Workshop
Leibniz Bioactives Cloud

2020-01-29
Leibniz Institut für Pflanzenbiochemie
Halle (Saale)

Inhalt

- Einführung
  - Zielstellung
  - Architektur
  - Status Quo
- Wo wollen wir hin
  - Multi-Cloud
  - CloudLIMS
  - Semantische Textanalyse?

Leibniz Forschungsverbund

https://www.leibniz-wirkstoffe.de

- Diversität der wissenschaftlichen Themen, Organisationsformen und verteilt über Deutschland
- Exzellenz durch Förderung von Kooperation
- Maßnahmen
  - Konferenzen
  - “Seed money”
  - **Leibniz Bioactives Cloud** als Werkzeug für Wissenschaftler

Zweck der Leibniz Bioactives Cloud

- Austausch von nichtöffentlichen Dokumenten (pdf)
- Vernetzung der Wissenschaftler
- (Semantische) Analyse der verfügbaren Dokumente
- Austausch von wissenschaftlichen Daten (chemische Strukturen, Assay-Ergebnisse, …)

Vorhandene Technologie

Existierende Softwareprojekte:

- Web basiert

- Fortschrittliche Technologie

- Verteilte Infrastruktur

- “Chemietauglich”

ckan

LAILAPS
THE PLANT SCIENCE SEARCH ENGINE

ZB MED Search Portal for Life Sciences
LIVIVO

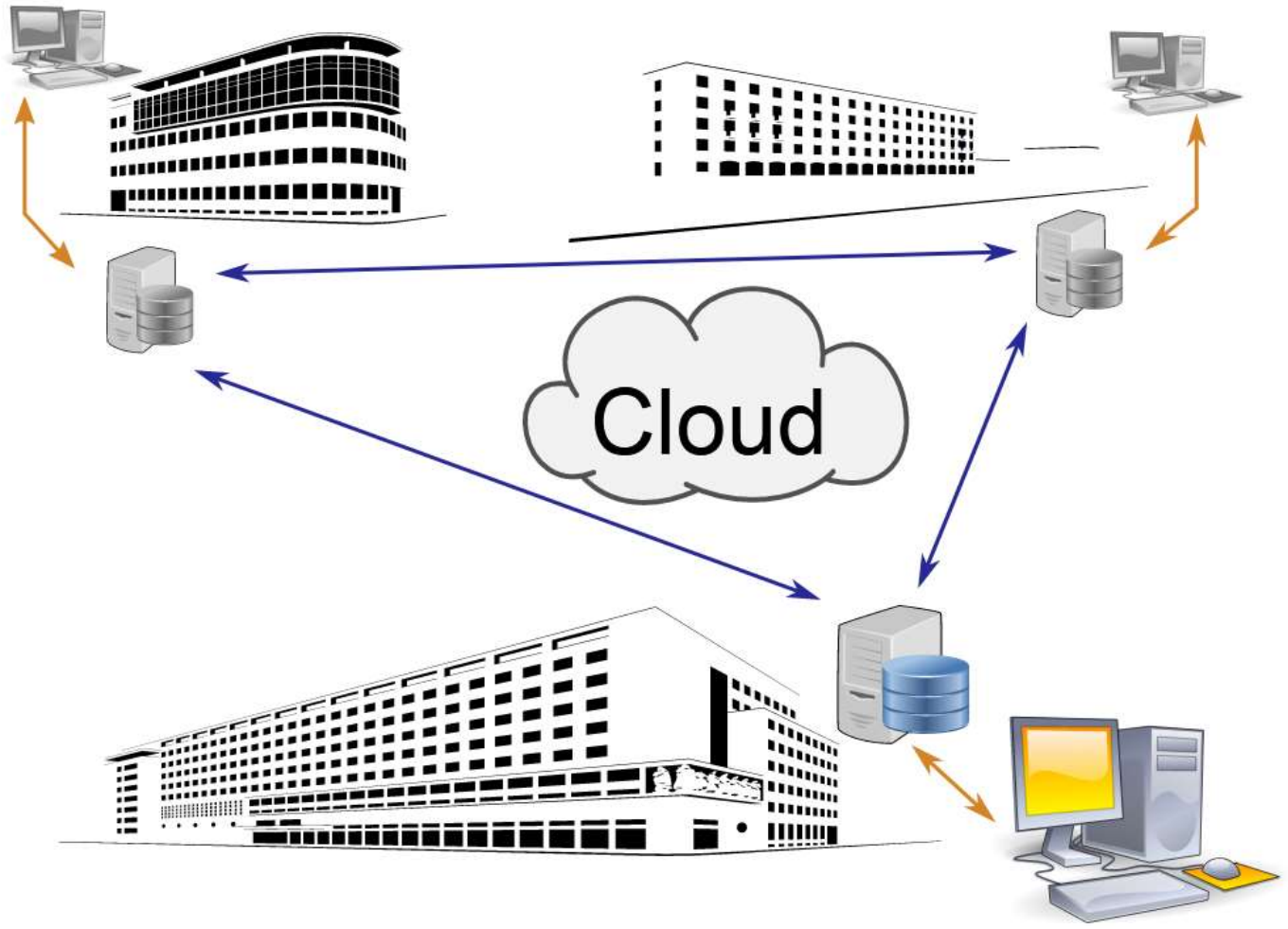
RADAR

eu openscreen

BioSolr

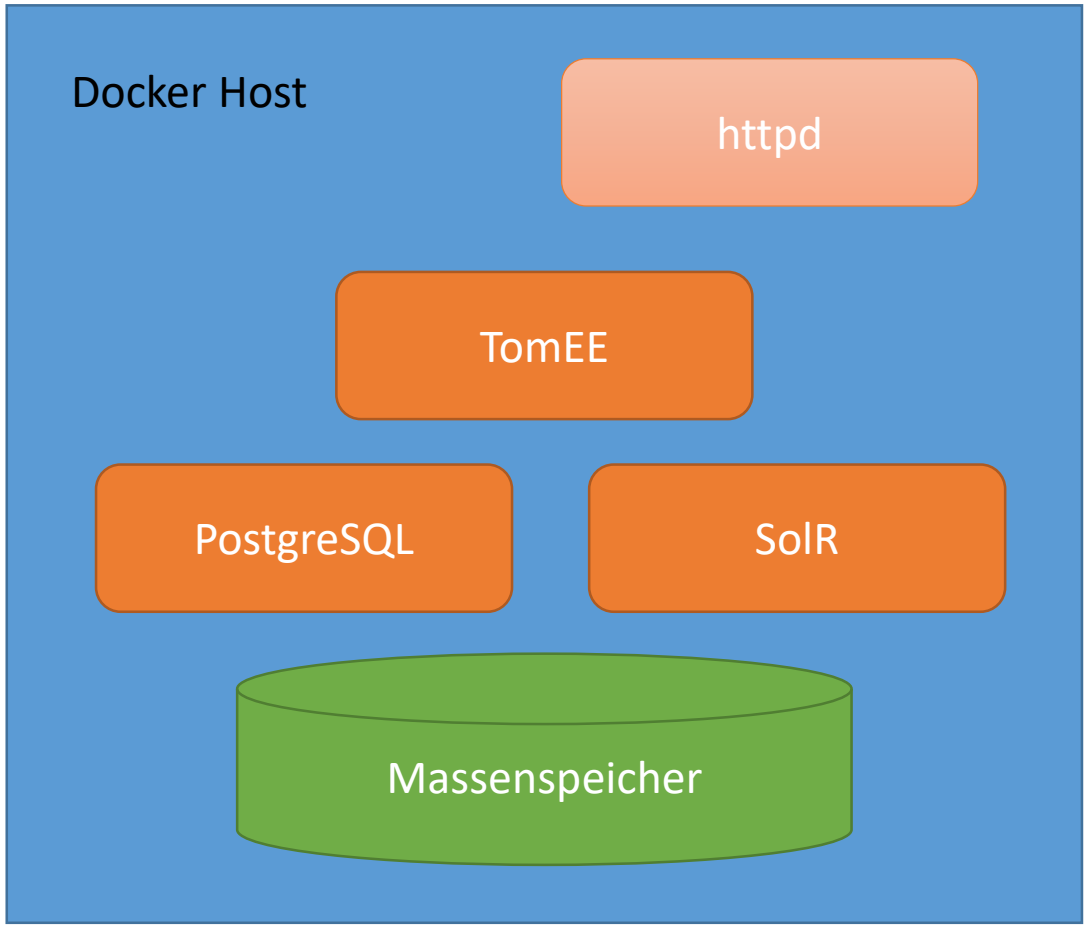
BASE
Bielefeld Academic Search Engine

Verteilte Architektur

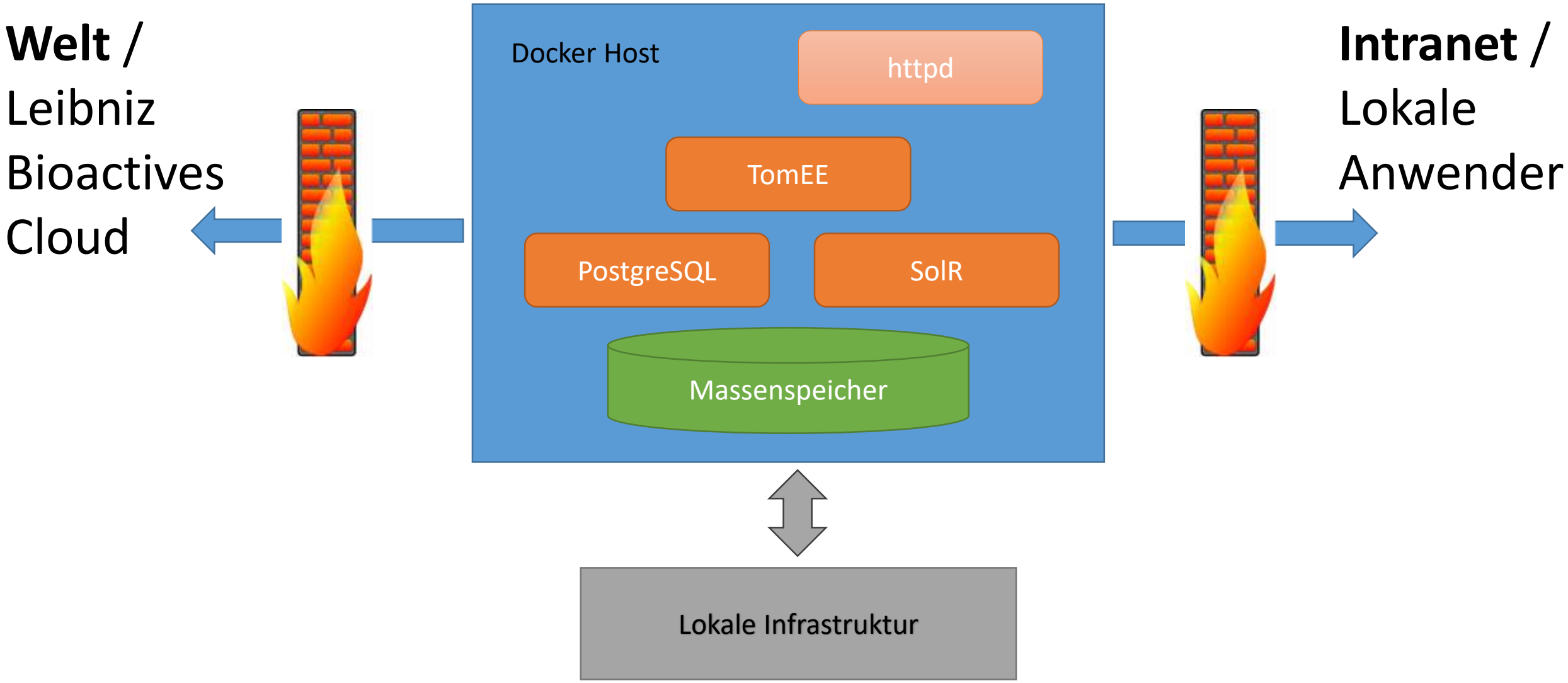


Architektur / Designprinzipien

- Linux / Docker / Java
- Aufbau aus Standardkomponenten
  - Apache TomEE 7
  - PostgreSQL
  - SolR
  - [Apache httpd]
- Geringe Ressourcenanforderungen (Technik **und** Personal)
- Absicherung durch Verschlüsselung
- Vermeidung unnötiger Netzwerkverbindungen



Architektur

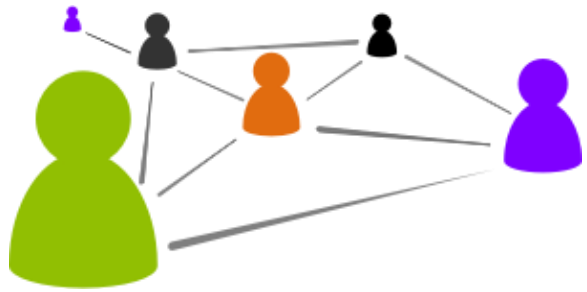


Entwicklung und Verwaltung

- Build-Umgebung (Compiler, IDE, Build-Werkzeuge, …)
- Sourcecode Repository, Wiki, Projektmanagement (BitBucket, Confluence Jira)
- Webserver für Newsletter, Softwareverteilung, Dokumentation („Handbücher“): https://www.leibniz-wirkstoffe.de/
- Mehrere Testsysteme
- Hilfsprogramme für die Ausstellung von Zertifikaten und Zertifikatssperrlisten, Softwareverteilung, …

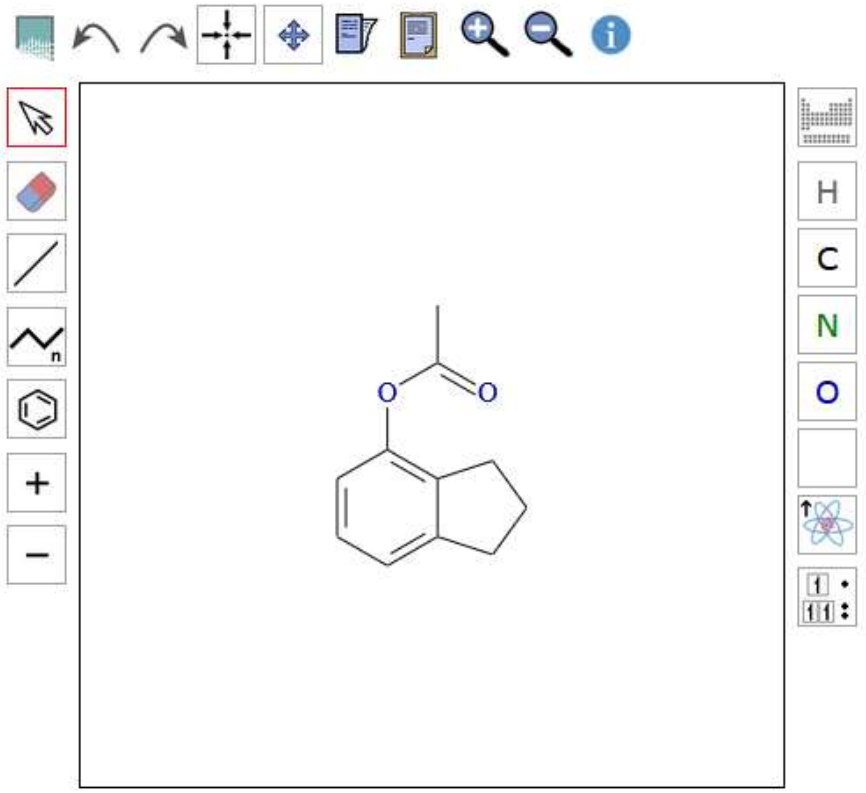
Leibniz Bioactives Cloud: aktueller Stand

- Volltextsuche nach verteilt gespeicherten Dokumenten
- Mehrsprachige Dokumentenprozessierung mit globaler Relevanzmetrik
- Einfaches (verteiltes) Nutzerforum
- Word-Cloud-Suche (hauptsächlich als Technologiestudie)
- Feingranulare Rechteverwaltung mit optionaler LDAP-Anbindung
- Internationalisiertes Nutzerinterface in „responsivem“ Design



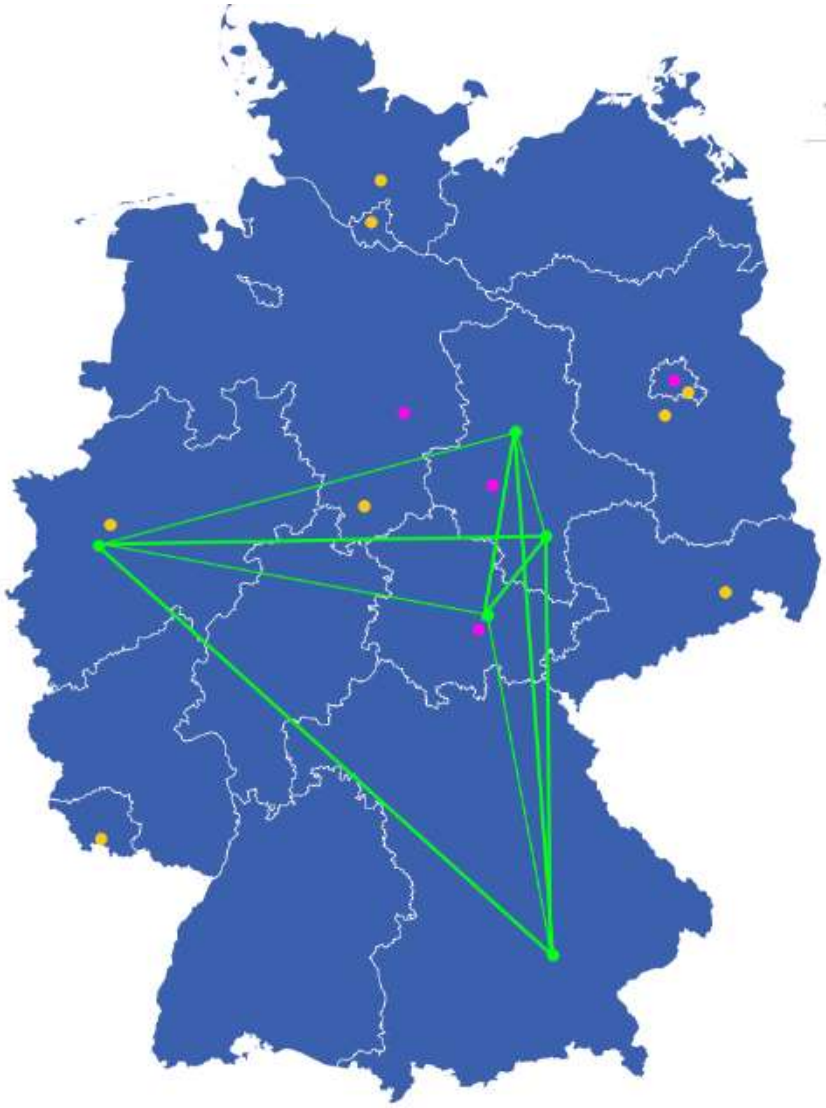
Leibniz Bioactives Cloud: aktueller Stand

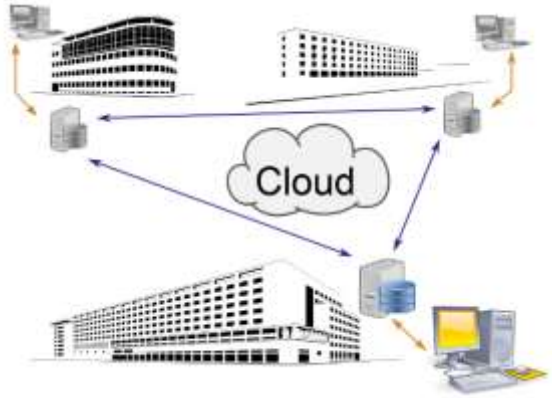
- Rund 70 Unit-Tests zur Qualitätssicherung
  Testabdeckung ca. 40 %
- Weitgehend automatisierte Installation
- Spin-Off: MolPaintJS
  https://github.com/ipb-halle/MolPaintJS



Leibniz Bioactives Cloud: Rollout

- IPB Halle ✓
- HKI Jena ✓
- ISAS Dortmund ✓
- LIN Magdeburg ✓
- Leibniz-LSB@TUM ✓
- IPK Gatersleben 🕐
- FLI Jena 🕐
- FMP Berlin 🕐
- DSMZ 🛑



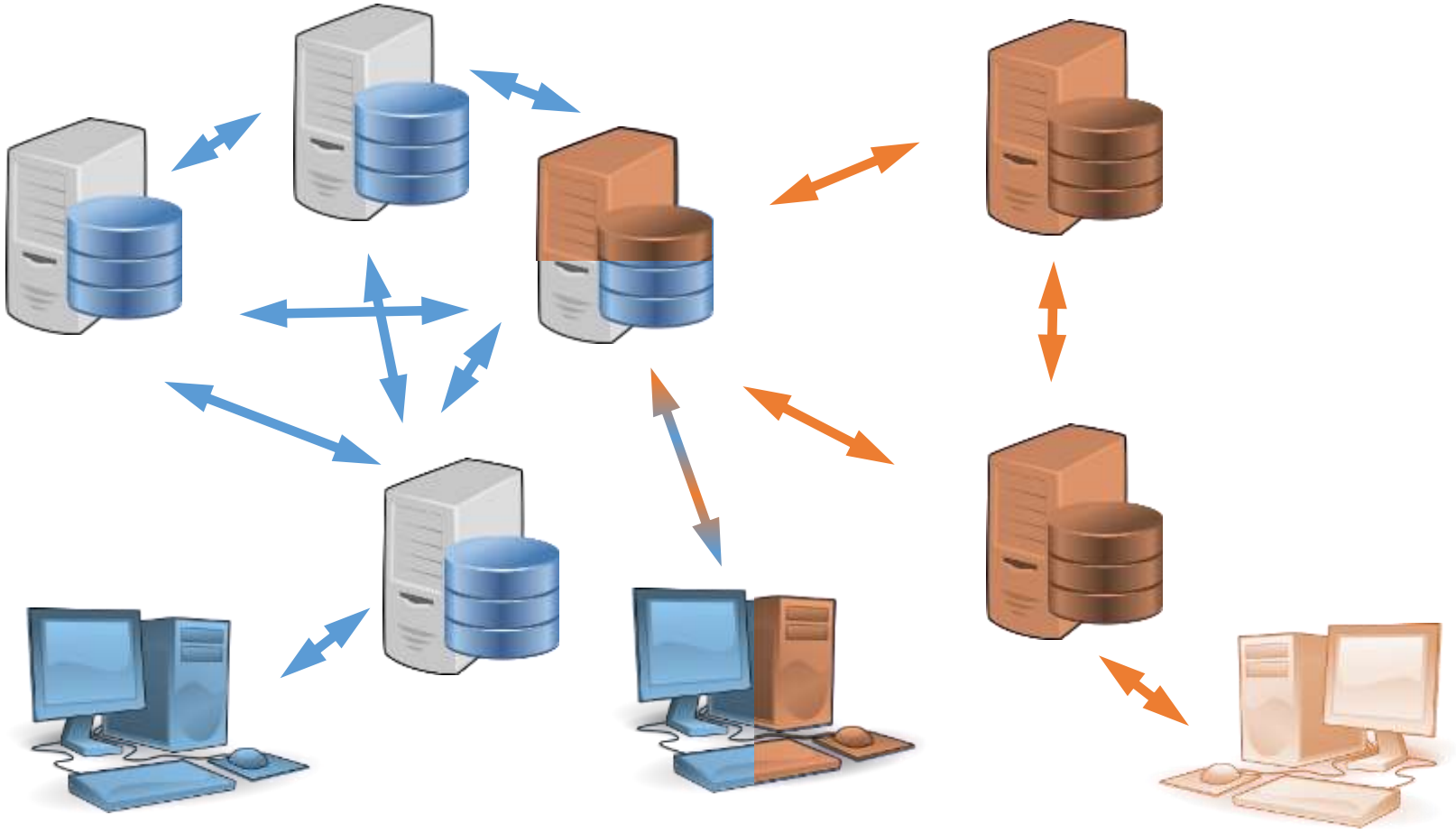


Zukünftige Entwicklung: Sicherheit

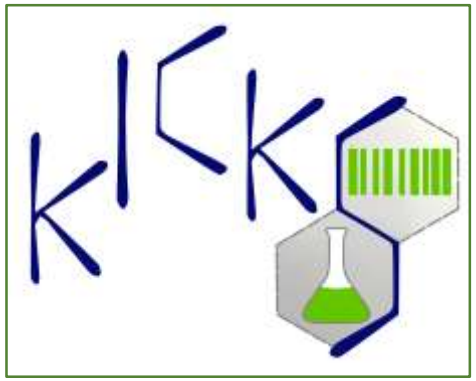
- Erhöhung der Sicherheit
  - HTTP Strict Transport Security (HSTS)
  - Content Security Policy Headers
  - Intruder Lockout
  - Lokale Spiegelung von JavaScript-Quellen (Bootsfaces, JQuery, …)
- Datenschutz-Compliance
  - Datenschutzerklärung
  - Datensparsamkeit
  - Anonymisierung

Zukünftige Entwicklung: Multi-Cloud

- Parallelbetrieb mehrerer Clouds
- Verbesserung der Wiederverwendbarkeit, Vermeidung mehrerer Instanzen pro Institut
- Strikte Trennung der einzelnen Clouds
- Zertifikatshierarchie / disjunkte CAs



Exkurs: KICKS



- Seit 2007: Verwaltung von kommerziellen Chemikalien (Gefahrstoffdatenbank)
- Java & PostgreSQL-basiert
- ca. 120.000 Verbindungen, > 11.000 Gebinde
- Chemikalienbörse, Barcodes, Versionierung / Historie, PDF-Reports
- Chemische Struktursuche und Substruktursuche
- Seit 2016: auch an der Universidade Federal do Rio Grande do Sul
- Codebasis enthält Vorarbeiten für ein LIMS / ELN

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Substanzdaten - Microsoft Internet Explorer bereitgestellt von IPB

http://kicks.ipb-halle.de/mf/user/subviews/substlist.jsf

Substanzdaten

Seite ▾ Sicherheit ▾ Extras ▾

Hilfe

Substanzliste

CAS/RN:
Summenformel:
Substanzbezeichnung:
Substanz-ID: 0
SMILES:
nach verfügbaren Gebinden filtern:
Substrukturesuche: Editor verbergen

Editor

Optionen

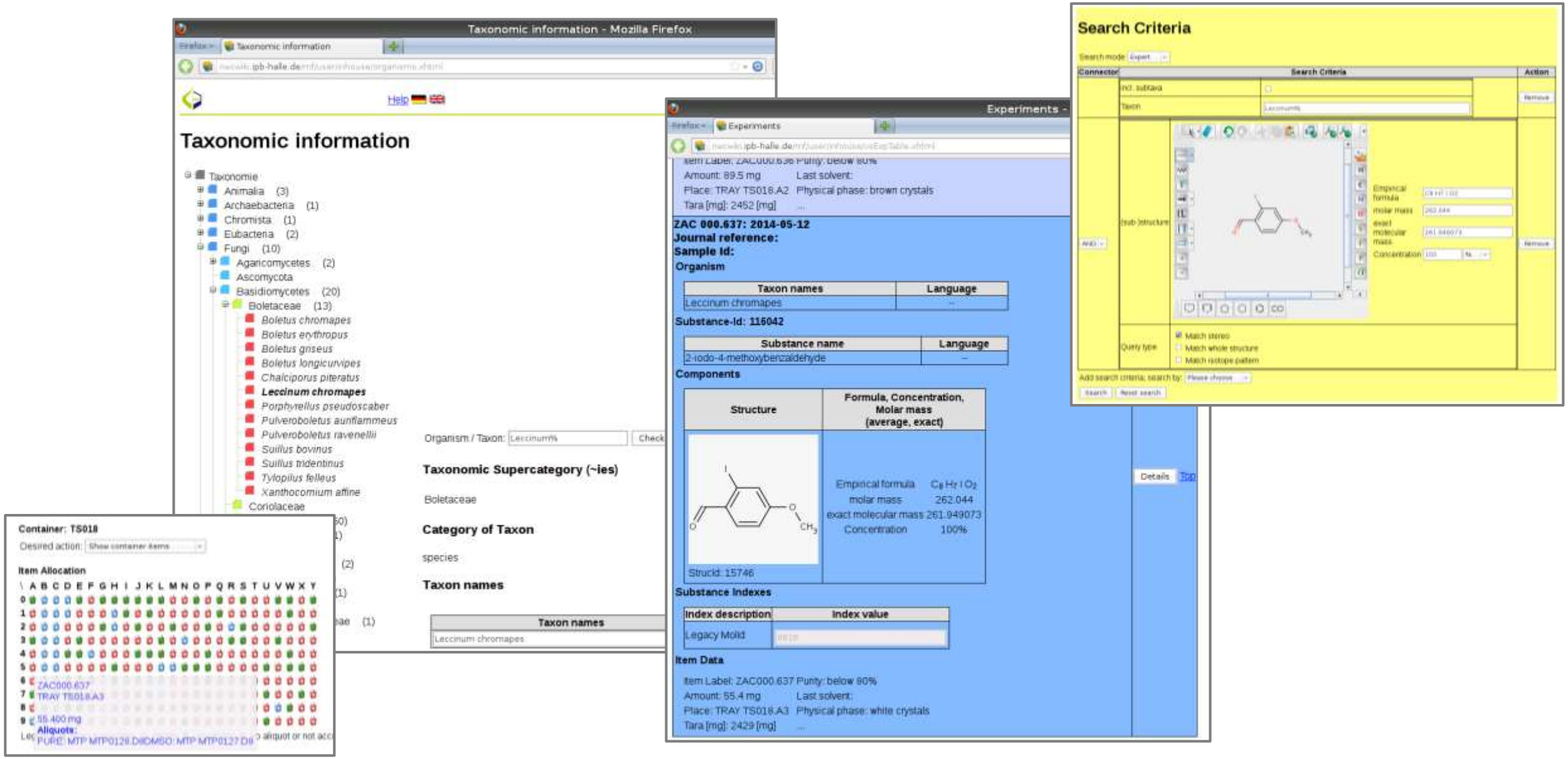
- Stereochemie berücksichtigen
- Nur übereinstimmende Strukturen
- Isotopenmuster berücksichtigen

Suchen | Formular zurücksetzen | Zurück | Neue Substanz

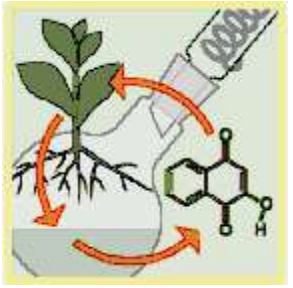
Substanz-ID	Summenformel	CAS/RN	Substanz	Gefahreninformation	Aktion
137	$C_{14}H_9Cl_5O$	115-32-2	Dicofol Dicofol (ISO) 2,2,2-Trichlor-1,1-bis(4-chlorphenyl)ethanc	R: 21/22-38-43-50/53 S: (2-)36/37-60-61	Details ... Neues Gebinde Gebinde anzeigen
4080	$C_{14}H_9Cl_5$	50-29-3	1,1-Bis(4-chlorphenyl)-2,2,2-trichlorethan DDT clofenotane	R: 25-40-48/25-50/53 S: (1/2-)22-36/37-45-60-61	Details ... Neues Gebinde Gebinde anzeigen

http://kicks.ipb-halle.de/mf/user/subviews/substlist.jsf

KICKS: ELN / LIMS Screenshots

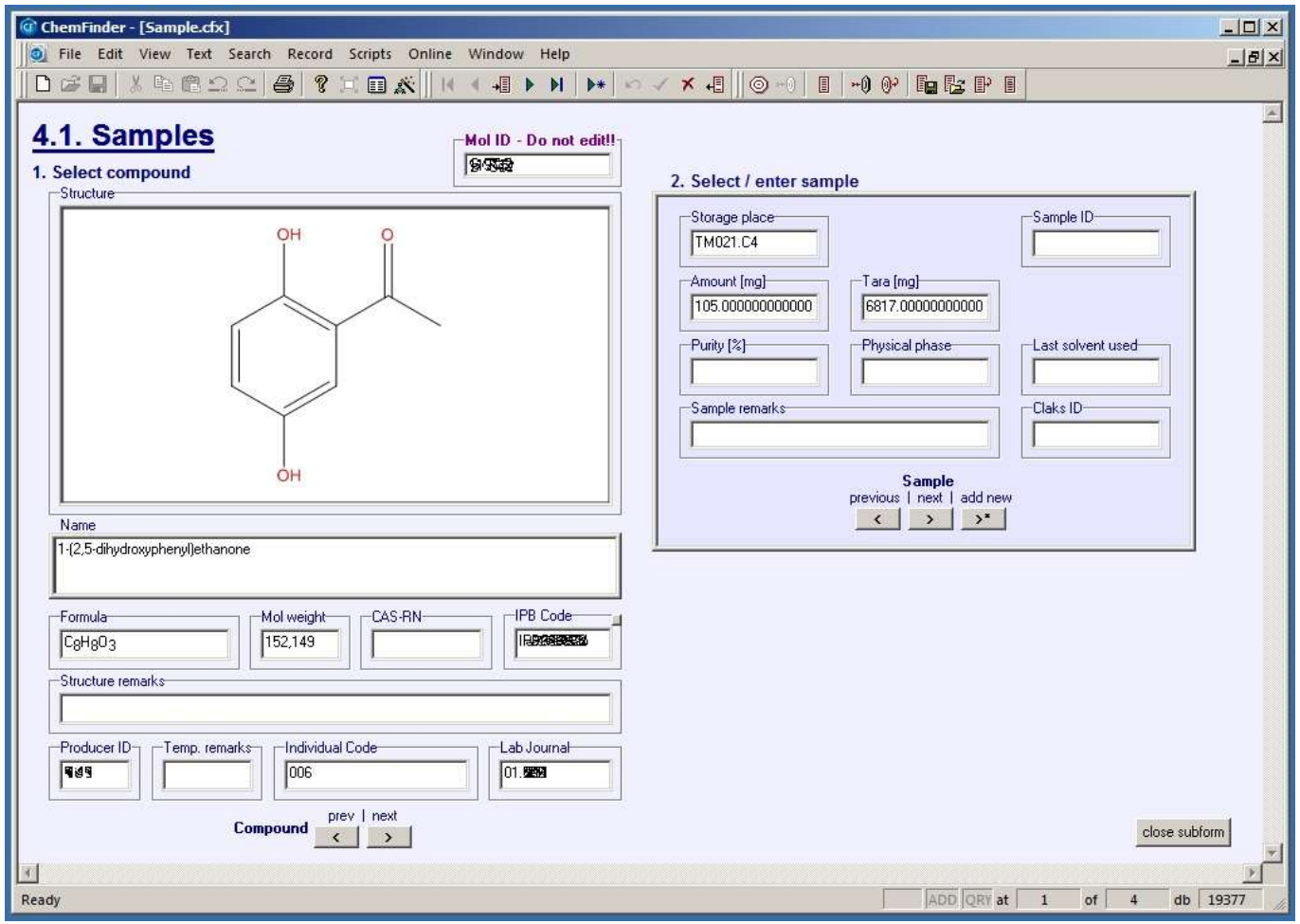


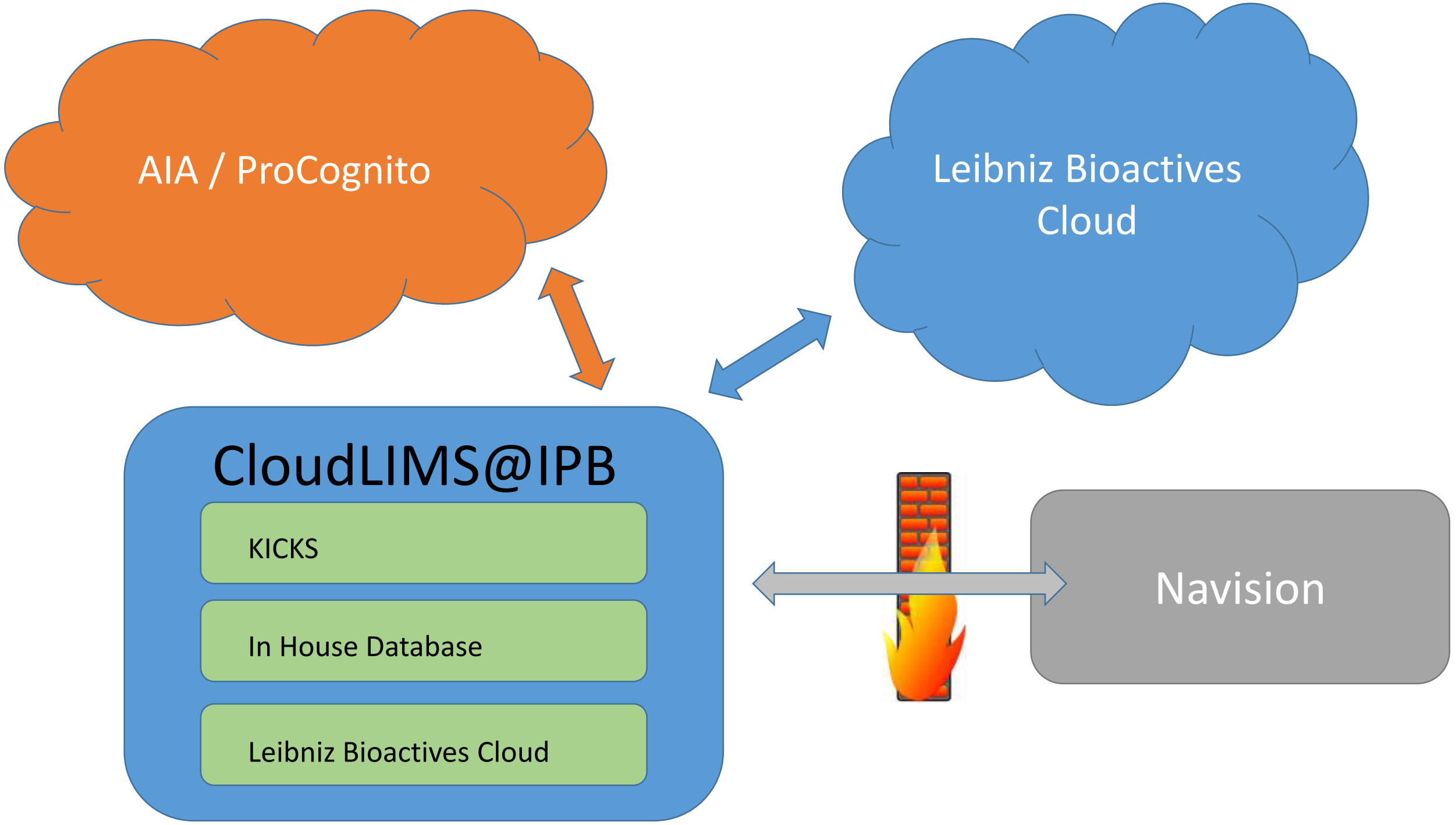
Exkurs: InhouseDB



- „LIMS“ der Abteilung Natur- und Wirkstoffchemie
- > 20.000 Verbindungen, > 26.800 Substanzproben
- > 3.600 Extrakte aus > 1.200 Organismen
- Verknüpfung zu Laborjournalen
- Registratur für Assay-Ergebnisse
- ChemFinder basiert (MS Access ☹)
- Basis für virtuelles Screening

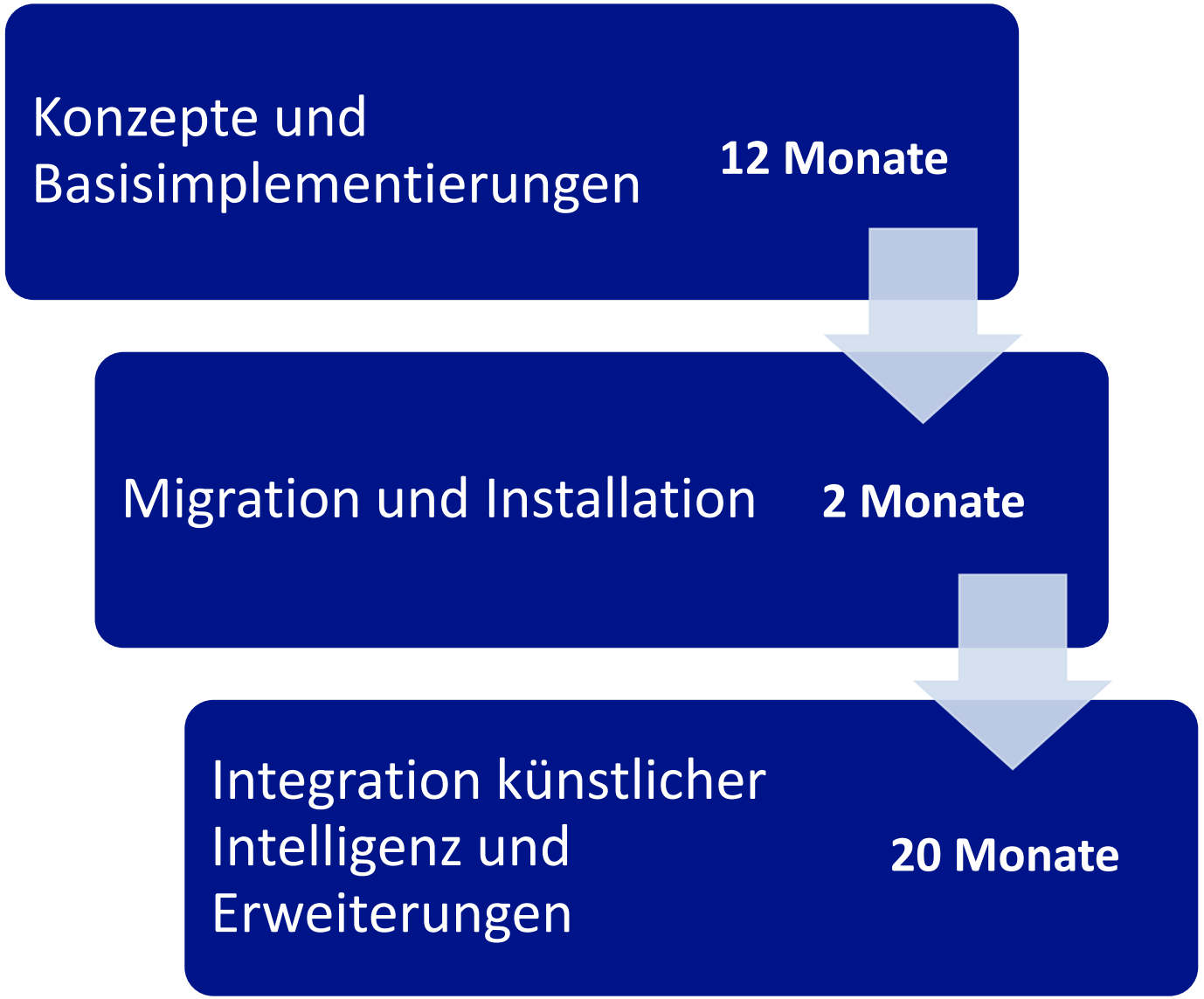
Exkurs: InHouseDB

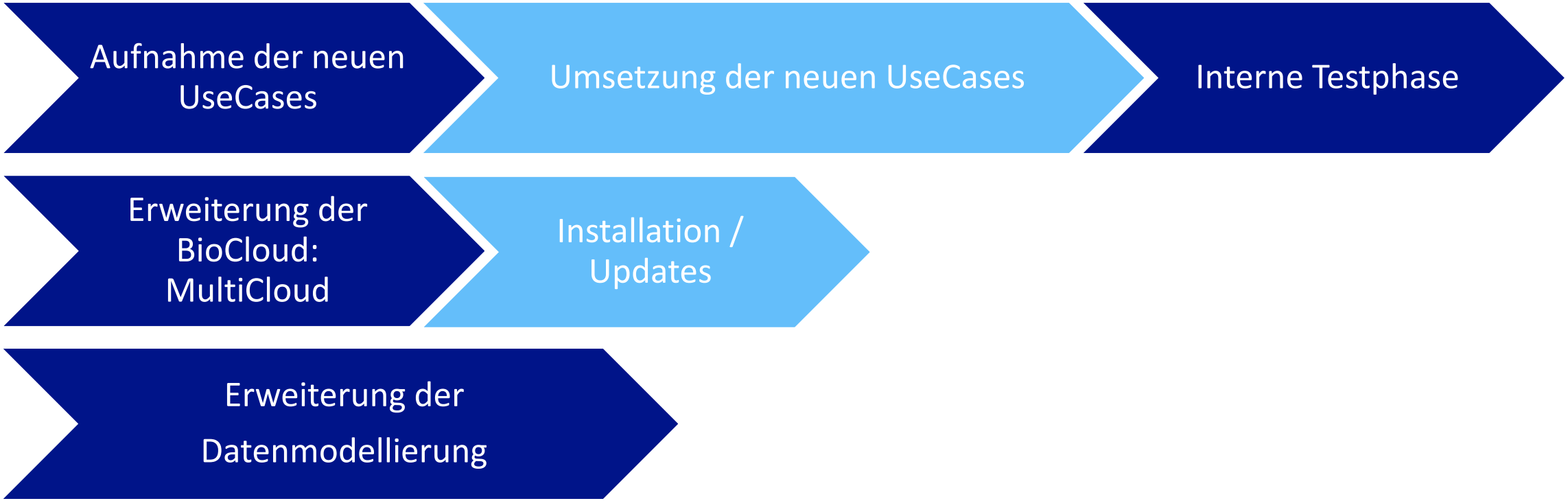



AIA / ProCognito
Leibniz Bioactives Cloud
CloudLIMS@IPB
KICKS
In House Database
Leibniz Bioactives Cloud
Navision

Unterstützung der wissenschaftlichen Arbeit durch eine innovative Software

- Entscheidungsunterstützung bei der Auswahl geeigneter Biotestsysteme
- Gesicherter Fernzugriff auf ausgewählte Bereiche der Substanzdatenbank durch Mitglieder
- Verbesserung der Textanalyse durch den Einsatz künstlicher Intelligenz


Konzepte und Basisimplementierungen
12 Monate
Migration und Installation
2 Monate
Integration künstlicher Intelligenz und Erweiterungen
20 Monate



- Die neuen UseCases umfassen hauptsächlich die der bestehenden Softwarebausteine **KICKS** und **InHouseDB**
- Die Unterstützung des Chemikalienlagers ist ebenfalls Bestandteil der neuen UseCases

Konzept für Datenmigration

Datenmigration

Installation bei einem „AiA“ Mitglied

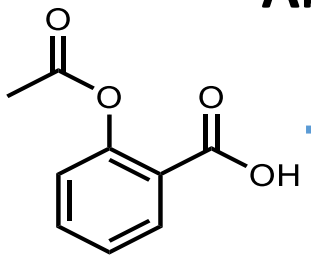
- Die Datenmigration soll verlustfrei und weitestgehend automatisiert geschehen

- Für die Installation ist das Leibniz-Institut für Neurobiologie (LIN) Magdeburg Wunschkandidat
  - Bestehende Installation der Software im Subnetz „BioActivesCloud“
  - Regelmäßige Kontakte mit Projektteilnehmern
  - Geografische Nähe

- Nach der Datenmigration ist zunächst eine Datensichtung mit dem Fokus zur Eignung für künstliche Intelligenz notwendig
- Für die Umsetzung der Entscheidungsunterstützungsfunktionalität müssen verschiedene Ansätze prototypisch getestet werden um eine fundierte Auswahl treffen zu können
- Weiterhin wird an der Erweiterung und Abrundung der Software gearbeitet

KI: Verteiltes SAR

Abfrage



Deskriptoren:
0, 0, 0.3, 0, 0.8, 0, 0, 0 ….

Trainierte künstl. neuronale Netzwerke

Ergebnisse

Institution A:
Evtl. entzündungshemmend

Institution B:
Nicht zytotoxisch

Institution C:
Kein XY-Proteaseinhibitor

Die Struktur wird im Prozess nicht offengelegt!

KI / Becoming Smart: Textanalyse

- Ontologien
  - Kontrolliertes Vokabular
  - Tripelbasierter Datenspeicher („… ist ein …“, „… wohnt in …“, „… ernährt sich von …“, …)
  - Abfrage z.B. mit „Apache Jena“
- Datenextraktion: Regelbasierte Ansätze (reguläre Ausdrücke)
  - Liefert mit überschaubarem Aufwand erste Ergebnisse
  - Komplexität nimmt schnell bis zur Unbeherrschbarkeit zu → Sackgasse
- Neuartige Herangehensweisen: Künstliche Neuronale Netzwerke
  - Google BERT
    Ressourcenbedarf: min. 64 Gbyte RAM
    Trainingszeit 30 Tage@24 Cores, 1 Tag@GPU Tesla V100

Context (Wikipedia: Herbal medicines)

Herbal medicine is also called phytomedicine or phytotherapy. Paraherbalism describes alternative and pseudoscientific practices of using unrefined plant or animal extracts as unproven medicines or health-promoting agents. Paraherbalism differs from plant-derived medicines in standard pharmacology because it does not isolate or standardize biologically active compounds, but rather relies on the belief that preserving various substances from a given source with less processing is safer or more effective – for which there is no evidence. Herbal dietary supplements most often fall under the phytotherapy category.

BERT: Question / Answer

1) What are alternative terms for herbal medicine?
   phytomedicine or phytotherapy
2) In which category fall herbal dietary supplements?
   phytotherapy
3) How would one describe the paraherbalistic practice of using unrefined extracts and unproven medicines?
   alternative and pseudoscientific

KI: Herausforderungen

- Überführung der Prototypen in anwendungsbereiten Code ist nicht trivial
- Training erfordert hohe Datenmenge **und –qualität**
- Anwendung und Setup werden wahrscheinlich komplex (z.B. viele Meta-Parameter)
- Ressourcenanforderungen zu hoch für normale Cloud-Knoten
  → Zukünftige Strategie: Offload zu externen Compute-Ressourcen