

Wolkenlose KI für zu Hause mit eigenen Dokumenten

Thomas Aglassinger, 26.04.2025



Über mich

Thomas Aglassinger

- Software-Entwickler
 - Selbstständig: Siisurit
 - Teilzeit: Providens Analytics
- Für verschiedene Branchen
- Schwerpunkt Daten und Prozesse
- Kein KI-Experte, ein Werkzeug von vielen



<https://providens.at/>



<https://siisurit.com>



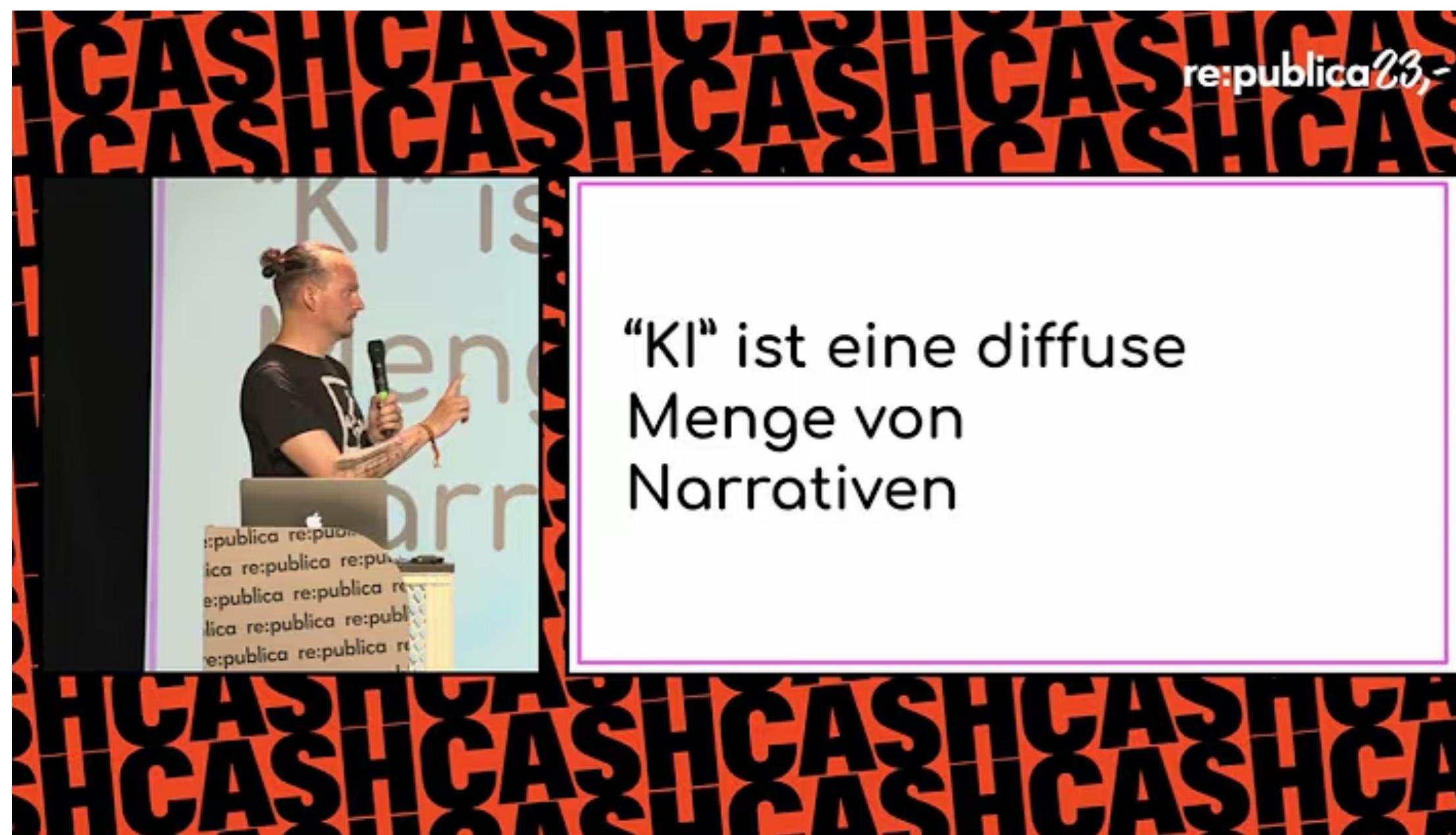
<https://roskakori.at>

**Künstliche Intelligenz,
ein Hype-Thema**

Kritische Aspekte

- Nützliche Aspekte

re:publica 2023: tante - I'm sorry HAL, I won't let you do that.



<https://youtu.be/3LlvHF-IX9Y>

Where AI Meets Code • Michael Feathers • GOTO 2024



<https://youtu.be/g9m3R0NMJ1Y>

**Wozu wolkenlose KI zu Hause
mit eigenen Dokumenten?**

Eigenermächtigung

- Dinge selbst tun,
- mit Software und Modellen, die ich wähle,
- mit meinen Geräte vor Ort,
- die mir gehören,
- statt Ressourcen mieten, die jemandem anderen gehören.

Datenhoheit

- Meine Daten liegen dort, wo ich sie will.
 - Zu Hause.
 - Server-Kammerl eines KMU.
 - Cloud-Dienstleister in Region/Land/EU.
- Ich kann entscheiden, wer die Daten wie verwendet.
 - "Meins meins meins!"
 - Individuelle Verträge mit Interessierten.
 - Keine "nimm oder geh" EULA eines Groß-Konzerns.

Regionale Wertschöpfung

- Große Teile meiner Investition bleibt in Region/Land/EU,
- anstatt Eskapaden von Multi-Milliardären zu finanzieren.

Was ist KI?

Drei Arten von KI

Drei Arten von KI



Drei Arten von KI



Drei Arten von KI





Pony-KI

Technisch: Datenstrukturen und Algorithmen

- Wie ein dressierte Pony im Zirkus
- Folgt Regeln, die ein menschlicher Experte vorgibt
- Beispiele:
 - Expertensysteme
 - Gegner in Videospielen



Pony-KI: Zombie-Shooter

- Wenn das Spiel startet, soll Zombie eine vorgegebenen Route entlang wandern
- Wenn Zombie lautes Geräusch hört, soll er dort hinlaufen
- Wenn Zombie den Spieler sieht, soll er ihn angreifen





Pony-KI: Pfad-Suche für Zombies

Goal: Deliver Robust Behavior Performances

- Reactive Path Following
 - Move towards “look ahead” point farther down path
 - Use local obstacle avoidance
 - Good
 - (Re)pathing is cheap
 - Local avoidance handles small physics props, other bots, corners, etc
 - Superposes well with mob flocking behavior
 - Resultant motion is fluid
 - Bad
 - Can avoid off path too much, requiring repath

A 3D grid-based simulation environment showing a zombie's pathfinding process. A green sphere represents the zombie at the bottom left. A pink arrow points from its current position towards a grey rectangular area representing a wall or obstacle. A series of grey arrows shows the path it has taken, which is curved around the obstacle, demonstrating local avoidance. The environment consists of various grey blocks of different sizes, some solid and some with internal voids, set against a light beige background.

Quelle: "The AI systems of Left 4 Dead", Michael Booth

<https://www.readkong.com/page/slides/the-ai-systems-of-left-4-dead-michael-booth-valve-9664541>

🐵 Affen-KI

Technisch: Maschinelles Lernen

- Affe trainiert in abgeschlossenem Gehege
- Lernt Grundfertigkeiten wie hüpfen, klettern, schwingen
- Lernt Hindernisse zu umgehen und vermeiden
- Wird belohnt, wenn Banane gefunden





Affen-KI

Finde beste Stelle für Banane



🐵 Affen-KI

Wenn gefunden: Belohnung



🐵 Affen-KI

Problem: Steckt in lokalem Maximum obwohl es mehr gäbe



🐵 Affen-KI

Problem: Weiss nicht, dass Äpfel gesünder sind



🐵 Affen-KI

Verschiedene Arten von Affen



Überwachtes Lernen



Unüberwachtes Lernen



Bestärkendes Lernen



Papagei-KI

Technisch: Generative KI

- Kann Gespräche zuhören
- Plappert nach, was es gehört hat
- Derzeit das, was **landläufig** als KI gesehen wird



Papagei-KI: Nächstes Wort mit höchster Wahrscheinlichkeit

<https://moebio.com/mind/>



**Wolkenloser Papagei-KI für zu
Hause**

Ollama

- Installation: <https://ollama.com/>
- Für macOS, Linux, Windows
- Einfache Terminal-Anwendung
- Bietet ein REST-API über HTTP (mehr dazu später)

Sprachmodelle laden

- Liste: <https://ollama.com/search>

- Gute Startpunkte:

- gemma3:4b

- phi4:14b

- deepseek-r1:14b

- Installation zum Beispiel mit:

```
ollama pull gemma3:4b
```

gemma3

The current, most capable model that runs on a single GPU.

vision 1b 4b 12b 27b

3.8M Pulls 21 Tags Updated 6 days ago

qwq

QwQ is the reasoning model of the Qwen series.

tools 32b

1.4M Pulls 8 Tags Updated 6 weeks ago

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

1.5b 7b 8b 14b 32b 70b 671b

39.5M Pulls 29 Tags Updated 2 months ago

mistral-small3.1

Building upon Mistral Small 3, Mistral Small 3.1 (2503) adds state-of-the-art vision understanding and enhances long context capabilities up to 128k tokens without compromising text performance.

vision tools 24b

50.7K Pulls 5 Tags Updated 2 weeks ago

llama3.3

New state of the art 70B model. Llama 3.3 70B offers similar performance compared to the Llama 3.1 405B model.

tools 70b

1.8M Pulls 14 Tags Updated 4 months ago

Demo: Ollama im Terminal

Fragen beantworten

Einzelne Frage:

ollama run gemma3:4b "Welche Farbe hat der Himmel?"

Die Farbe des Himmels ist ein faszinierendes Phänomen, das auf der Streuung des Sonnenlichts durch die Atmosphäre beruht. Hier ist eine detaillierte Erklärung: [...]

Chat mit Folgefragen:

ollama run gemma3:4b

>>> "Welche Farbe hat der Himmel?"

Die Farbe des Himmels ist ein faszinierendes Phänomen, das auf der Streuung des Sonnenlichts durch die Atmosphäre beruht. Hier ist eine detaillierte Erklärung: [...]

>>> Heute war der Himmel grau. Wie kann das sein?

Du hast vollkommen Recht, dass der Himmel grau sein kann, und das ist ein sehr wichtiger Punkt! Die Erklärung, warum der Himmel grau sein kann, [...]

>>>

Bild beschreiben

Nur wenn Modell das Etikett "vision" hat

ollama run gemma3:4b

>>> Beschreibe folgendes Bild: https://de.wikipedia.org/wiki/Blume#/media/Datei:Sonnenblume_1.jpg

Das Bild zeigt eine wunderschöne, große Sonnenblume.
[...]

Insgesamt ist das Bild eine ansprechende Darstellung einer gesunden und prächtigen Sonnenblume. Es vermittelt ein Gefühl von Wärme, Lebensfreude und der Schönheit der Natur.



CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=29603>

Exkurs: Open Source Software & Modelle

- Open-Source-**Software** enthält alles, was ich brauche, um die Software zu Hause selbst zu bauen (Quellcode, Build-Scripts, Linker-Anweisungen, ...)
- Open-Source-**Modelle** bieten nur das Ergebnis, aber nicht die Basisdaten und Anweisungen, die zum Bau verwendet wurden.
 - Entspricht eher Shareware- oder Freeware-Software, welche nur das fertige Programm enthält.
 - Bessere Bezeichnung: **Open-Weight** ("offene Gewichtungen")

Exkurs: Open-weight Modelle

- Derzeit sind viele Modelle gratis verfügbar
- Im Laufe der Zeit veraltet das in ihnen gespeicherte Wissen
- Neuaufbau sehr aufwendig und teuer
(oft in der Größenordnung 10 oder gar 100 Millionen Euro)
- Nur gratis, weil derzeit großer Hype, Konkurrenzkampf zwischen Herstellern, und viele naive Investoren

Exkurs: Open-weight Modelle

- Derzeit sind viele Modelle gratis verfügbar
- Im Laufe der Zeit veraltet das in ihnen gespeicherte Wissen
- Neuaufbau sehr aufwendig und teuer
(oft in der Größenordnung 10 oder gar 100 Millionen Euro)
- Nur gratis, weil derzeit großer Hype, Konkurrenzkampf zwischen Herstellern, und viele naive Investoren
- Kann sich jederzeit ändern → "Enshittification"



I PUT ON MY ROBE

AND WIZARD HAT



Quellcode-Beispiele

- Github-Repo: <https://github.com/roskakori/wolkenlose-ki-fuer-zu-hause>
- Enthält:
 - Jupyter-Notebook mit Beispielen und Erklärungen
 - Docker compose.yaml für Ollama und Datenbank
 - Hinweise zur lokalen Einrichtung in README.md
- Hauptsächlich Python und SQL

Demo: Ollama über REST-API

Eigene Dokument

Beispiel: Gemeindezeitung

- Verfügbar von: <https://www.edelschrott.gv.at/buergerservice/gemeindezeitung/>
- Original liegt als **PDF** vor
- Umgewandelt nach **Markdown**
 - Textformat
 - enthält nur einfache Formatierungen
 - KIs können es gut verarbeiten
- **Kurze Artikel** zu Gemeindethemen wie Kundmachungen, Veranstaltungen, besondere Ereignisse
- Menge: **25 Dokumente** mit insgesamt **458 Artikeln**

Beispiel: Lärmschutzverordnung

Lärmschutzverordnung

Die Verwendung von motorbetriebenen Rasenmähern sowie die Durchführung von vergleichbaren lärmverregenden Arbeiten (Verwenden von Kreissägen, Presslufthämmern und dergl.) ist von

Montag bis Freitag nur in der Zeit von **06:00 Uhr bis 21:00 Uhr**
Samstag nur in der Zeit von **06:00 Uhr bis 18:00 Uhr** gestattet.
An Sonn- und Feiertagen sind diese Arbeiten ganztägig verboten.

Land- und forstwirtschaftliche Tätigkeiten sowie Arbeiten der gewerblichen Gärtnereien und solche der kommunalen Betriebe im Rahmen der Betreuung der öffentlichen Anlagen sind von dieser Regelung ausgenommen.

Lärmschutzverordnung

Die Verwendung von motorbetriebenen Rasenmähern sowie die Durchführung von vergleichbaren lärmverregenden Arbeiten (Verwenden von Kreissägen, Presslufthämmern und dergl.) ist von

Montag bis Freitag nur in der Zeit von **06:00 Uhr bis 21:00 Uhr**
Samstag nur in der Zeit von **06:00 Uhr bis 18:00 Uhr** gestattet. **An Sonn- und Feiertagen sind diese Arbeiten ganztägig verboten.**

Land- und forstwirtschaftliche Tätigkeiten sowie Arbeiten der gewerblichen Gärtnereien und solche der kommunalen Betriebe im Rahmen der Betreuung der öffentlichen Anlagen sind von dieser Regelung ausgenommen.

PDF

Markdown

Demo: Eigene Dokument

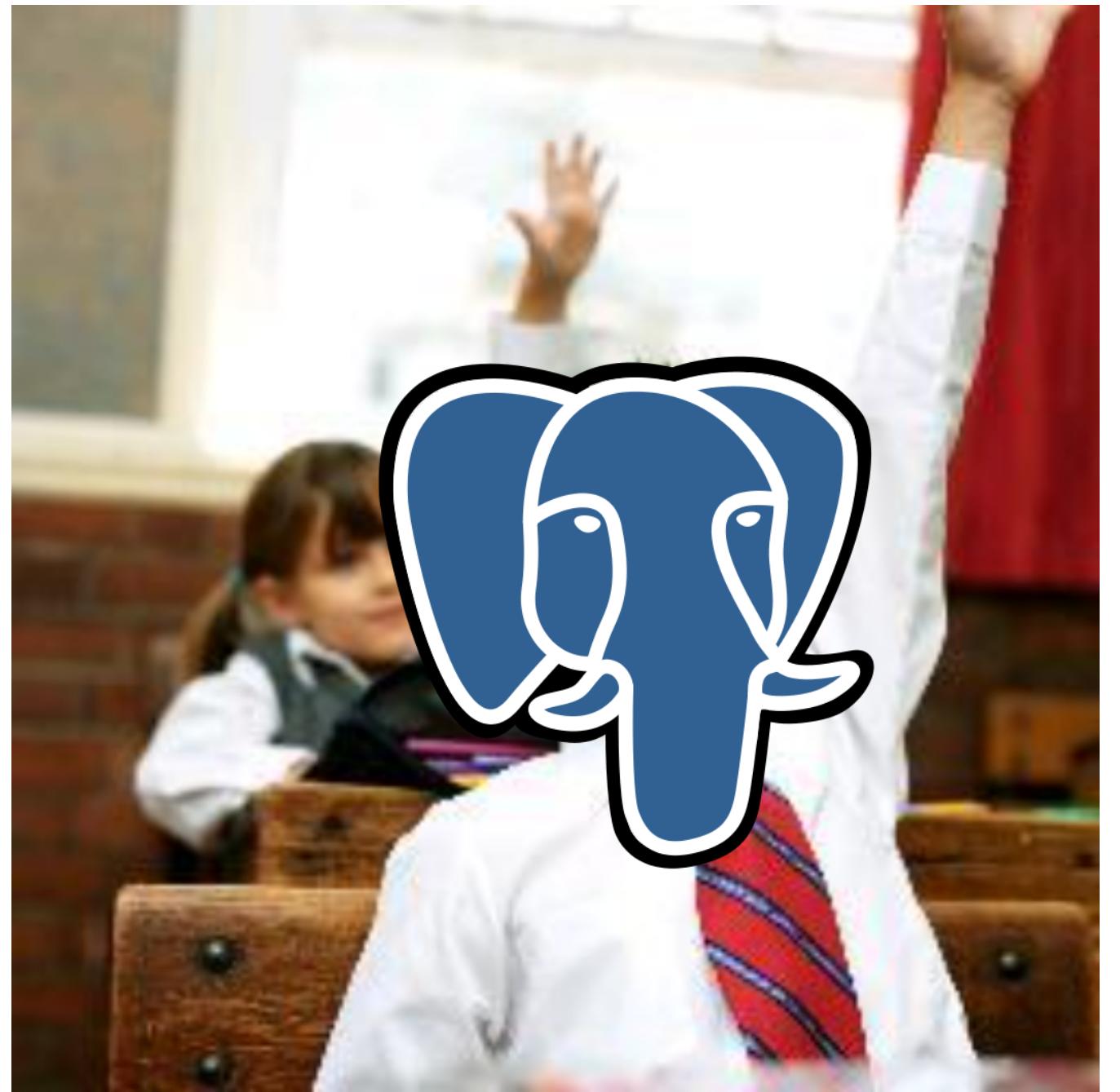
Datenbank mit eigene Dokument

PostgreSQL



@donalshijan5615 1 month ago

Postgress is like that kid whom we thought would have peaked in high school

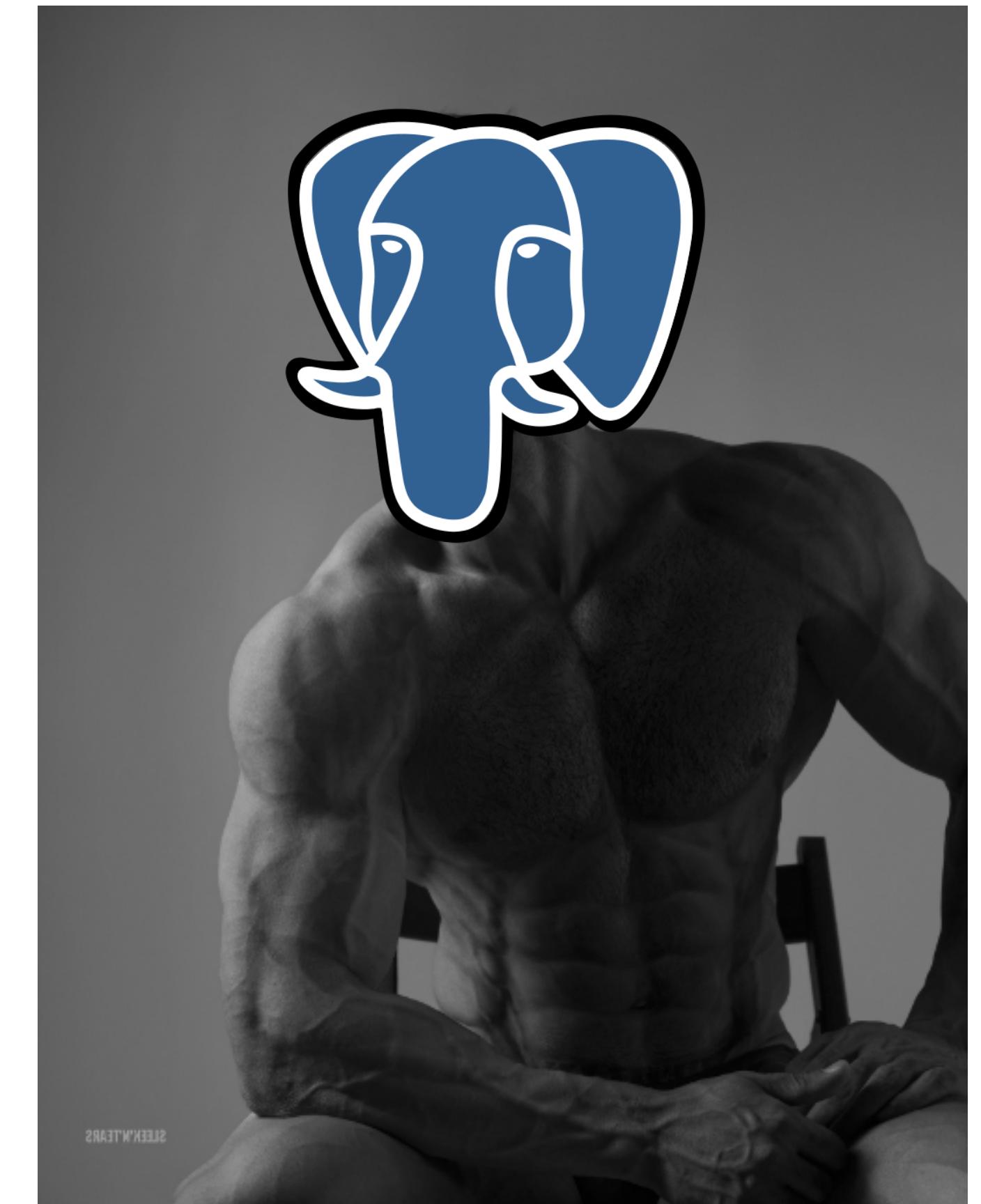
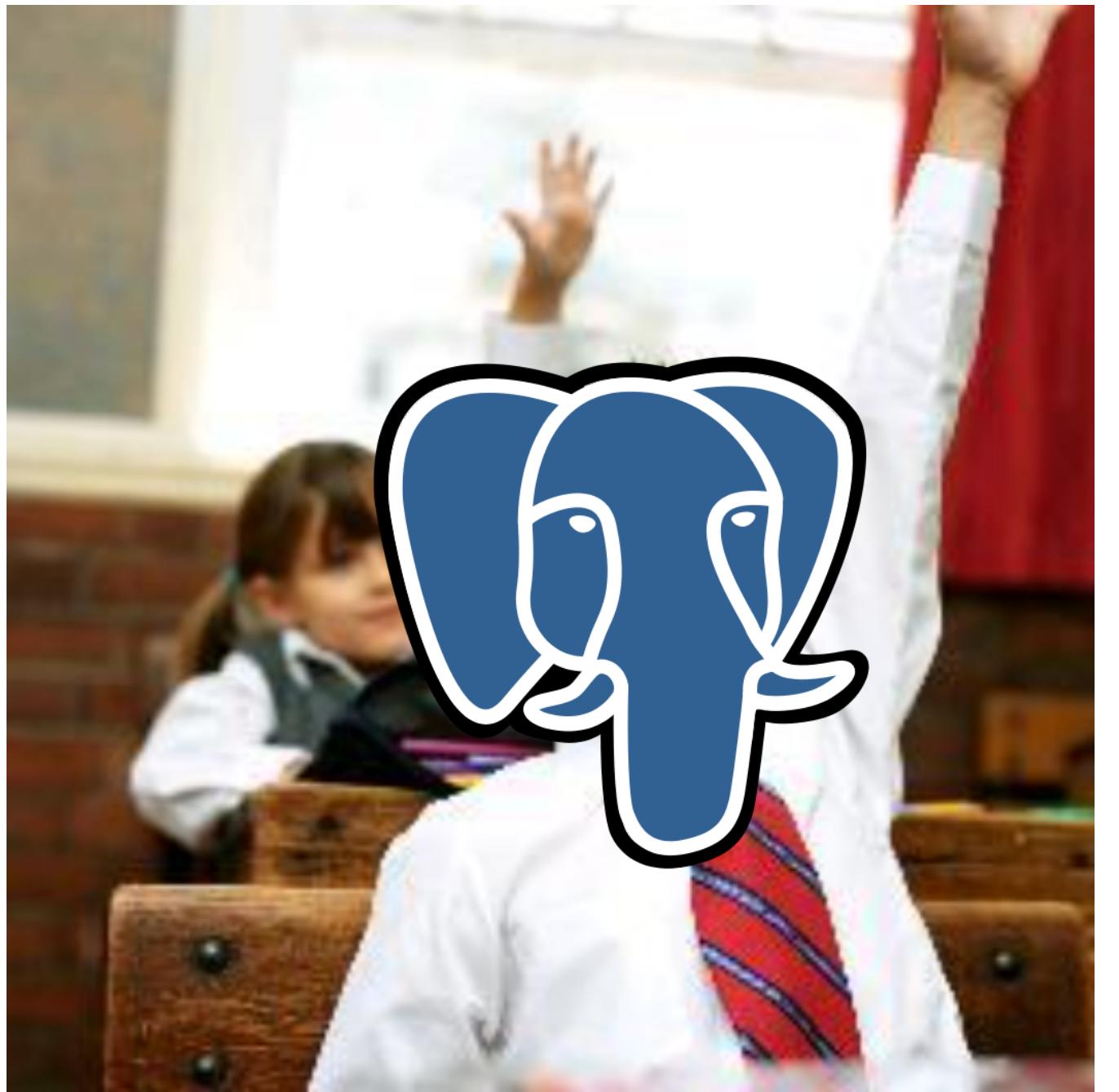


PostgreSQL



@donalshijan5615 1 month ago

Postgress is like that kid whom we thought would have peaked in high school but years later, turns out an absolute looksmaxxed giga chad.

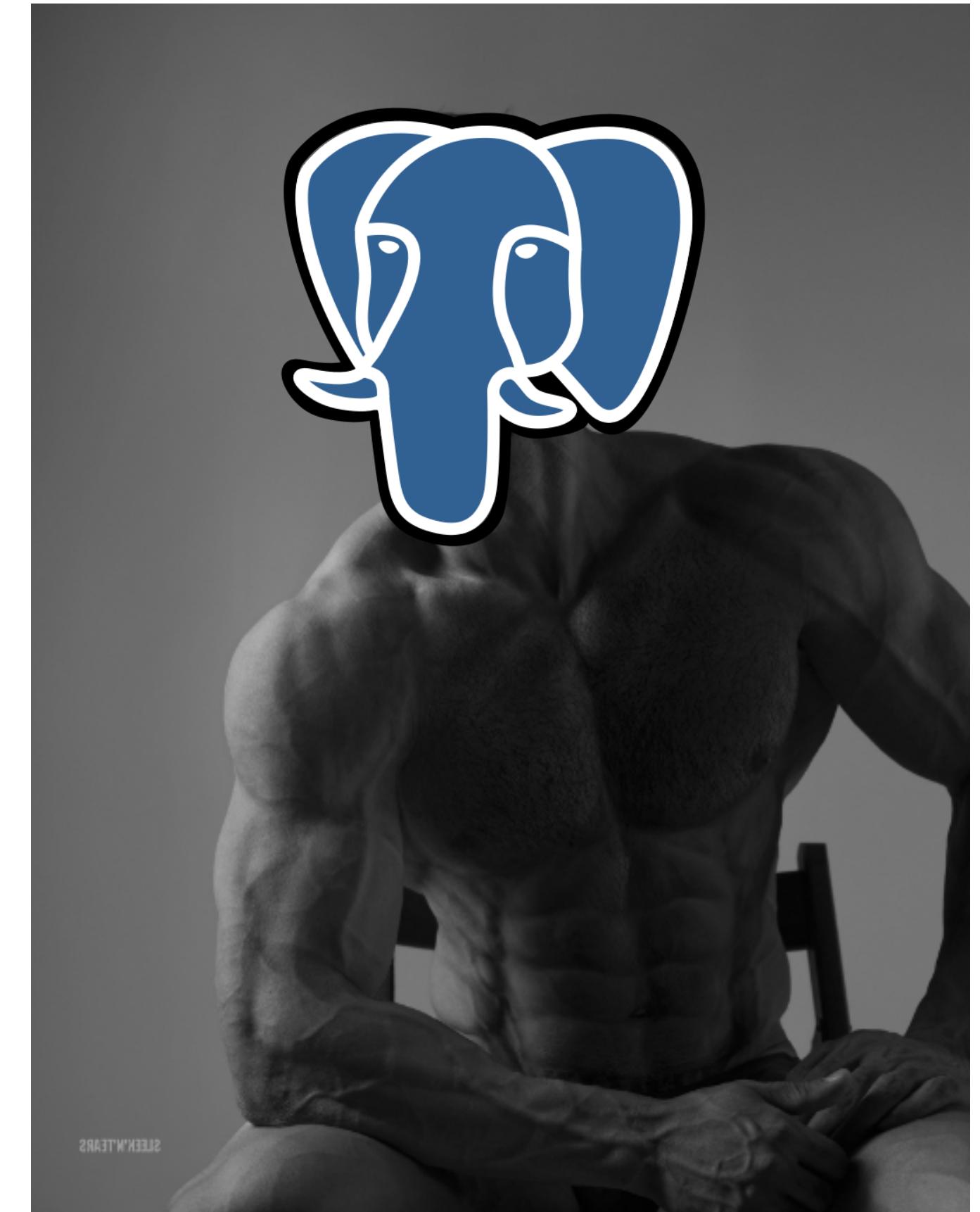
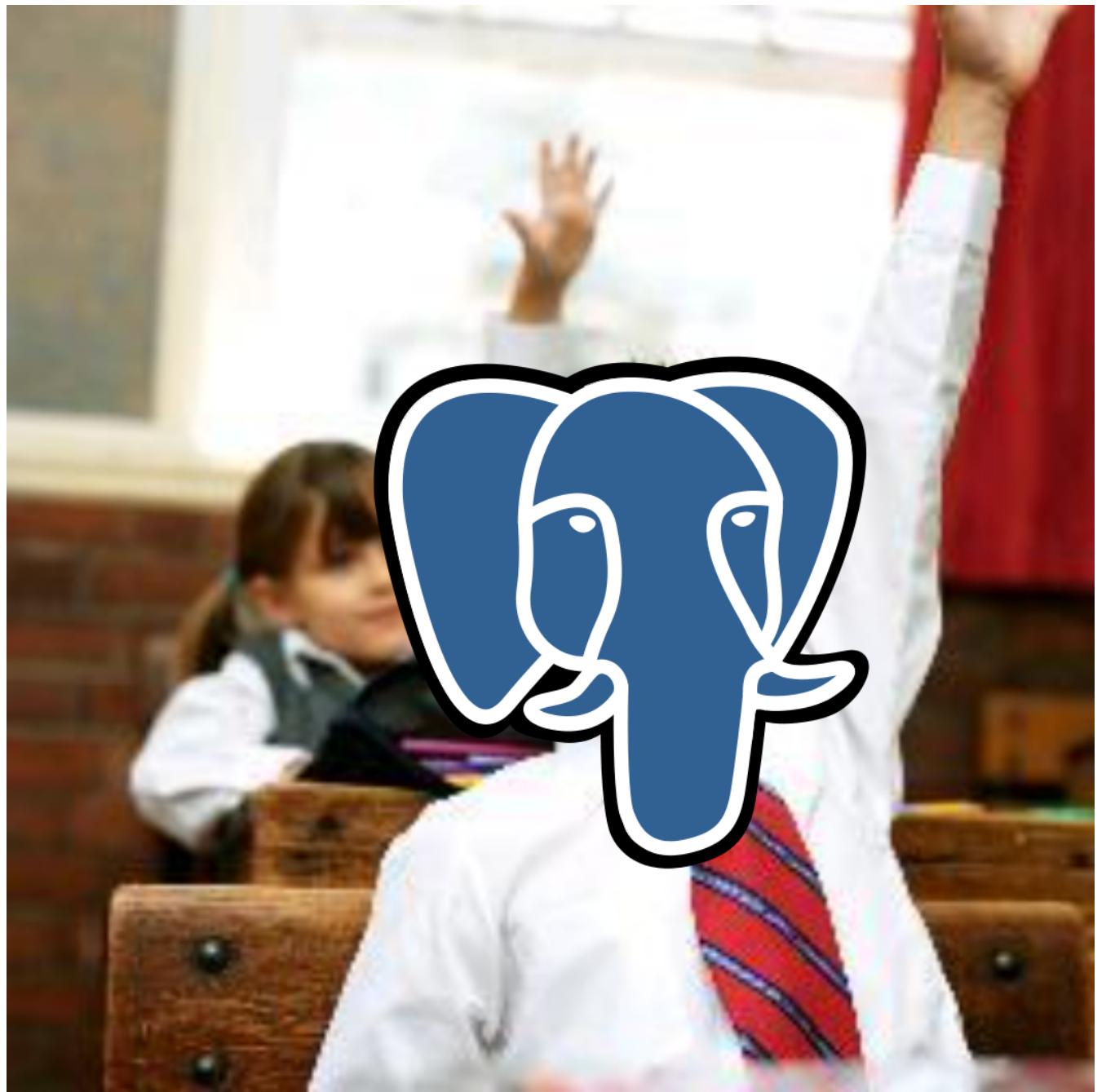


PostgreSQL



@donalshijan5615 1 month ago

Postgress is like that kid whom we thought would have peaked in high school but years later, turns out an absolute looksmaxxed giga chad.



Video: <https://youtu.be/3JW732GrMdg>

**Wolkenloser Papagei-KI mit
eigenen Dokumenten**

Grundprinzip

Bisher:

Anfrage

?



KI



Antwort

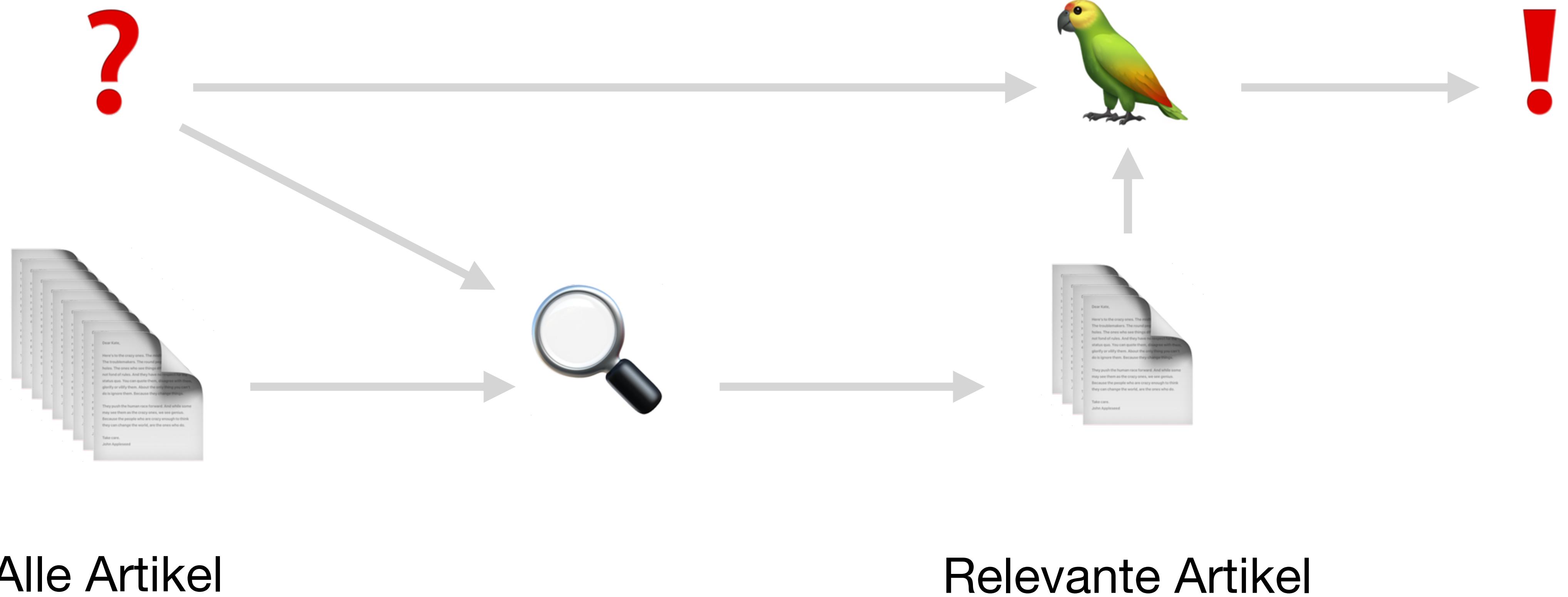
!

Grundprinzip Retrieval-Augmented Generation (RAG)

Anfrage

KI

Antwort



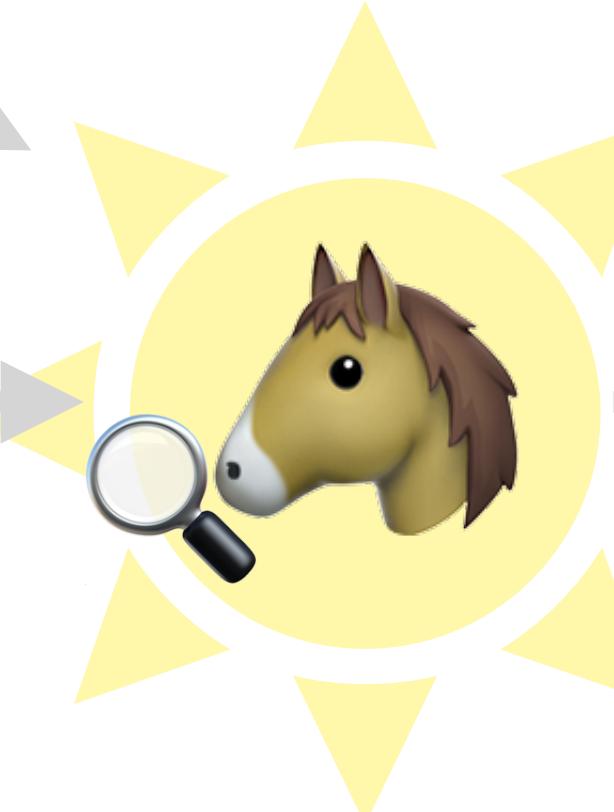
RAG mit Volltextsuche

Anfrage

KI

Antwort

?



!

Alle Artikel

Relevante Artikel

Volltextsuche

Grundprinzipien Volltextsuche

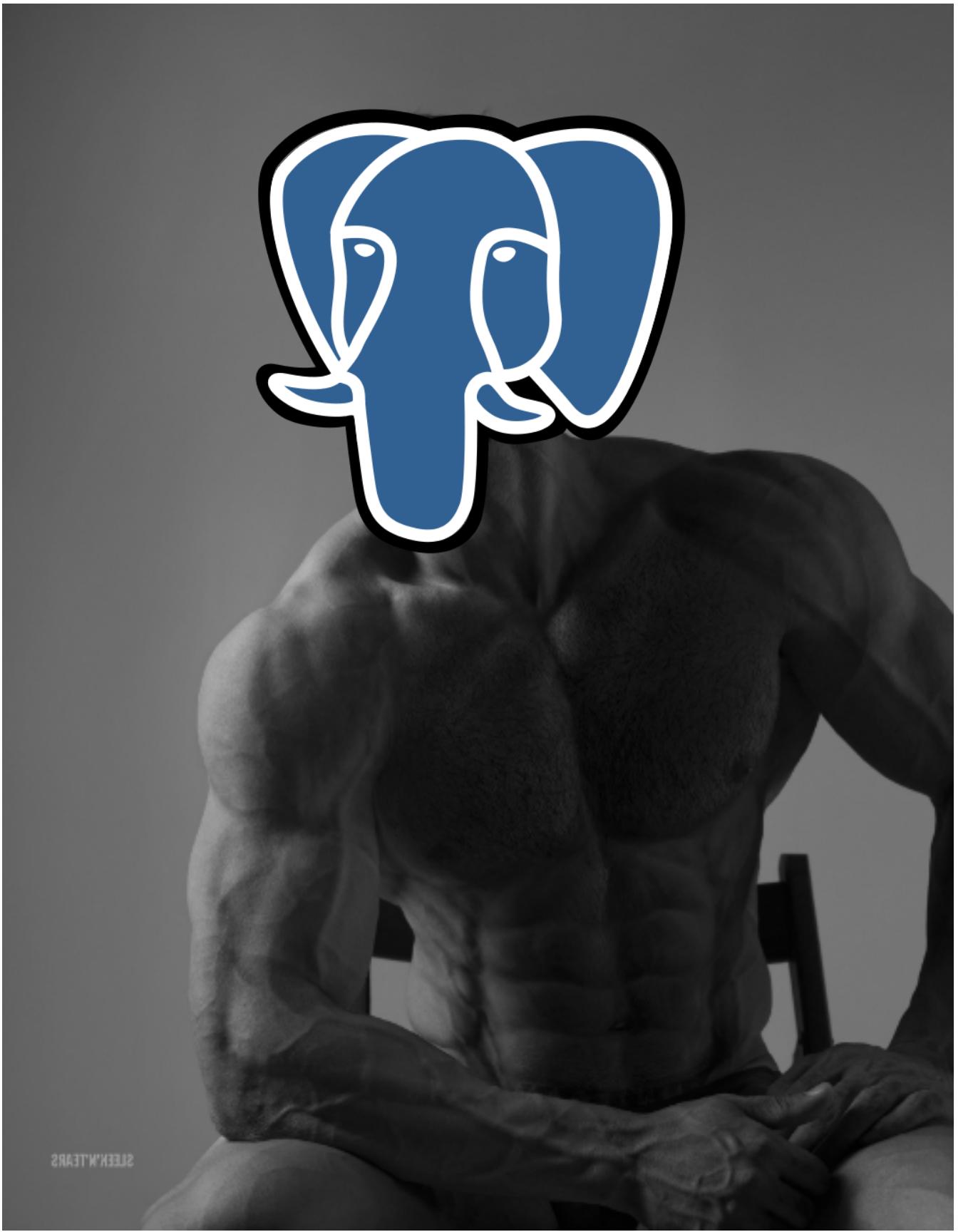
- Sucht nicht nach Texten sondern nach Token
(Token: in etwa: Zahlen, die Wörter darstellen; Beispiel folgt)
- Ignoriert für Suche unwichtige Wörter wie der, die, das, auf, um, wo, ...
(Fachbegriff: Stoppwörter)
- Vereinheitlich Kleinschreibung und Umlaute.
- Arbeitet mit Stammwörtern, z.B. Suche nach "Haus" findet "Häuser"
- Bewertet gefundene Artikel nach Relevanz
- ...und einiges mehr.

Gängige Systeme für Volltextsuche

- Elasticsearch: <https://www.elastic.co/elasticsearch>
- Solr: <https://solr.apache.org/>

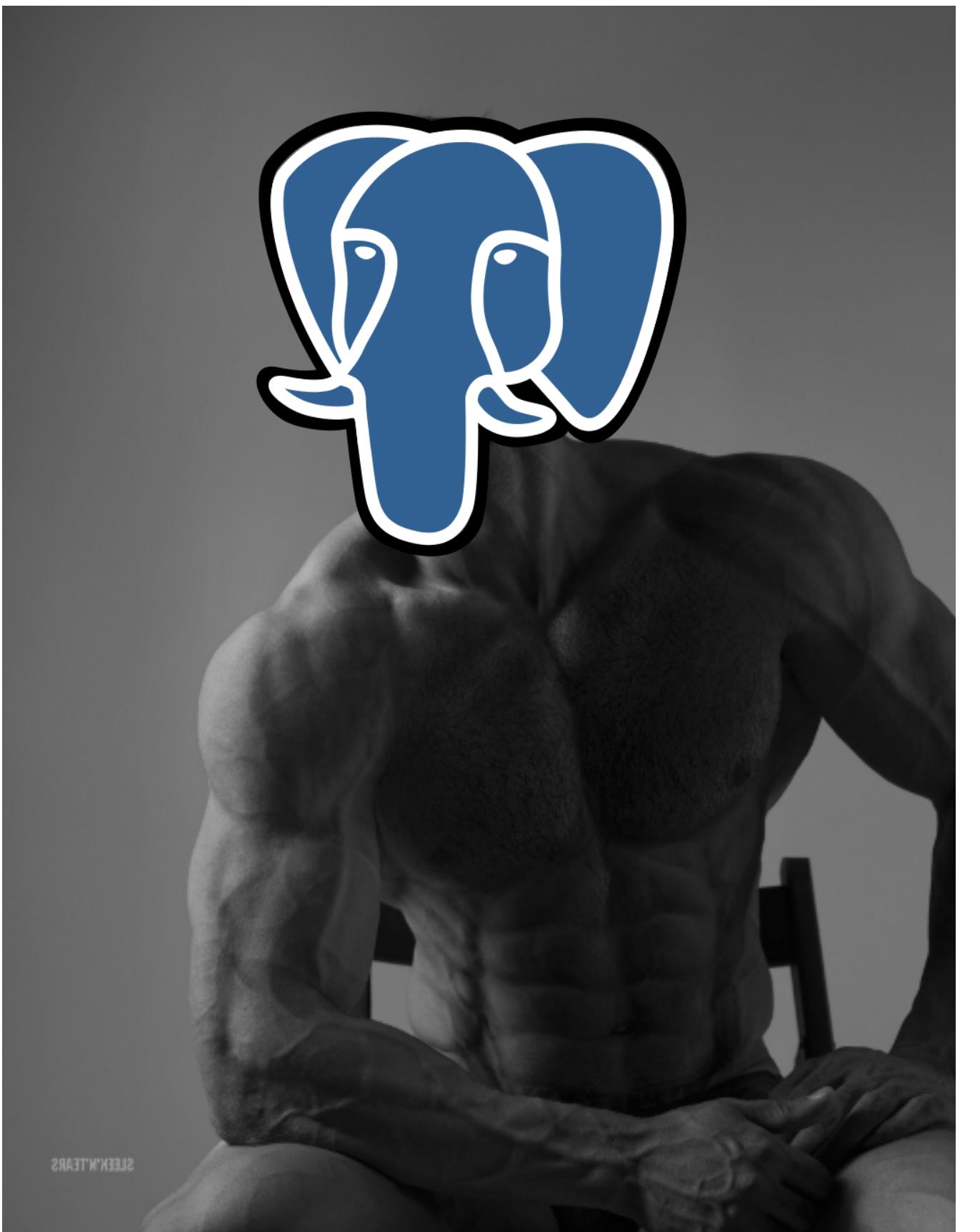
Gängige Systeme für Volltextsuche

- Elasticsearch: <https://www.elastic.co/elasticsearch>
- Solr: <https://solr.apache.org/>
- PostgreSQL 😊



Volltextsuche mit PostgreSQL

- ts_vector
- ts_rank
- to_tsquery
- feld @@ suchausdruck



Demo: Relevante Artikel mit Volltextsuche

Grenzen der Volltextsuche

- Findet nur Wörter, die im Text vorkommen.
- Inhaltlich ähnliche Begriffe werden nicht als solche erkannt.
- Beispiel: Lärmschutzverordnung erwähnt Rasenmäher, Kreissägen und Presslufthämmer, aber nicht Bohrmaschinen

Vektoren und Einbettungen (Embeddings)

Vektorschule

Grundprinzip

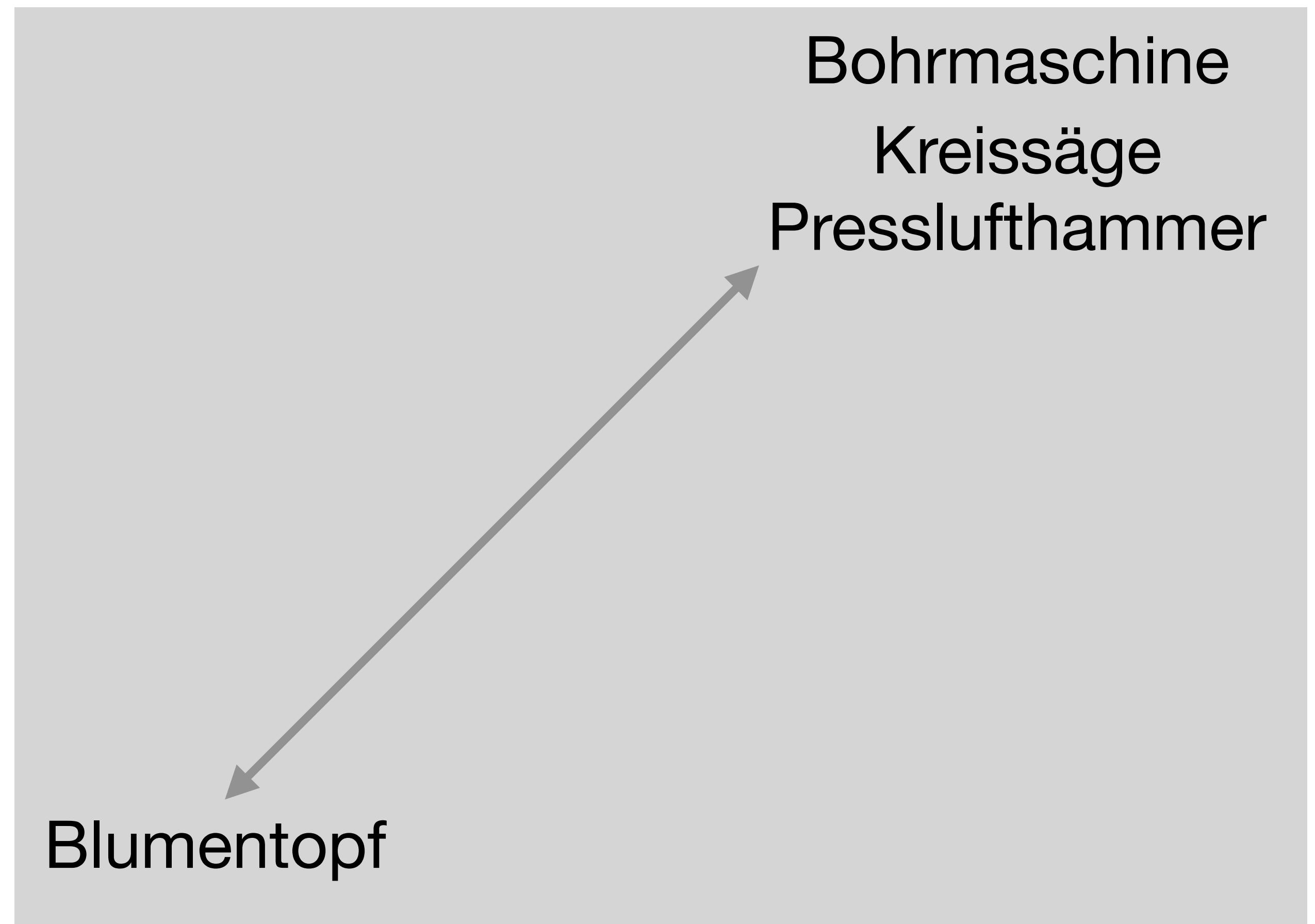
- Begriffe sind mehrdimensional angeordnet.
- Inhaltlich ähnliche Begriffe sind näher zusammen.
- Beispiel für 2-dimensionalen Vektor.

Bohrmaschine
Kreissäge
Presslufthammer

Vektorschule

Grundprinzip

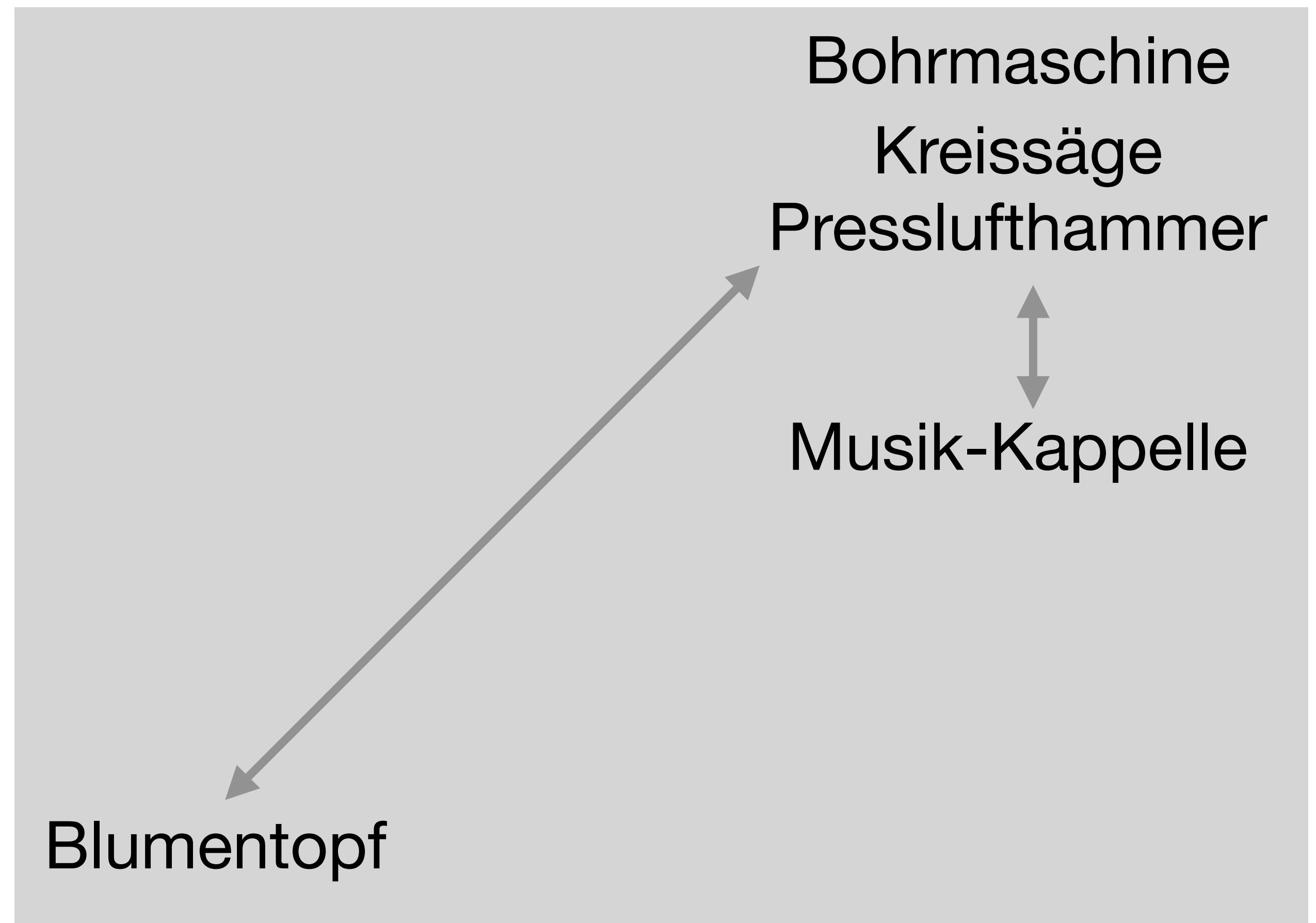
- Begriffe sind mehrdimensional angeordnet.
- Inhaltlich ähnliche Begriffe sind näher zusammen.
- Beispiel für 2-dimensionalen Vektor.



Vektorsuche

Grundprinzip

- Begriffe sind mehrdimensional angeordnet.
- Inhaltlich ähnliche Begriffe sind näher zusammen.
- Beispiel für 2-dimensionalen Vektor.



RAG mit Vektorschreibe

Anfrage

KI

Antwort

?



!

Alle Artikel

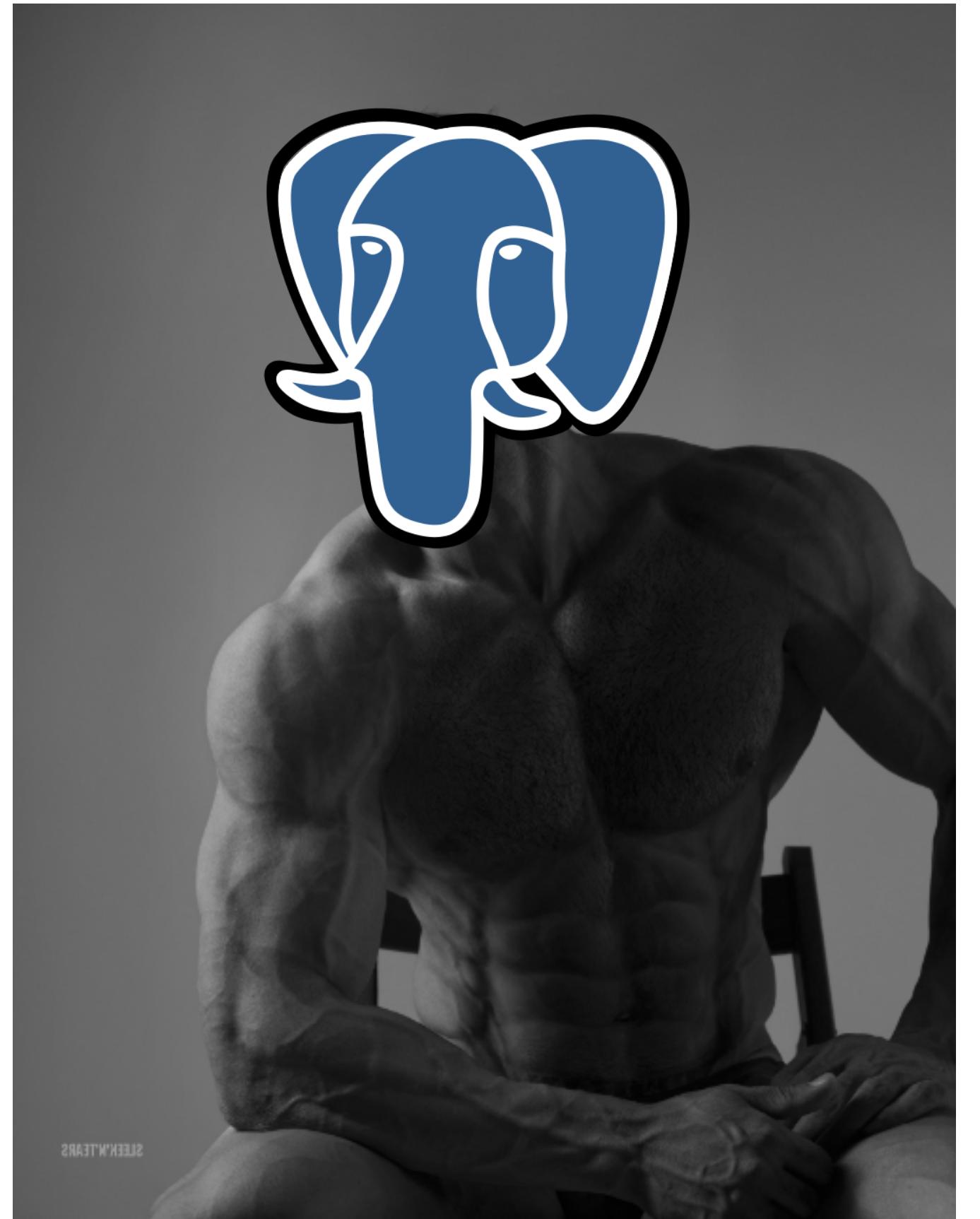
Relevante Artikel

Gängige Systeme für Vektorschre

- ChromaDB: <https://github.com/chroma-core/chroma>

Gängige Systeme für Vektorschre

- ChromaDB: <https://github.com/chroma-core/chroma>
- PostgreSQL 🎉



Demo: Einbettungen (embeddings)

Was kommt als nächstes?

LangChain

<https://github.com/langchain-ai/langchain>

- Bibliothek mit "schönen Funktionen" statt wildes REST und SQL
- Abstraktionen verstecken zwar Komplexität, machen aber auch Verständnis schwieriger
- Flexible Strukturen für:
 - Verkettung von Arbeitsschritten
 - Prompts mit Templates, Embeddings, Dateiformate
 - Unterstützt verschiedene KI's, nicht nur Ollama
 - Chunking: Aufteilen langer Texte in kleinere Häppchen
 - ...

LangChain

<https://github.com/langchain-ai/langchain>

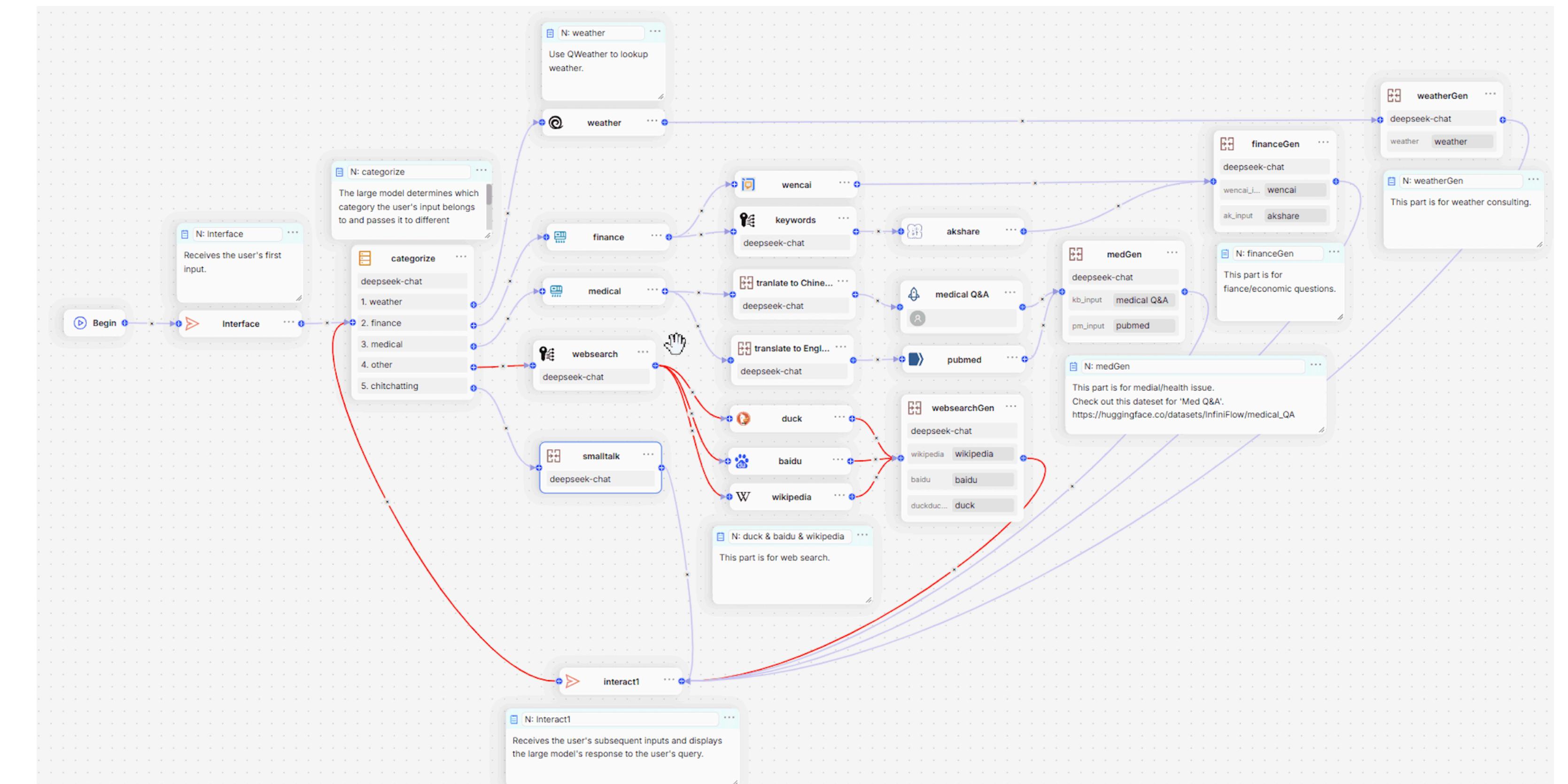
- Bibliothek mit "schönen Funktionen" statt wildes REST und SQL
- Abstraktionen verstecken zwar Komplexität, machen aber auch Verständnis schwieriger
- Flexible Strukturen für:
 - Verkettung von Arbeitsschritten
 - Prompts mit Templates, Embeddings, Dateiformate
 - Unterstützt verschiedene KI's, nicht nur Ollama
 - Chunking: Aufteilen langer Texte in kleinere Häppchen
 - ...



RagFlow

<https://github.com/infiniflow/ragflow>

- Visuelle Darstellung
- Geführte Oberfläche
- Einfachere Benutzbarkeit

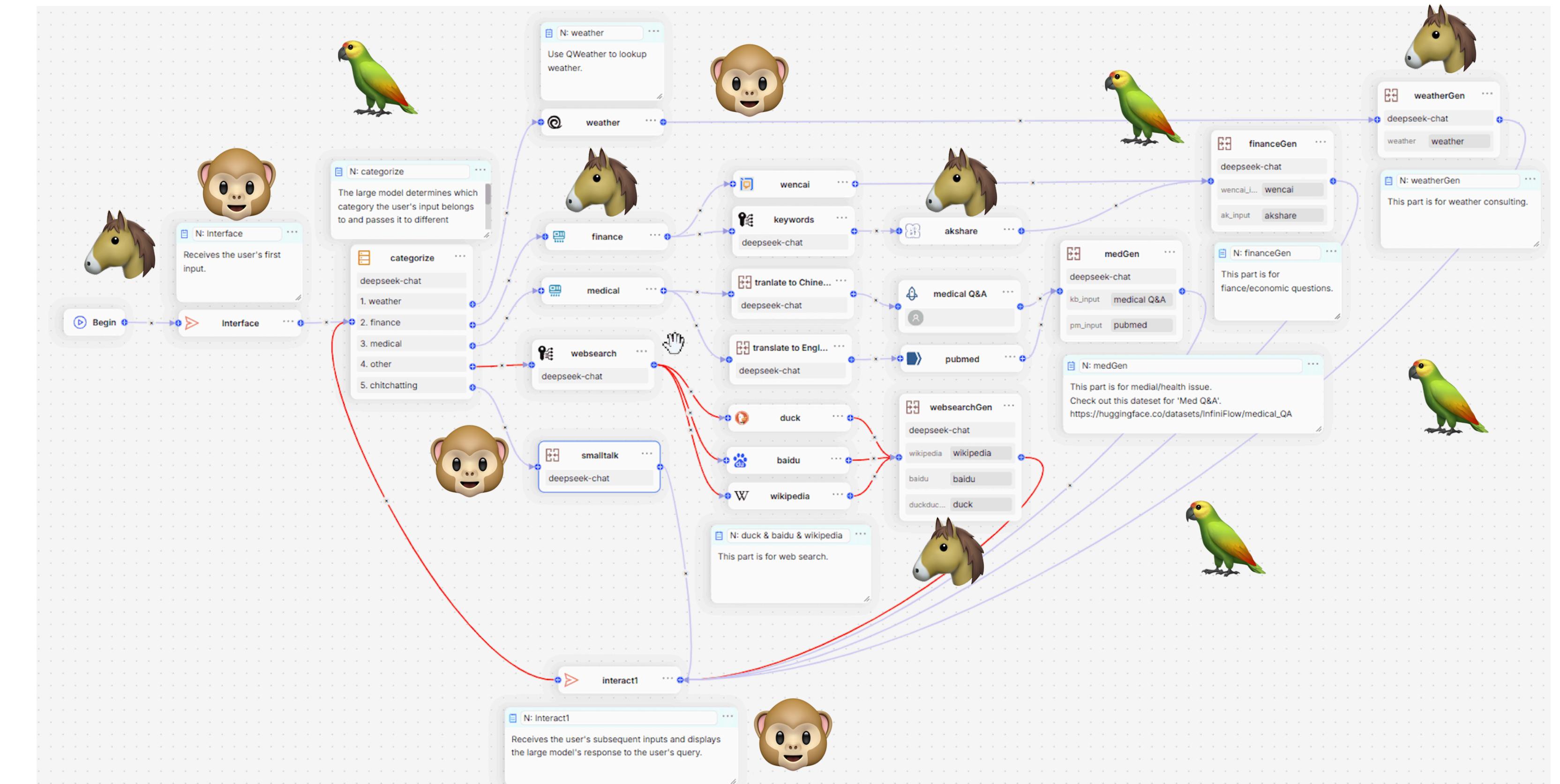


Quelle: <https://github.com/infiniflow/ragflow>

RagFlow

<https://github.com/infiniflow/ragflow>

- Visuelle Darstellung
- Geführte Oberfläche
- Einfachere Benutzbarkeit

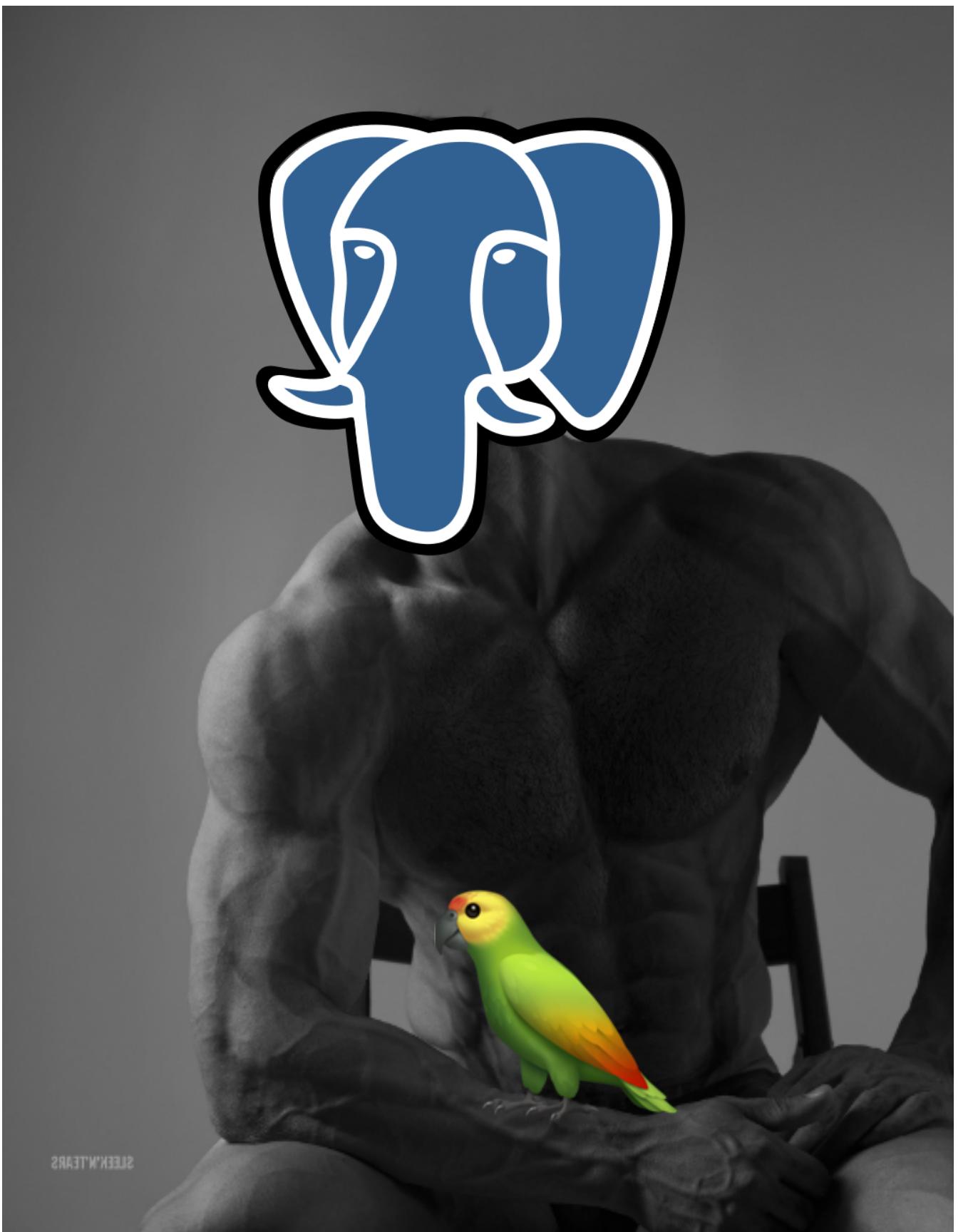


Quelle: <https://github.com/infiniflow/ragflow>

pgai

<https://github.com/timescale/pgai>

- RAG, direkt in SQL
- Direkter Aufruf gängiger KI
- Eingebaute Werkzeuge zum Laden von Daten



Schluss

Schluss

- **Eigenermächtigung:**
 - Grundverständnis
 - Quellcode-Beispiele als Basis für eigene Experimente
- **Datenhoheit:** Alles am eigenen Laptop möglich.
- **Regionale Wertschöpfung:** Liegt jetzt an euch.
 - Firmeninterne Prototypen
 - Nutzung lokale Interessengruppen
 - Diplomarbeiten oder Schulprojekte

Schluss

- **Eigenermächtigung:**
 - Grundverständnis
 - Quellcode-Beispiele als Basis für eigene Experimente
- **Datenhoheit:** Alles am eigenen Laptop möglich.
- **Regionale Wertschöpfung:** Liegt jetzt an euch.
 - Firmeninterne Prototypen
 - Nutzung lokale Interessengruppen
 - Diplomarbeiten oder Schulprojekte

