

## Simon K.

# Python Software Engineer with data engineering skills

### SUMMARY

- 2+ years of experience with Python as a Data Engineer and Deep/Machine Learning Intern
- Experience with Data Vault modeling and AWS cloud services (S3, Lambda, and Batch)
- Cloud Services: Sagemaker, Google BigQuery, Google Data Studio, MS Azure Databricks, IBM Spectrum LSF, Slurm
- Data Science Frameworks: PyTorch, TensorFlow, PySpark, NumPy, SciPy, scikit-learn, Pandas, Matplotlib, NLTK, OpenCV
- Proficient in SQL, Python, Linux, Git, and Bash scripting.
- Had experience leading a BI development team and served as a Scrum Master.
- Native English
- Native German

### SKILLS

<b>Main Technical Skills</b>	Python
<b>AI &amp; Machine Learning</b>	AWS SageMaker (Amazon SageMaker), NumPy, OpenCV, PyTorch, Scikit-learn, TensorFlow
<b>Programming Languages</b>	C++, Java
<b>Python Frameworks and Libraries</b>	Matplotlib, NLTK, NumPy, Pandas, PySpark, PyTorch, Scikit-learn, SciPy, TensorFlow
<b>Data Analysis and Visualization Technologies</b>	Apache Spark, Databricks, Jupyter Notebook, MapReduce
<b>Databases &amp; Management Systems / ORM</b>	Apache Hadoop, Apache Spark, Google BigQuery, Greenplum, MongoDB, MySQL, NoSQL, PostgreSQL, SQL
<b>Cloud Platforms, Services &amp; Computing</b>	AWS, IBM Spectrum LSF, Slurm
<b>Amazon Web Services</b>	AWS Batch, AWS Lambda, AWS S3, AWS SageMaker (Amazon SageMaker)
<b>Azure Cloud Services</b>	Databricks
<b>Google Cloud Platform</b>	Google BigQuery

<b>Virtualization, Containers and Orchestration</b>	Docker
<b>Version Control</b>	Git
<b>Operating Systems</b>	Linux
<b>Third Party Tools / IDEs / SDK / Services</b>	PyCharm
<b>Scripting and Command Line Interfaces</b>	Shell Scripts
<b>Other Technical Skills</b>	Multi-threading, YAML

## EXPERIENCE

### Deep Learning Intern – Multi-task learning, Bosch Center for Artificial Intelligence (BCAI)

03/2022 – 09/2022 (Renningen, Germany)

- Developed novel loss weighting methods for Multi-task learning that outperformed state-of-the-art methods
- Compared loss balancing methods from the literature on tasks such as semantic segmentation, depth, and normal surface estimation on scene understanding datasets such as Cityscapes and NYUv2
- Registered two novel loss weighting methods as patents
- Documented the results within the master thesis

**Technologies:** Python, PyTorch, MTL, Git, IBM Spectrum LSF

### Technical Solutions Specialist/Data Engineer, Scalefree International GmbH

10/2020 – 01/2022 (Hanover, Germany)

- Led the internal BI development team as a Scrum Master
- Established connection between different source systems and the enterprise data warehouse
- Developed processes for loading the staging area and the raw data vault by employing AWS services such as S3, Lambda, and Batch
- Created XML documents using T-SQL and XQuery for an external customer project
- Containerized jobs using docker and yaml to load the enterprise data warehouse and deployed them using AWS batch.

**Technologies:** SQL, Python, Linux, Git, Bash Script, Data, Vault, AWS, YAML



## **Machine Learning Intern – Cloud ML Services, Novatec Consulting GmbH**

11/2019 – 06/2020 (Hanover, Germany)

- Developed a prototype application for churn prediction and evaluating different machine learning algorithms such as Random Forest, SVM, Gradient Boosted Decision Trees, and Logistic Regression using sci-kit-learn
- Compared the Cloud Machine Learning Services MS Azure Databricks with PySpark, AWS Sagemaker, and Google Cloud BigQuery
- Documented the results of the comparison within the bachelor thesis Python Scikit-learn

**Technologies:** PySpark, MS Azure, AWS GCP

## **ACADEMIC PROJECTS**

### **Student Research Project, University of Hildesheim**

12/2020 – 03/2022 (Hildesheim, Germany)

- Conducted image-to-image translation between the domains of regular images and artworks with Deep Generative Adversarial Networks using Tensor-Flow
- Enhanced CycleGAN by introducing a two-objective discriminator as regularization, incorporating adversarial self-defense for better cycle consistency, and applying differentiable augmentation on the target domain with fewer data
- Employed agile intercultural project management techniques to manage the project successfully

**Technologies:** Python, TensorFlow, GANs, Git, Slurm

## **COURSEWORK**

### **Machine Learning, University of Hildesheim**

04/2020 – 09/2021 (Hildesheim, Germany)

- Implemented various machine learning models such as ridge regression with SGD, LASSO with coordinate descent, least-angle regression, logistic regression with Newton method, gradient-boosted decision tree, and AdaBoost from scratch in Python and NumPy on real-world datasets like Rossmann sales and Wine quality data. Employed data preprocessing techniques such as one-hot encoding, stratified sampling, PCA, and KNN data imputation
- Conducted performance comparison of the implemented models with a sci-kit-learn implementation
- Performed exploratory data analysis on various real-world datasets using Pandas and Matplotlib
- Developed a recommender system by applying matrix factorization with SGD on a movie lens 100k dataset

**Technologies:** Python, NumPy, Pandas, sci-kit-learn, Matplotlib



## Deep Learning/Computer Vision, University of Hildesheim

04/2021 – 09/2021 (Hildesheim, Germany)

- Trained a CNN end-to-end on a self-driving dataset (camera view from the car) using regularization techniques such as cutout and mixup and implemented a custom batch normalization layer and residual connections to predict the steering angle in PyTorch
- Computed the saliency map for an input image using an ImageNet pre-trained model
- Compared metric learning techniques such as learned embedding of a simple classification model, contrastive loss, and triplet loss with an embedding layer for MNIST data using TensorFlow
- Implemented transfer learning for training a U-Net model on a real-world weed field image dataset with a custom categorical cross-entropy loss. Pretrained the first half of the model on the classification dataset DeepWeeds using TensorFlow, improving the test accuracy by 1.5% compared to a vanilla U-net model, and visualized the predicted segmentation map
- Generated adversarial examples using the Carlini-Wagner attack against a CNN trained on MNIST data and created sparse perturbations with the Hoyer-Square regularizer using PyTorch

**Technologies:** Python, PyTorch, TensorFlow

## Distributed Computing, University of Hildesheim

04/2020 – 03/2021 (Hildesheim, Germany)

- Performed exploratory data analysis using PySpark on the movie lens 10m dataset and used the Hadoop MapReduce framework on BTS flight data
- Conducted distributed K-means clustering and distributed linear regression using SGD on KDD Cup 1998 dataset and VirusShare executables with OpenMPI, including a performance analysis on the speed-up with different numbers of used cores
- Implemented Naive Bayes and SVM classifiers from scratch to categorize news items on 20 newsgroups text datasets using preprocessing techniques such as bag-of-words and TF-IDF feature representation and the Hadoop MapReduce framework
- Employed distributed matrix factorization using coordinate descent with the Hadoop MapReduce framework on the movie lens 10m dataset

**Technologies:** Python, Hadoop, MapReduce, PySpark, OpenMPI, mpi4py

## Reinforcement Learning, University of Hildesheim

10/2022 – 03/2023 (Hildesheim, Germany)

- Utilized PyTorch to develop both the Deep Q-Learning model and the REINFORCE algorithm with policy gradients from scratch to solve the Gym environment Mountain Car



**Technologies:** Python, PyTorch

## **EDUCATION**

### **M.S. Data Analytics, University of Hildesheim**

04/2020 – 01/2023 Hildesheim, Germany

GPA: 3.5/4.0

### **B.S. Business Information Systems, University of Applied Sciences and Arts Hanover**

03/2016 – 06/2020 Hanover, Germany

GPA: 3.5/4.0

## **CERTIFICATES**

- Certified Data Vault 2.0 Practitioner (CDVP2)
- Professional Scrum Master (PSM I)

