Kurzprofil Stefan Lang

Datenwissenschaftler, Bioinformatiker und Systembiologe (M.Sc.)

Kontaktdaten

Mühlenstr. 45
D-07745 Jena

kontakt@slang-it.de

www.slang-it.de

+49 3641 2690220

Sprachen

Deutsch

Englisch

Russisch

Programmiersprachen

Python

Java

R

C++

Skills

KI / Maschinelles Lernen

Datenanalysen

Bioinformatik

DevOps / MLOps

Labor

Mikrobiologie

Molekularbiologie

Mikroskopie

Besonderheiten

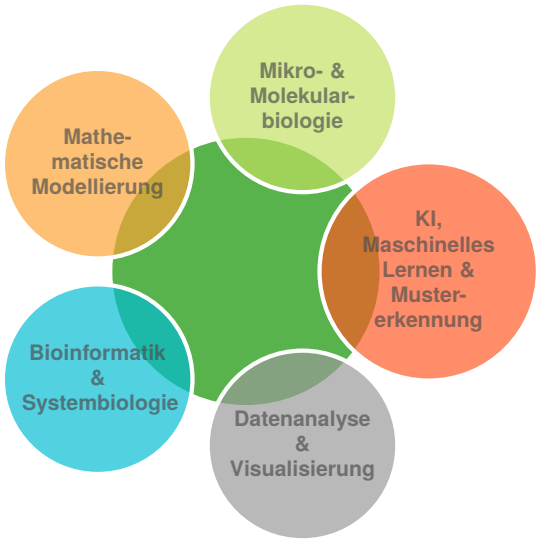
Interdisziplinäre Ausbildung & Arbeit

- Schnittstelle Mathematik, Biologie, Physik, Chemie
- Spezialisierung: Maschinelles Lernen, KI und Systembiologie
- Vertiefungsbereich: Mikro- und Molekularbiologische Methoden

Systemisches und innovatives Denken

- Neben dem Verhalten interessiert mich vor allem das Verhältnis der Elemente eines Systems

Forschungsprofil



Projekte

04/2024 - 07/2024

Bereich: Natürliche Sprachverarbeitung (NLP), große Sprachmodelle (LLMs), Chat-Bots.

Resultate: Entwicklung eines Llama-3 Chat-Bots, welcher verschiedene Werkzeuge (Kunden-Datenbanken, Web-Suche, Chat, ...) benutzen kann um Nutzerfragen zu beantworten. Der Bot kann komplett auf kundeneigenen Servern gehostet werden, was auch die Verarbeitung sensibler Daten ermöglicht.

Methoden:

- **Daten-Pipeline:** Pytorch-Lightning Modul um Text-Dokumente mittels eines (kleinen) Sprachmodells in einen Vektorraum zu überführen (Jina-V2 Text-Embeddings) und zusammen mit beliebigen (auch verschachtelten) Metadaten in einer Vektordatenbank zu speichern
- **Chat-Bot:** LangGraph-Agent mit Llama-3 LLM um Nutzerfragen zu beantworten. Der Agent hat verschiedene Werkzeuge zur Verfügung um Informationen über die Frage aus den verknüpften Datenbanken und dem Internet zu sammeln oder Rückfragen an den Nutzer zu stellen, wenn die Frage unklar ist. Anhand der gesammelten Informationen erstellt der Agent dann eine Antwort (Retrieval-Augmented Generation, RAG)

02/2024 - 03/2024

Bereich: Computer Vision, Optical Character Recognition (OCR)

Resultate: Modul um Text in Bildern zu erkennen und mit einem kundenspezifischen Template zu vergleichen.

Methoden:

- **OCR:** Identifikation von Text-Blöcken in Bildern & Erkennung des Textes
- **Matching:** Lokale Ähnlichkeitsbestimmung des erkannten Textes mit einem kundenspezifischen Template unter Berücksichtigung möglicher OCR-Fehler

10/2023 - 12/2023

Bereich: Computer Vision, Segmentierung von Bildern, Objekterkennung

Resultate: Tool, um einzelne Objekte verschiedener, durch den Kunden vorgegebenen, Klassen in Bildern zu identifizieren und einzufärben

Methoden:

- **Daten-Pipeline:** API-Anbindung an Annotations-Tool des Kunden für die Erstellung der Trainingsdaten
- **Modell:** Neuronales Netzwerk für Multi-Klassen, Multi-Objekt Segmentierung von Bildern
- **Bereitstellung:** Microservices für Training & Inferenz

11/2022 - 08/2023

Bereich: Musik-Informationsgewinnung, Audio-Fingerprinting & -Matching

Resultate: Automatisierte Erkennung von Musikstücken in Live-Aufnahmen. Die entwickelte Methode kann u.a. Instrumental- / Gesang-Versionen, Variationen in Gesang oder Instrumental (bis hin zu geänderten Instrumenten oder Gesang in einer anderen Sprache) und Ausschnitte von Musikstücken in einer Datenbank von Audio-Aufnahmen identifizieren

Methoden:

- **Musik-Detektion:** Klassifikation von Musik bzw. Extraktion einzelner Musikstücke aus gemischten Aufnahmen (z.B Fernsehsendungen, Live-Konzerte, Alben, ...)
- **Musik-Dekomposition:** Dekomposition der Musikstücke in die Audiospuren ”Gesang” und ”Instrumental”
- **Matching:** Erstellung eines elektronischen Fingerabdruckes der beiden dekomponierten Audiospuren und lokale Ähnlichkeitsbestimmung der Fingerabdrücke um die Musikstücke in einer Datenbank zu identifizieren.

02/2023 - 05/2023

Bereich: Natural-Language-Processing, Named-Entity-Recognition, Relation-Tagging

Resultate: Entwicklung eines Tools zur Erkennung und Verknüpfung kundenspezifischer Begriffs-Klassen aus Fließtext

Methoden:

- **Modelle:** Named-Entity-Recognition (NER) Modell mit Transformer-Embeddings um die Begriffe zu annotieren, Relation-Tagging Modell um die Begriffe zu verknüpfen (Bibliotheken: PyTorch)
- **Annotations-Pipeline:** Import- / Export-Funktionen um Beispiele für die zu erlernenden Begriffs-Klassen und Relationen manuell mittels eines grafischen Annotations-Tools (INCEpTION) auf Websites zu markieren
- **Trainer:** Modul um die KI-Modelle an die manuell annotierten Daten anzupassen, also die kundenspezifischen Begriffs-Klassen und Relationen zu erlernen

06/2022 - 10/2022

Bereich: Implementierung eines Neuronalen Netzwerks für multimodale Autokodierung

Resultate: Aufbau einer Bibliothek von Eingabe- und Ausgabe-Adaptern für die generische perceiverIO Architektur. Implementierte Modalitäten (Datentypen): Text, Audio, Bilder, Videos, Zeitreihen

Methoden:

- **Eingabe-Adapter:** Modalitätsspezifische Restrukturierung der Eingabedaten als 2-Dimensionales Array und Konkatenierung der Modalitäten als Eingabe für perceiverIO
- **Ausgabe-Adapter:** Entwicklung von Abfragen (query arrays) für die Rekonstruktion (Autokodierung), Klassifikation und Vorhersage der Eingabedaten
- **Modell:** Methoden zur Vorbereitung der Daten, Konfiguration von Modellen (je nach Eingabedaten und Aufgabe), Training der Modelle und Verwendung der Modelle

09/2021 - 05/2022

Bereich: Datenwissenschaftliche Analyse und Vorhersage der SARS-CoV-2-Epidemie in Thüringen

Resultate: Zeitliche sowie räumliche Vorhersage epidemiologischer Kennwerte (Neuinfektionen, R-Wert) durch Verknüpfung und Interpretation unterschiedlicher Datenquellen (Infektionszahlen, soziodemographische Daten, Mobilität, ...)

Methoden:

- **Aufbau der Dateninfrastruktur:** Zusammenführung & Aufbereitung der unterschiedlichen Datenquellen in einer Graph-Datenbank (ArangoDB). Pipeline zur Aktualisierung der Daten
- **Datenanalyse:** Frequenzanalyse & Filterung (Glättung). Bestimmung von zeitlichen Abhängigkeiten (Kreuzkorrelation) zwischen den Zeitreihen (innerhalb von Orten und zwischen Orten). Bestimmung des Effektes von getroffenen Maßnahmen auf die Zeitreihen
- **Modellierung:** Neuronales Netzwerk für Multivariate Zeitreihenanalyse (Temporal Fusion Transformer) unter Berücksichtigung statischer Kovariablen (Ort, Einwohnerzahl, ...). Bestimmung von partiellen Abhängigkeiten der statischen und dynamischen Kovariablen auf die Zielvariable
- **Bereitstellung:** Docker-Container mit Datenbanken, Modellen und API

07/2020 - 08/2021

Bereich: Speech Recognition, Natural Language Processing

Resultate: Automatisierte Transkription von Audio-Dateien in deutscher Sprache. Überwachung der Transkriptionsqualität und Training von neuen / unerkannten Wörtern. Klassifikation / Interpretation der transkribierten Texte

Methoden:

- **Speech-to-text (STT):** KaldiASR-Modell, trainiert auf deutschsprachigem Datensatz. Bestimmung von Wort-Erkennungs-Wahrscheinlichkeiten zur Qualitätsabschätzung
- **Trainer:** Modul, um dem STT-Modell neue Wörter beizubringen. Prüfen der Erkennungsrate für gegebene Stichwörter. Phonemisierung schlecht erkannter Wörter durch ein separates Graphem-zu-Phonem Modell (g2p). Scraping von Beispieltexten zur Berechnung von Wort-Übergangswahrscheinlichkeiten. Einbindung der neuen Wörter und Phonemisierungen in die Grammatik und Phonem-Klassen des Modells und Retraining des Modells. Synthetisierung einiger Beispieltexte durch ein separates text-to-speech Modell (CoquiTTS) und Rückübersetzung in Text als Validierung
- **Natural-Language-Processing:** Dokument-Indexierung der transkribierten Texte, semantische Suche von Stichwörtern und Klassifikation der Texte anhand vorgegebener Klassen

02/2020 - 06/2020

Bereich: Datenwissenschaftliche Analyse von Produkt-Verkäufen & Bedarfs-Vorhersage von benötigten Materialien

Resultate: Identifizierung von Produktgruppen mit ähnlichem Verkaufsverhalten. Analyse von Trends und Saisonalitäten der Verkäufe. Abschätzung des zukünftigen Materialbedarfes für mehrere Monate

Methoden:

- **Datenvorbereitung:** Anbindung an die Verkaufs-Datenbank. Datensatz mit Produktverkäufen als Zeitreihen und Metadaten zu den Produkten (Einzelteile, Farben, Größe, ...)
- **Datenanalyse:** Auffinden von Korrelationen im Verkaufsverhalten, Gruppierung der Produkte. Frequenzanalyse und saisonale Dekomposition der Zeitreihen
- **Bedarfsvorhersage:** Vorhersage der Verkaufszahlen von Produkten bzw. Produktgruppen unter Berücksichtigung der Produkteigenschaften, aktuellen Trends und saisonalen Verkaufsmustern (Prophet und NBeats ensemble). Abschätzung des zukünftigen Materialbedarfs

10/2019 - 01/2020

Bereich: Analyse und Vorhersage von Fehler-Propagation in Transport-Netzwerken

Resultate: Abschätzung der Auswirkung von Fehlern / Verzögerungen im Ablauf an spezifischen Orten auf das restliche Transport-Netzwerk

Methoden:

- **Datenvorbereitung:** Strukturieren der Daten in Orte, Bewegungen zwischen Orten und Abläufe an Orten. Berechnung von zeitlich statischen und dynamischen Eigenschaften der Orte (Kapazitäten, Auslastungen, ...)
- **Datenanalyse:** Analyse von Bewegungen und Störungen im Netzwerk, Abschätzung der Auswirkung von Störungen auf nachfolgende Stationen (Identifikation von Störungsketten)
- **Simulation:** Simulation des Effektes geänderter Transportwege / -Zeiten oder geänderter Abläufe / Parameter an den Orten auf das Gesamtnetzwerk

09/2015 - 07/2018

Bereich: Charakterisierung und Modellierung von immunologischen Reaktionen auf mikrobielle Infektionen mit Schwerpunkt Autoimmunität

Resultate: Krankheitserreger müssen sich im Körper tarnen, um nicht als fremd erkannt und entfernt zu werden. Die Tarnung kann nicht perfekt sein. Das Immunsystem muss abwägen, ab welchem "Schwellwert" von Eigen-Ähnlichkeit es möglicherweise getarnte Krankheitserreger angreift (ein niedriger Schwellwert bedeutet wenig Autoimmunität, aber schlechtere Abwehr von getarnten Krankheitserregern, ein hoher Schwellwert gute Abwehr, aber eventuelle Autoimmunität). Identifikation von Zielproteinen für eine medikamentöse Intervention von Autoimmunität

Methoden:

- **Literaturrecherche:** (Angeborenes) Immunsystem, Komplementsystem, soziale Systemtheorie, Mimikry / Krypsis, mathematische / spieltheoretische Modelle von Mimikry, mathematische / metabolische Modelle des Komplementsystems
- **Modellierung:** Transfer verhaltensbiologischer Modelle zur Beschreibung von Mimikry und Krypsis bei Tieren auf die mikrobiologische Ebene (molekulare Krypsis). Verknüpfung der Modelle mit Modellen der angeborenen Immunantwort (speziell Komplementsystem). Modellierung des Kompromisses zwischen Autoimmunität und Abwehr getarnter Krankheitserreger
- **Veröffentlichung:** Publikation der relevanten Ergebnisse in wissenschaftlichen Fachzeitschriften

04/2015 - 08/2015

Bereich: Appentwicklung & Geräteentwicklung für die Vor-Ort Analyse in der Lebensmittelsicherheit

Resultate: Implementierung eines Protein-Mikroarrays und Fluoreszenzfilters in einen Smartphone-Aufsatz um vor Ort spezifische (z.B. unerwünschte) Proteine in Proben aufzuspüren (z.B. Wachstumshormone in Milch). Effiziente Analyse direkt auf dem Smartphone durch Methoden des Rechnersehens

Methoden:

- **Datenvorbereitung:** Eingabebilder (farbige Spots des Microarrays, aufgenommen mit Smartphones, also stark variierenden Qualitäten) standardisieren, interpolieren und entzerren (orthogonalisieren)
- **Bilderkennung:** Spots lokalisieren und Ränder markieren. Spots der positiv / negativ Kontrollen identifizieren. Farbintensitäten der weiteren Spots bestimmen und anhand der Kontrollen kalibrieren um die Konzentration des Ziel-Proteins in der Probe zu berechnen

02/2014 - 03/2015

Bereich: Soziale Interaktionen bei Mikroorganismen

Resultate: Charakterisierung von Kooperationsformen in Biofilmen. Insbesondere Modellierung artgleicher und artübergreifender Crossfeeding-Interaktionen. Untersuchung der evolutionären Stabilität der Kooperation bezüglich Parasitismus

Methoden:

- **Literaturrecherche:** soziale Systemtheorie, evolutionäre Spielthorie, Kooperations- und Kommunikationsformen von Mikroorganismen, Crossfeeding
- **Modellierung:** Agenten-basiertes Modell zur Simulation von Crossfeeding-Interaktionen zwischen einzelligen Pilzen. Modellierung des Effektes von Kommunikation über in die Umgebung abgegebene Moleküle oder direkte Verbindung der Individuen durch Nanotubes
- **Veröffentlichung:** Publikation der relevanten Ergebnisse in wissenschaftlichen Fachzeitschriften

06/2013 - 01/2014

Bereich: Maschinelle Erkennung und Verfolgung (Tracking) von Zellen in mikroskopischen Video-Aufnahmen

Resultate: Entwurf und Implementierung von Algorithmen zur Trennung von Zellaggregaten (Segmentation), Tracking der einzelnen Zellen und Extraktion zelltypischer Parameter. Später: Weiterentwicklung zur Analyse von Daten aus konfokaler Laser-Scanning-Mikroskopie (5-Dimensional)

Methoden:

- **Datenvorbereitung:** Dekonvolution der Bilder mit mikroskop-spezifischem Kernel (spezifische Lichtstreuungsmuster entfernen), Interpolation, Standardisierung
- **Segmentierung:** Vordergrund (fokusierte Zellen) von Hintergrund (Rauschen, Makromoleküle, nicht fokussierte Zellen, ...) trennen
- **Bilderkennung:** Einzelne Zellen und Zellcluster erkennen. Zellcluster trennen. Form der einzelnen Zellen rekonstruieren
- **Merkmalsvektoren extrahieren:** Spezifische Merkmale der Zelltypen erkennen und gegebene Eigenschaften charakterisieren (Größe, Bewegungsmuster, Geschwindigkeit, ...)