Stefan Lang

Data Scientist, Bioinformatician and Systems Biologist (M.Sc.)

Personal Data

Mühlenstr. 45
D-07745 Jena

contact@slang-it.de

www.slang-it.de/en

+49 3641 2690220

Languages

- German
- English
- Russian

Programming

- Java
- Python
- R
- C++

Skills

- AI / Machine Learning
- Data Analysis
- Bioinformatics
- DevOps / MLOps

Lab

- Microbiology
- Molecular Biology
- Microscopy

Specialities

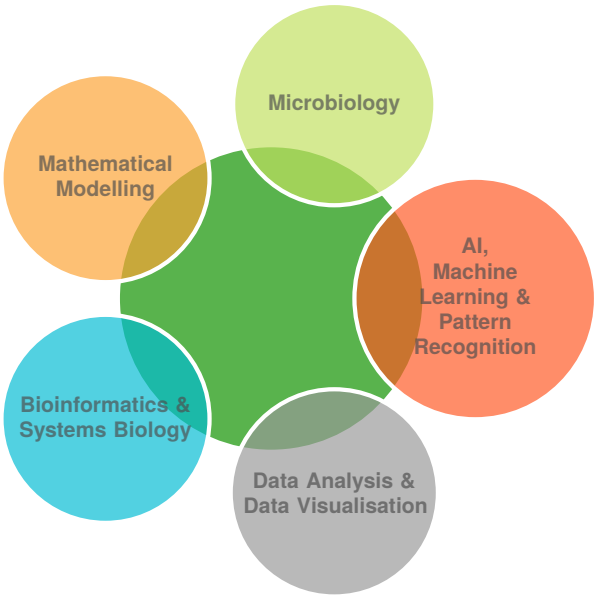
Interdisciplinary Education & Experience

- Interface mathematics, biology, physics, chemistry.
- Specialisation: machine learning, AI & systems biology.
- Broadened skills: microbiological methods and techniques of molecular biology.

Systemic and Innovative Reasoning

- Interested in the interactions between elements forming a system.

Research Profile



Projects

10/2023 - 12/2023

Field: Computer vision, image segmentation, object recognition

Results: Tool for identifying and coloring individual objects of different classes specified by the customer in images

Methods:

- **Data-pipeline:** API connection to the customer's annotation tool for the creation of training data
- **Model:** neural network for multi-class, multi-instance segmentation of images
- **Deployment:** microservices for training & inference

11/2022 - 08/2023

Field: music information retrieval, audio fingerprinting & matching

Results: Automated recognition of music tracks in live recordings. The developed method can identify instrumental / vocal versions, variations in vocals or instrumentals (up to changed instruments or vocals in a different language), and excerpts of music pieces in a database of audio recordings

Methods:

- **Music detection:** classification of music or extraction of tracks from mixed recordings (e.g., television broadcasts, live concerts, albums, ...)
- **Music decomposition:** decomposition of musical pieces into vocal and instrumental channels
- **Matching:** creation of an electronic fingerprint of the decomposed audio channels and local similarity analysis of the fingerprints to identify the music pieces in a database

02/2023 - 05/2023

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

Results: development of a tool for recognizing and linking custom term classes from continuous text. The tool can be used, for example, to recognize people and activities / subjects on websites and to associate activities with individuals

Methods:

- **Models:** Named-Entity-Recognition (NER) model using transformer embeddings to annotate the terms, Relation-Tagging model to link the terms (libraries: PyTorch / FlairNLP)
- **Annotation pipeline:** import / export functions to manually tag examples of the term classes and relations to be learned using a graphical annotation tool (INCEpTION)
- **Trainer:** module to adapt the AI models to the manually annotated data, i.e. to learn the customized term classes and relations

06/2022 - 10/2022

Field: implementation of a neural network for multimodal autocoding

Results: Construction of a library of input and output adapters for the generic perceiverIO architecture. Implemented modalities (data types): Text, audio, images, videos, time series

Methods:

- **Input adapters:** modality-specific restructuring of input data as a 2-dimensional array and concatenation of modalities as input to perceiverIO
- **Output adapter:** development of queries (query arrays) for reconstruction (autocoding), classification, and prediction of the input data
- **Model:** methods for data preparation, configuration of models (depending on input data and task), training of models and use of models

09/2021 - 05/2022

Field: datascience analysis and prediction of the SARS-CoV-2 epidemic in Thuringia

Results: Temporal as well as spatial prediction of epidemiological parameters (new infections, R-value) by linking and interpreting different data sources (infection numbers, socio-demographic data, mobility, ...)

Methods:

- **Building the data infrastructure:** merging & processing the different data sources in a graph database (ArangoDB). Pipeline for updating the data
- **Data analysis:** frequency analysis & filtering (smoothing). Determination of temporal dependencies (cross-correlation) between time series (within locations and between locations). Determination of the effect of measures taken on the time series
- **Modeling:** Neural network for multivariate time series analysis (Temporal Fusion Transformer) taking into account static covariates (place, number of inhabitants, ...). Determination of partial dependencies of the static and dynamic covariates on the target variable
- **Deployment:** Docker container with databases, models and API

07/2020 - 08/2021

Field: Speech Recognition & Natural Language Processing

Results: Automated transcription of audio files in German language. Monitoring of transcription quality and training of new / unrecognized words. Classification / interpretation of the transcribed texts

Methods:

- **Speech-to-text (STT):** KaldiASR model trained on German language dataset. Determination of word recognition probabilities for quality estimation
- **Trainer:** module to teach new words to the STT model. Testing recognition rate for given keywords. Phonemization of poorly recognized words using a separate grapheme-to-phoneme model (g2p). Scraping sample texts to calculate word transition probabilities. Incorporation of the new words and phonemes into the grammar and phoneme classes of the model and retraining of the model. Synthesizing some example texts through a separate text-to-speech model (CoquiTTS) and retranslating them into text as validation
- **Natural language processing:** document indexing of the transcribed texts, semantic search of keywords and classification of the texts based on given classes

02/2020 - 06/2020

Field: datascience analysis of product sales & demand forecasting of required materials

Results: Identification of product groups with similar sales patterns. Analysis of trends and seasonality of sales. Estimation of future material requirements for several months

Methods:

- **Data preparation:** connection to sales database. Data set with product sales as time series and metadata about the products (single parts, colors, size, ...)
- **Data analysis:** finding correlations in sales behavior, grouping of products. Frequency analysis and seasonal decomposition of time series
- **Demand forecasting:** predicting sales of products or product groups, taking into account product characteristics, current trends and seasonal sales patterns (Prophet and NBeats ensemble). Estimation of future material demand

10/2019 - 01/2020

Field: analysis and prediction of fault propagation in transportation networks

Results: Estimation of the impact of errors/delays in processes at specific locations on the remaining transportation network

Methods:

- **Data preparation:** structuring data into locations, movements between locations, and processes at locations. Calculation of temporal static and dynamic properties of locations (capacities, load factors, ...)
- **Data analysis:** analysis of movements and disturbances in the network, estimation of the effect of disturbances on subsequent stations (identification of error-chains)
- **Simulation:** simulation of the effect of changed transport routes / times or changed processes / parameters at the locations on the overall network

09/2015 - 07/2018

Field: characterization and modeling of immunological responses to microbial infections with emphasis on autoimmunity

Results: Pathogens must camouflage themselves in the body to avoid being recognized as foreign and beeing removed. The camouflage cannot be perfect. The immune system must weigh at what ”threshold” of self-similarity it might attack camouflaged pathogens (a low threshold means little autoimmunity, but poorer defense against camouflaged pathogens, a high threshold means good defense but possible autoimmunity). Identification of target proteins for drug intervention of autoimmunity

Methods:

- **Literature review:** (innate) immune system, complement system, social systems theory, mimicry/crypsis, mathematical / game theoretical models of mimicry, mathematical / metabolic models of complement system
- **Modeling:** transfer of behavioral models describing mimicry and crypsis in animals to the microbiological level (molecular crypsis). Linking crypsis models to models of the innate immune response (specifically complement system). Modeling of the trade-off between autoimmunity and defense against camouflaged pathogens
- **Publication:** Publication of relevant results in scientific journals

04/2015 - 08/2015

Field: app development & device development for on-site analysis in food safety

Results: Implementation of a protein microarray and fluorescence filter in a smartphone attachment for on-site detection of specific (e.g., unwanted) proteins in samples (e.g., growth hormones in milk). Efficient analysis directly on the smartphone using computer vision methods

Methods:

- **Data preparation:** standardize, interpolate, and rectify (orthogonalize) input images (colored spots of the microarray taken with smartphones, i.e., highly varying qualities)
- **Image recognition:** localize spots and mark edges. Identify spots of positive / negative controls. Determine the color intensities of the other spots and calibrate them against the controls to calculate the concentration of the target protein in the sample

02/2014 - 03/2015

Field: social interactions among microorganisms

Results: Characterization of forms of cooperation in biofilms. In particular, modeling of intra-species and inter-species crossfeeding interactions. Investigation of the evolutionary stability of cooperation with respect to parasitism

Methods:

- **Literature review:** social systems theory, evolutionary game theory, forms of cooperation and communication in microorganisms, crossfeeding
- **Modeling:** agent-based model to simulate crossfeeding interactions between unicellular fungi. Modeling the effect of communication via molecules released into the environment or direct connection of individuals by nanotubes
- **Publication:** Publication of relevant results in scientific journals

06/2013 - 01/2014

Field: Automated detection and tracking of cells in microscopic videos

Results: Design and implementation of algorithms for separation of cell aggregates (segmentation), tracking of single cells and extraction of cell typical parameters. Later: further development to analyze data from confocal laser scanning microscopy (5-dimensional)

Methods:

- **Data preparation:** deconvolution of images with microscope-specific kernel (remove specific light scattering patterns), interpolation, standardization
- **Segmentation:** separate foreground (focused cells) from background (noise, macromolecules, non-focused cells, ...)
- **Image recognition:** recognize single cells and cell clusters. Separate cell clusters. Reconstruct shape of single cells
- **Extract features:** Recognize specific features of cell types and characterize given properties (size, movement pattern, speed, ...)