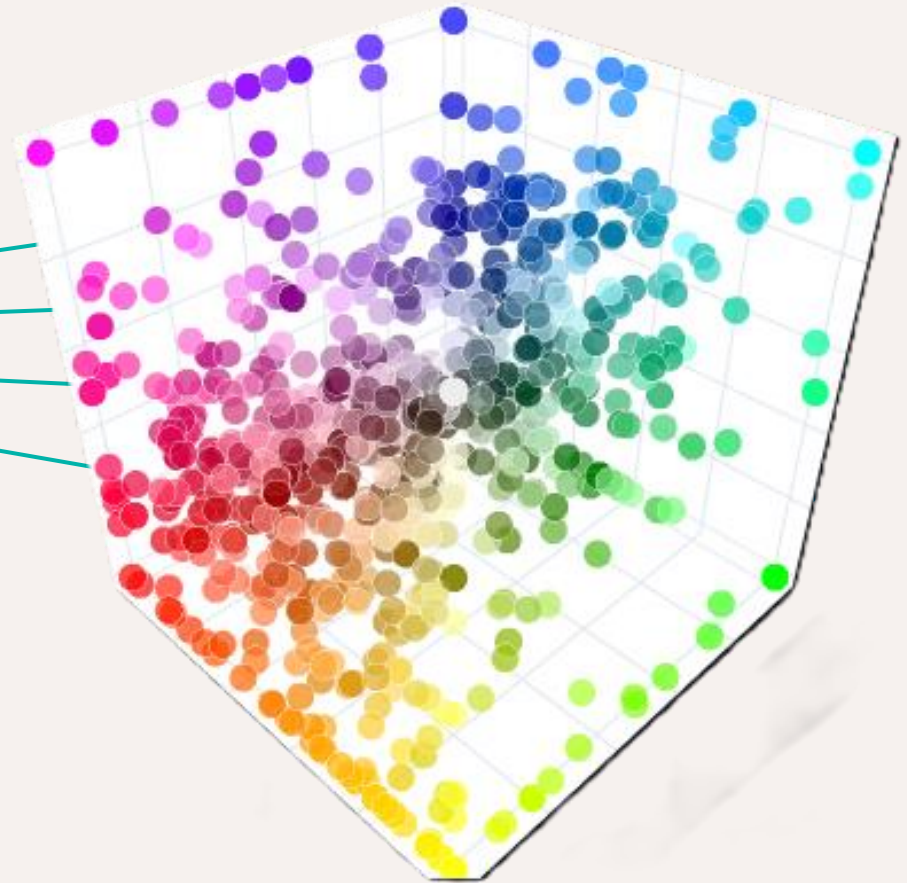
Object 2

Object 3

Object 4

Set an Objekten

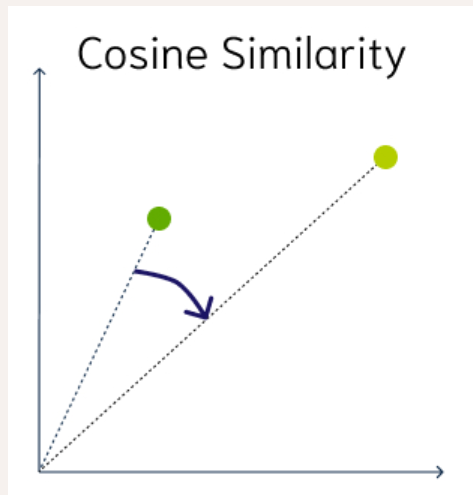
Embedding Model



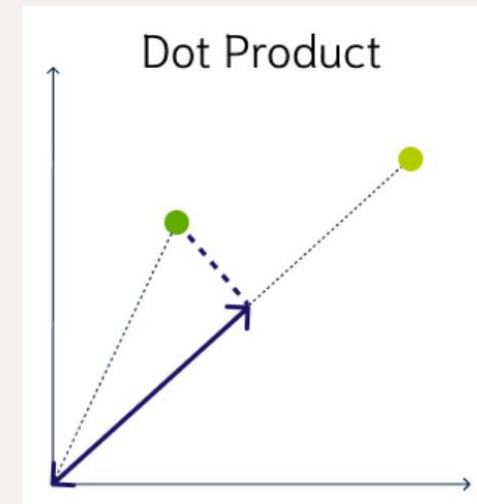
Vector Search



Basiert auf verschiedenen Nearest Neighbour Algorithmen



- Die Cosinus-Ähnlichkeit berechnet den Cosinus des Winkels zwischen zwei Vektoren. Ihr Wertebereich liegt zwischen $[-1, 1]$.

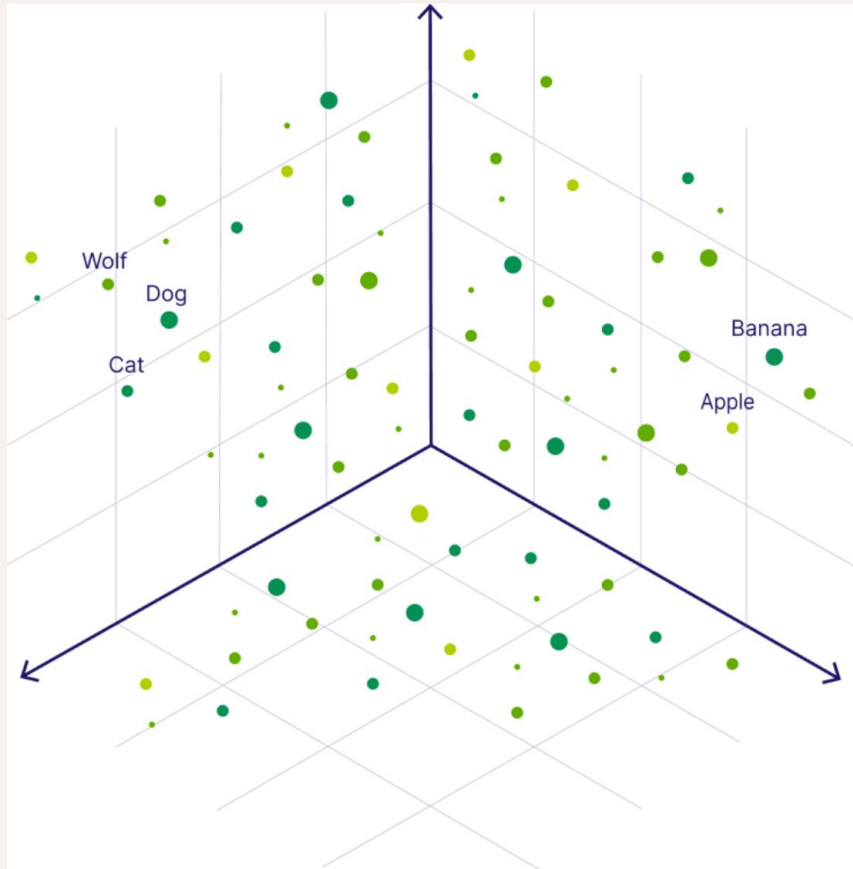


- Das Skalarprodukt berechnet das Produkt aus den Beträgen zweier Vektoren und dem Cosinus des Winkels zwischen ihnen. Sein Wertebereich ist $[-\infty, \infty]$.

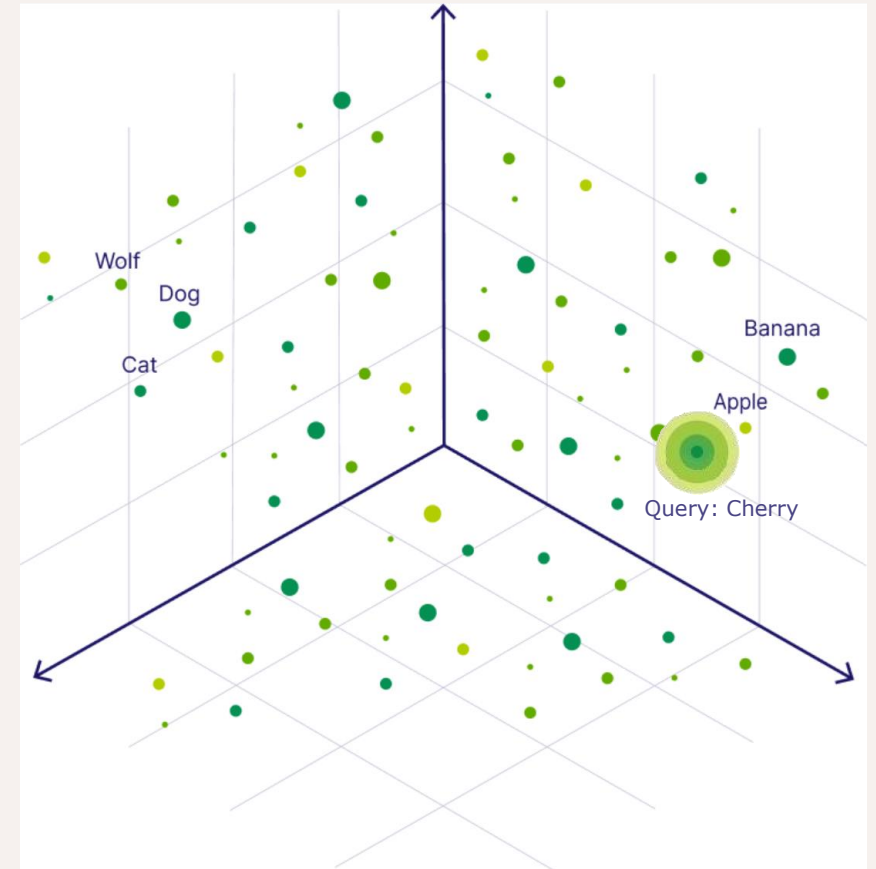
Vector Search



Vector Search basiert auf den Embeddings und den Nearest Neighbour Search Algorithmen



Vector Search
"Cherry"





Vector Search

Eine Technologie/Methodik aus dem Bereich Information Retrieval und semantische Suche

- **Einordnung in die KI-Landschaft**

- Machine Learning / Deep Learning: für die Erzeugung von Vektoren/Embeddings
- Information Retrieval: für das effiziente Suchen in grossen Datenmengen

- **Hauptaufgabe:**

- Anstatt exakte Schlüsselwörter zu vergleichen, sucht Vector Search nach *semantisch ähnlichen Inhalten*, indem Daten in Vektorräumen verglichen werden
- Ähnlichkeitssuche (Cosine Similarity, Dot Product, Euclidean Distance, Manhattan Distance) findet die nächsten Nachbarn (nearest neighbors)

- **Einsatzgebiete:**

- Semantic Search - z. B. Google-Suche mit Kontextverständnis (AI-Mode)
- RAG (Retrieval Augmented Generation) – Kombination von LLM und lokalen Daten/Dokumenten.
- Empfehlungssysteme – z. B. ähnliche Produkte, Songs, Filme
- Bild- und Videosuche - z. B. Suche per Beispielbild
- Plagiats- und Duplikatenerkennung
- Doublettensuche in Datenmengen

Wie kann ich Vector Search implementieren?



Live-Demo

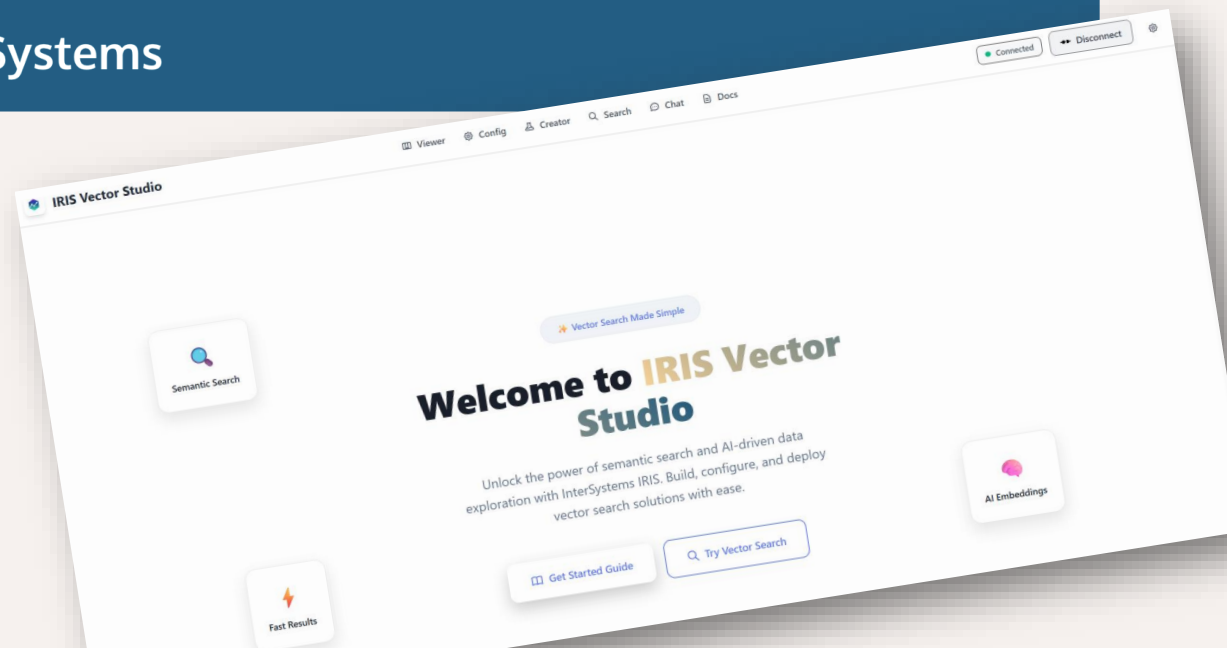
VectorSearch – Semantische Suche für bestehende Anwendungen

Raum Globe 2 - 3

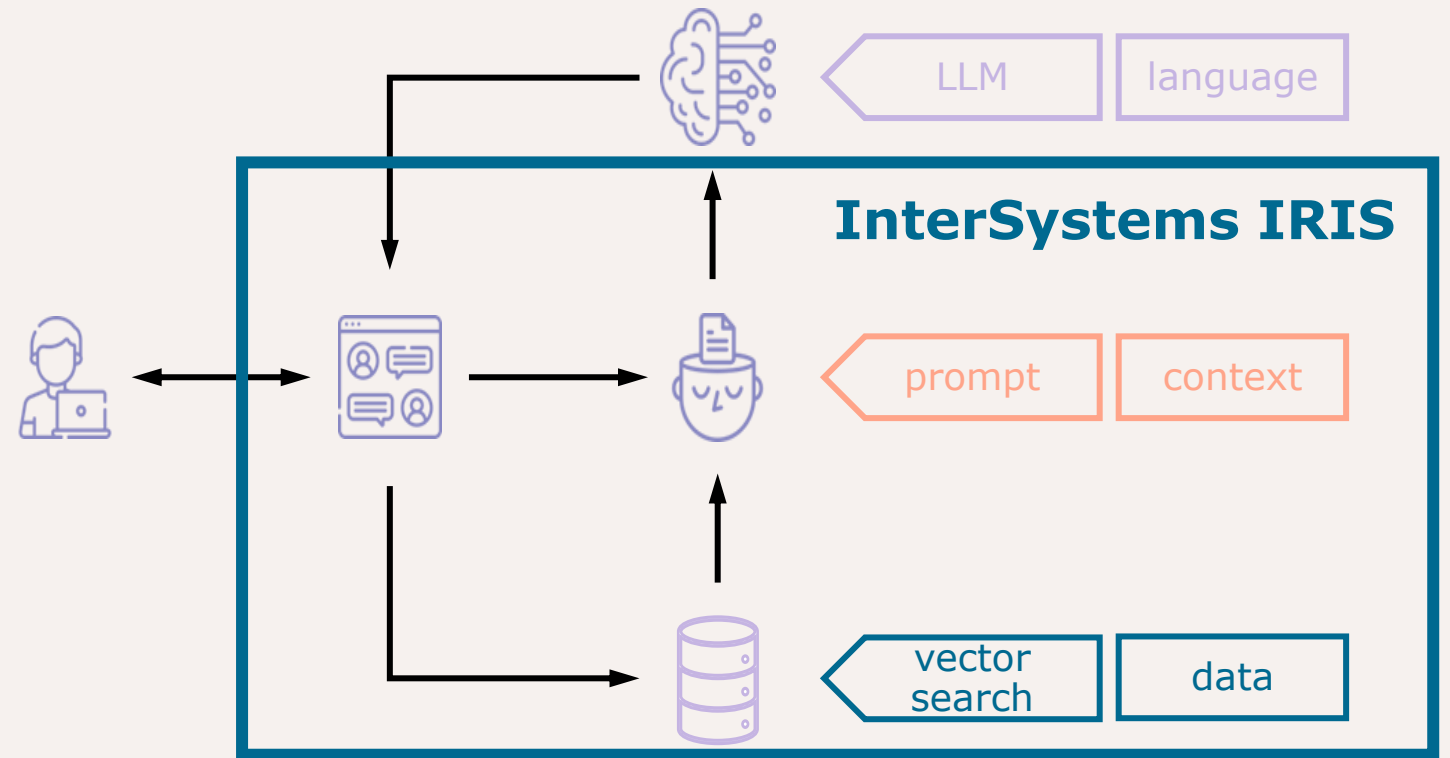
14:00 – 14:30

Die Session zeigt, für welche Use-Cases VectorSearch relevant ist und wie man sie bestehenden Anwendungen hinzufügen kann.

Benjamin Kiwitz | InterSystems



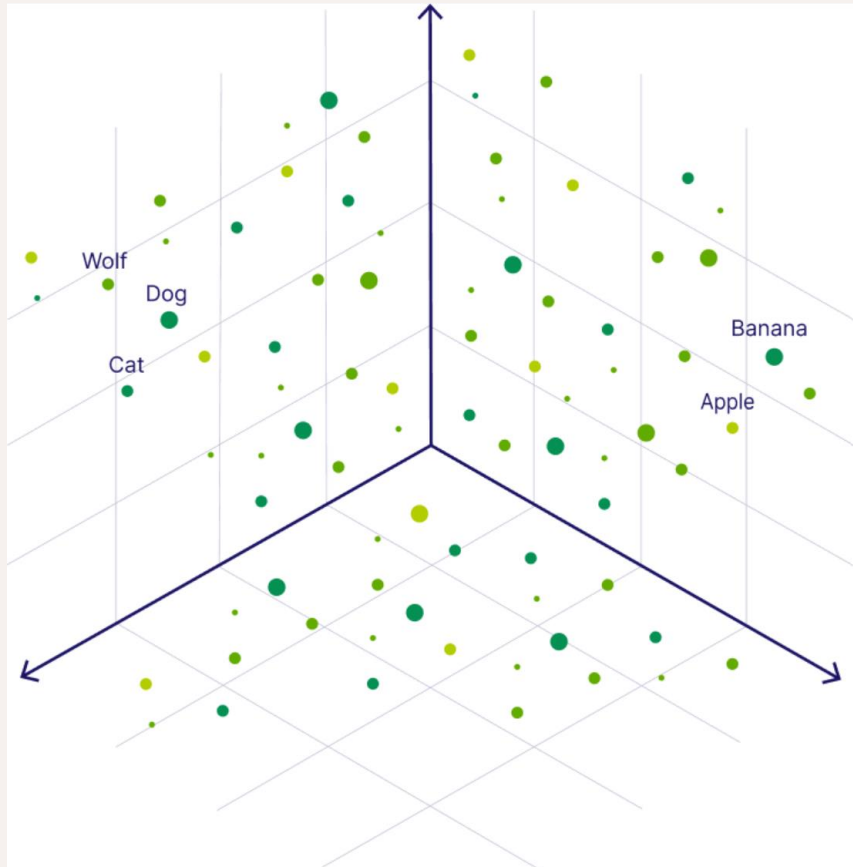
Retrieval Augmented Generation



Retrieval Augmented Generation



Beispiel-LLM ist ein Lexikon nur mit begrenztem Wissen zur Pflanzen- und Tierwelt

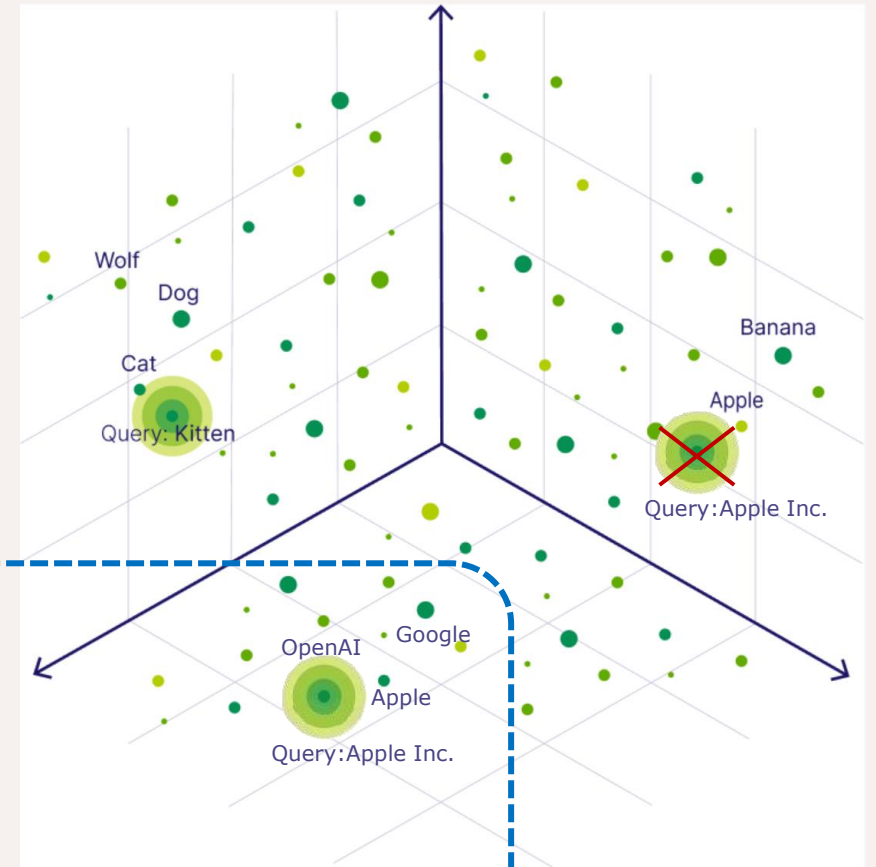


Vector Search
"Kitten"

Vector Search
"Apple Inc."
?

Augmentation
with company
data

Retrieval Augmented Generation



Retrieval Augmented Generation



Status von Large Language Modellen

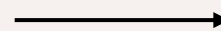


LLMs werden anhand einer Vielfalt öffentlicher Quellen trainiert



Generelles Wissen, breit und nicht allzu tief

LLMs fehlt oftmals spezialisiertes Domain-Wissen



Unvollständige, falsche oder halluzinatorische Antworten

LLMs haben Wissens-Cutoff



Veraltetes Wissen und damit Unfähigkeit zu aktuellen Ereignissen zu antworten

LLMs greifen auf keine externen Datenquellen zu



Beschränkt die Möglichkeit fehlendes Wissen abzugreifen

Retrieval Augmented Generation (RAG)



LLMs haben keinen Zugriff auf private Unternehmens-Daten und spezialisiertes Domain- und/oder Business-Wissen



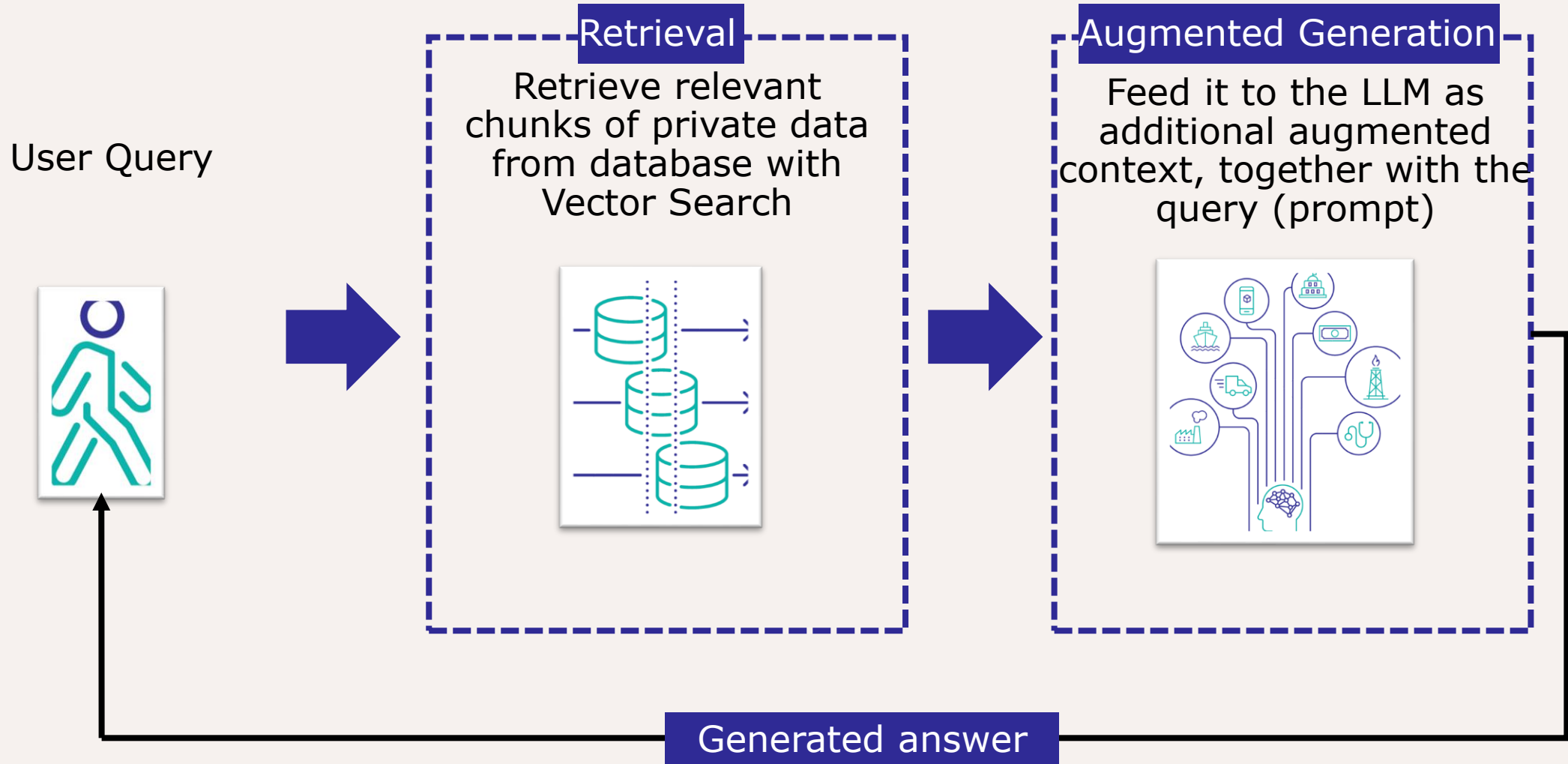
Was tun?

Fine Tuning von bestehenden LLMs durch Anreicherung von zusätzlichem Wissen

Retrieval Augmented Generation



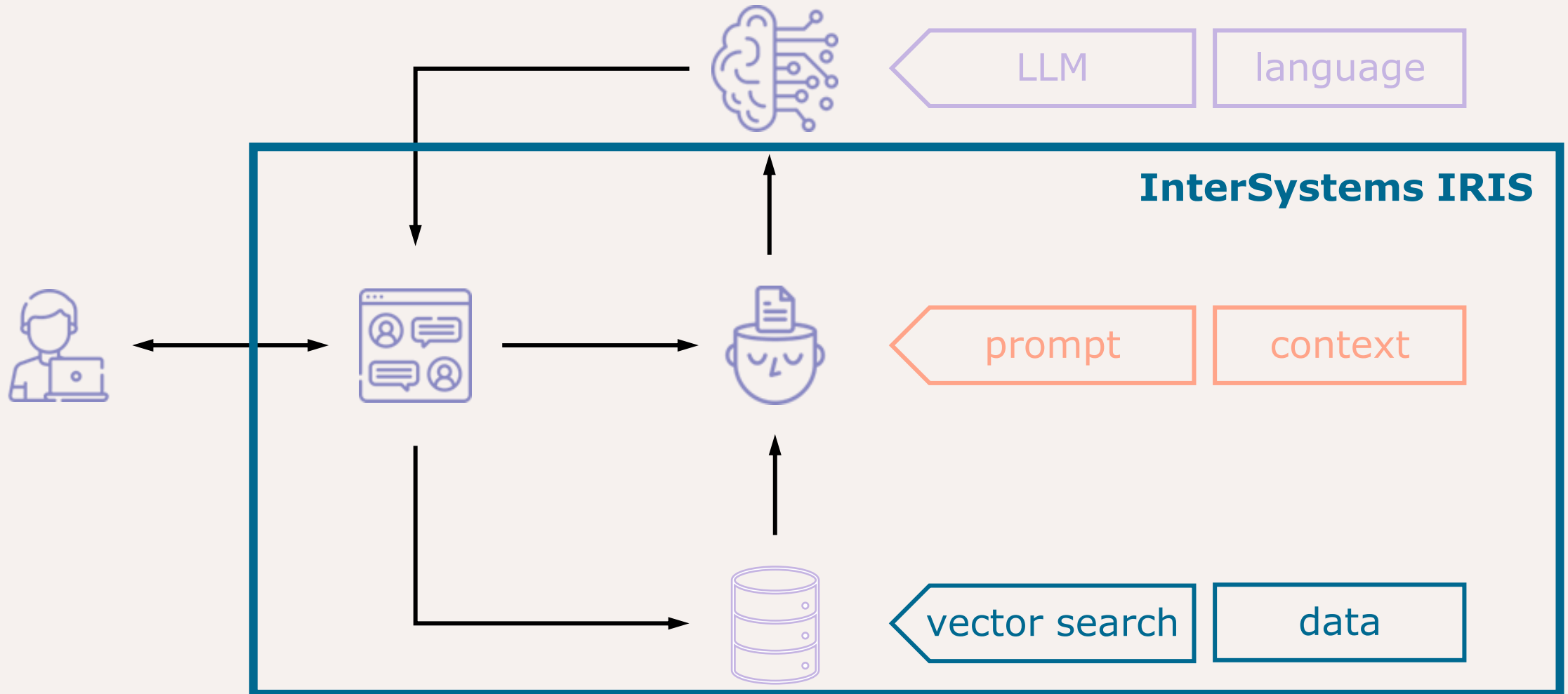
Retrieval (Abruf) Augmented (Anreicherung) Generation (Antworterstellung)



Retrieval Augmented Generation



InterSystems IRIS orchestration with RAG



Orchestrierung von KI mit InterSystems IRIS



Enterprise Ready

Die Entwicklung generativer KI-Anwendungen im Bereich GenAI mit integriertem Vector Search in InterSystems IRIS bietet folgende Vorteile:

- Volle Kontrolle über die Daten – Multi-Model Persistenz und Suche in InterSystems IRIS
- Umfassende Auditierung und Rückverfolgbarkeit
- Orchestrierung der KI-Prozesse mit InterSystems IRIS Interoperability Produktionen - Erstellen von zusammengesetzten Anwendungen, die sich über mehrere KI-Modelle und KI-Dienste erstrecken, unabhängig davon, wo diese ausgeführt werden
- LLMs lokal ausführen, sodass sensiblen Daten vollständig lokal und sicher bleiben
- LLMs, KI-Dienste und Agenten-Frameworks nach Wahl einbinden
- Ein riesiges Ökosystem an KI-Modelle und KI-Diensten einbinden und nutzen, unabhängig davon, wo sie ausgeführt werden

Retrieval Augmented Generation



Live-Demo

RAG mit Chatbot

RAG mit PDF-Dokumenten

KI-Orchestrierung

IRIS Vector Studio

Viewer Config Creator Search Chat Docs

Connected Disconnect

IRIS RAG and Vector Search Demo

User: Describe in one sentence the company InterSystems

LLM: InterSystems is a technology company that specializes in advanced data management, integration, and analytics solutions for healthcare, finance, and other industries.

User: Describe in one sentence the company Logistics

LLM: Logistic is a technology company that specializes in advanced data management, integration, and analytics solutions for healthcare, finance, and other industries.

User: How many sustainable log...

LLM: Logystic H...

User: How man...

LLM: Logystic H...

User: You:

InterSystems
IRIS Data Platform

Management Portal

Home About Help Logout

Server d293e8150c1d Namespace RAG User _SYSTEM Licensed To InterSystems IRIS Community Instance IRIS

Interoperability > Production Configuration - (Demo.DemoProduction)

Production Configuration

Start Stop Try the new UI

Sort: Name Status Number View: [List Icon] [Grid Icon] [Table Icon]

Production Running

Category: All Legend Production Settings

Services

Processes

Operations

Start PDF Import

Streamlit Service

Injection Process

LLM Process

Embedding Operation

OpenAI Operation

PDF Operation

Prompt Operation

RAG Operation

OpenAI Operation

Settings Queue Log Messages Jobs Actions

Apply Search:

Basic Settings

Enabled

OpenAI

Model

gpt-4o

CustomModel

llama3.2:latest

ChatBotInstructions

You are a helpful assistant.

APIKeyCredentials

OpenAIKey

KI und Coffee – Weitere Sessions am Nachmittag



Interoperabilität in Aktion – die ultimative 3D-Simulation

Raum Globe 2 - 3

13:00 – 13:30

Tauchen Sie gemeinsam mit uns in die Welt der virtuellen Realität ein: Unsere Live-Demo zeigt, wie eine moderne Game Engine mit InterSystems IRIS verbunden wird, um eine Fabrik in Echtzeit zu visualisieren und zu überwachen. Wir simulieren Maschinenausfälle und Predictive-Maintenance-Szenarien – interaktiv sowohl auf einem Tablet als auch in einem VR-Headset für ein vollständig immersives Benutzererlebnis.

Stephan Mohr | InterSystems



VectorSearch – Semantische Suche für bestehende Anwendungen

Raum Globe 2 - 3

14:00 – 14:30

Die Session zeigt, für welche Use-Cases VectorSearch relevant ist und wie man sie bestehenden Anwendungen hinzufügen kann.

Benjamin Kiwitz | InterSystems



ChatFHIR: GenAI trifft FHIR, Agentenbasierter Datenzugriff mit MCP

Raum Globe 1

14:45 – 15:30

FHIR-Daten strukturiert nutzen war gestern - heute sprechen wir mit ihnen. Diese Session stellt den MCP-Server als Brücke zwischen KI und FHIR vor. Ein MCP-Server eröffnet KI-Modellen über standardisierte Schnittstellen den Zugriff auf externe Quellen. Mit dem InterSystems FHIR SQL Builder werden bestehende FHIR-Repositories als Tools verfügbar gemacht. Die Teilnehmer erfahren, wie sich ein Agent erstellen lässt, der FHIR-Daten über MCP abrufen, strukturiert und verständlich nutzbar macht - und wie sich mit wenig Aufwand eigene Agenten bauen lassen. Ein praxisnaher Einstieg in agentenbasierte GenAI-Anwendungen im Gesundheitswesen.

Sylwester Boldt, Shubham Sumalya & Henning Siewert | InterSystems



Fragen?



Vielen Dank

Gerne beantworte ich im Anschluss Ihre Fragen. Bitte sprechen Sie mich an.

