Sanyogita P, PhD

# >LANGUAGES AND TECHNOLOGIES

* **Techniques:** Convolution Neural Networks, unsupervised techniques (Clustering, PCA, t-SNE, Autoencoders), Optimization Algorithms, Natural Language Processing (Tokenization, LDA, NER, Language models), LSTM, Recurrent Neural Networks, Graph Neural Networks, Diffusion models, Multi-Task Learning, multimodal learning, composite AI, Linear and logistic regression, sampling algorithms, statistical techniques.
* **Tools and frameworks:** ImageJ, OpenCV, Python, PyTorch, PyTorch lightning, PyTorch Geometric, FairSeq, Weights and Biases, Keras, TensorFlow, Scikit-learn, Numpy, Pandas, Matplotlib, Seaborn, Bokeh, Plotly, Jupyter, C#, MATLAB, OriginPro, Google Cloud technologies (GCP), AWS, TensorFlow, Gitlab and GitHub, Spektral (graph deep learning), Deep Graph Library, Docker, TorchServe, MLFlow, Optuna, Hugging Face, Hydra, VSCode, PyCharm.
* **Exposure:** MySQL, Flask, Oracle, SQL, NVIDIA XGBoost, NVIDIA RAPIDS (GPU Acceleration), NVIDIA TensorRT (low precision inference), TFlite, CoreML, MultiGPU, reinforcement learning, PySpark, Spark, GeoPandas, ChatGPT (prompt design).

# >WORK EXPERIENCE

**Amgen | Scientist - R & D | SF Bay area , CA** May 2022 -Present The goal of this project is to reduce the drug discovery pipeline time from 10 years to a few months.

* **ML architect:** Spearheaded planning, design, and implementation of a composite AI system. I performed requirement analysis for domain and processes through meetings with cross functional teams. I researched and evaluated state-of-the-art generative AI technologies (Diffusion models, Graph Neural Networks, BERT based Language models, Multi-Task Learning, multimodal learning) to architect a technical solution.
  + This scalable solution resulted in reducing the time to generate and recommend sequence text from 6 months to ~3-4 minutes and it is also expected to significantly reduce the expenses for conducting multiple wet lab experiments.
  + Created best practices for building and maintaining scalable AI systems.

## ML researcher:

* + Invented a generative AI architecture based on business and domain requirements. Invention disclosure record is filed for this work.
  + Attended scientific conferences to stay abreast with the developments in deep learning and machine learning. Explored academic collaboration opportunities with top tier institutes in the United States.
* Performed data collection, cleaning, and preprocessing for public and proprietary data (both sequence text and 3D point cloud). Presented actionable insights using t-SNE plot of sequence text embeddings obtained using MSA Transformer language model, Matplotlib, Seaborn, Bokeh.
* Designed and implemented scalable end-to-end pipeline from scratch using PyTorch, PyTorch lightning, MLFlow, Optuna, VSCode, python, Docker, Gitlab, TorchServe on AWS for generative AI models.
* Made production releases in AWS for Amgenʼs internal AI based system (Inference) supporting research across various groups using Tensorflow, jax, python.
* Worked on a computer vision based AI system using a medical imaging dataset of biobank that is part of the world's largest imaging study for anomaly detection in fat distribution.
* Provided estimation for tasks, create and maintain smartsheets based task tracking for different projects to update leadership teams.
* Mentored and trained interns for building deep learning systems for various projects.

**Absci | Artificial Intelligence Scientist - R & D | SF Bay area, CA** May 2021 - Feb 2022 Worked on a cutting edge proprietary end-to-end multimodal deep learning platform to reduce drug candidate identification timeline from months to few minutes and save expenses on multiple iterations of wet lab experiments.

* **ML researcher/architect:** Researched and evaluated state-of-the-art ML technologies to identify an appropriate technical solution for drug candidate classification. Designed and trained customized Graph neural network based architecture for regression tasks and trained it using wetlab measurements.
  + Trained and evaluated Recurrent Neural Networks for predicting sequence text in the production process using proprietary dataset.
  + Attended scientific conferences to stay abreast with the developments in deep learning and machine learning.
  + Conducted cross functional meetings to identify and collect data from wetlab experiments for training and improving performance of models. Performed domain specific data preprocessing.
  + Established best practices for pipeline using PyTorch lightning, Hydra, Weights and Biases. Fine tuned RoBERTa for sequence classification on 600 million records and tracked the experiments using Weights and Biases on NVIDIA A100 clusters.
  + Worked on a cutting edge platform to develop generative AI models based on geometric deep learning (3D point cloud and sequence text) using PyTorch lightning, PyTorch Geometric, Weights and Biases, FairSeq, Gitlab, Docker, Hugging Face, NVIDIA A100 clusters.
  + Presented actionable insights by clustering of sequence text embeddings obtained using MSA Transformer language model, Plotly, Seaborn, Matplotlib.

**Outco | Machine Learning Tech Coach | SF Bay area, CA** April 2020 – May 2021 | March 2022 - August 2022 Worked as a machine learning tech coach to train and mentor ML engineers, data scientists, data engineers, and recent graduates. Several of my students are currently placed in senior positions in multinational companies.

* Designed, implemented, and delivered educational content for architecture and implementation of scalable machine learning systems using linear and logistic regression, unsupervised techniques (Clustering, PCA,

t-SNE, Autoencoders); Natural Language Processing (Arc90 Readability algorithm for data extraction from url documents, LDA, NER to recognize disease name or treatment, tokenization, BERT) for healthcare ad generation; LSTM, Recurrent Neural Networks, Graph Neural Networks, gitlab, tensorflow, keras, numpy, pandas, Jupyter, matplotlib, Colab, AWS, GCP for various industry verticals for applications such as search, ranking, and recommendations using public and proprietary datasets.

* Designed and implemented a price forecasting system using an ensemble of models using proprietary dataset of 10 years of historical transactions.
* Implemented a paragraph concept classification system: extracted text from pdf contract documents using OCR, prepared and used domain specific ontology for removing stop words, compared different

pre-processing techniques, and visualized the performance of the model using mglearn.

* Provided hands-on coding material in github for advanced machine learning and deep learning techniques such as Graph convolutional networks, transformer neural networks, and deep reinforcement learning using python, TensorFlow, and deep learning libraries.

**Passio | Senior Machine Learning Engineer | Menlo Park, CA** Feb 2020 - Mar 2020 The goal of this project was to detect food and provide nutrition information using mobile devices for airline food industry customers.

* **ML architect/Senior Machine Learning engineer:** Architected a technical solution for a scalable machine learning system for food detection in GCP.
  + Designed and implemented a scalable end-to-end machine learning pipeline using Google Cloud (VM Instance, Google Buckets, Google AI Platform, TFRecords Shard) to efficiently store 50+ GBs of food recognition data.
  + Designed custom image compression techniques to reduce storage size for images in Google Cloud by over 50% implemented in Python with TFRecords Shard.
  + Accelerated data access and processing from Google Cloud Buckets using Python and TensorFlow, reducing
  + Reduced training time for big data images for 700+ classes from 60 hours to 2.5 hours without affecting model accuracy.
  + Trained and tuned Convolution Neural Networks models for image classification for beta testing and researched new algorithms for improving performance in custom use cases of object detection (knowledge graphs, embedding based approach with Elasticsearch).
  + Tested and analyzed performance of deep learning models for obtaining actionable insights.
  + Performed model conversion and quantization in CoreML and TFLite for serving on iOS and Android platform.

## University of Illinois ECE Department | Machine Learning Scientist, Research Assistant Jan 2012 - May 2019

* **ML researcher:** Conducted commercially sponsored (Cisco Systems, Inc) research work for anomaly detection in optical microscope images to improve semiconductor yield. I invented a novel likelihood-based generative machine learning algorithm (few-shots, computer vision) for detecting and for predicting details of rarely occurring critical nanoscale anomalies in microscope images. The main concept of this algorithm is similar to that of diffusion models.
  + Used image pre-processing techniques for camera images: performed filtering to reduce spatial and temporal noise in images. Used PCA to reduce dimensionality and noise. Used ImageJ , openCV, statistical and sampling techniques in originPro for data collection and preprocessing.
  + Evaluated and identified gaps in existing deep learning techniques (GAN and Autoencoder). Implemented a baseline for defect detection and classification using Python.

## Inventions:

* + - **Machine learning algorithms:** Reduced selection bias and invented a likelihood-based generative algorithm for rarely occurring nanoscale anomalies in noisy images. The algorithm uses few training samples (zero shot/few shots) and it is interpretable by design.
    - Implemented this technique to detect and characterize 9nm node defects in microscope images using MATLAB and parallel computing toolkit, optimized for generic hardware.
    - **Optimization algorithm:** Invented first-order gradient based optimization algorithm for sharply varying analytically non-differentiable functions with potential generic applications as an alternative to ADAM and SGD.

Developed mathematical framework for convergence analysis of optimization algorithm, applying it for mathematical functions, machine learning loss functions, and optical imaging phase retrieval techniques.

* **Publication/patents:** Generated funding for ML research through Cisco Systems, Inc to cover for assistantship, tuition waiver, and project expenses, leading to 2 provisional patents and first author journal publication. This paper was considered as a seminal paper in the scientific community for its unique blend of optics and machine learning.

**Byer California | Software Engineer (Consultant)** Jun 2008 - Nov 2009

* Designed, developed, and deployed inventory audit application on wearable devices using C# and Oracle, significantly improving quality control and resulting in appreciation from users and customers.
* Collaborated with diverse stakeholders (developers, management team, finance, marketing, warehouse supervisors, hardware teams) for requirement analysis; received employee of the year award.

**Johnson & Johnson | Project Leader (Consultant- TCS)** Jan 2006 - Oct 2007

* Hired, trained, and led team of 4 to create automated DNA synthesis app in C# and multi-threading; researched and developed new mathematical algorithms for load balancing DNA synthesis hardware, optimizing production time by over 3X. Requirement analysis, design, development, testing, project management, and estimation.

**Tata Research Development and Design Center-TCS | Senior Software Engineer** Jan 2003 - Jan 2006

* Researched and developed unique algorithm for phrase search multiple indexing for in-house search and case based artificial intelligence product for improving accuracy and efficacy in phrase search engine in C#.
* Developed wearable device-based applications for centralized inventory control, audit, and seamless integration to improve overall efficiency of operations for Singapore Airlines using C#.Net, SOAP, and XML.
* Researched and developed unique mathematical algorithm for automating and characterizing insurance case analysis for American International Group with C#, ASP.NET, ADO.NET SQL Server, and Oracle.

**>AI Fellow | Insight Data Science** Jun 2019 – Jan 2020

## TrademarkRADAR | Artificial Intelligence | [code](https://github.com/snygt2007/Gita_Insight_Project2019)

*Search engine designed to detect trademark infringement in similar looking trademarks. Used median filtering to remove salt and pepper noise, sharpened the blurred images using OpenCV. Built end to end pipeline in AWS for training*

*and inference of models using Convolutional Neural Networks, k-nearest neighbors, S3, boto, tensorflow, keras, and github.*

**City-Scale Multi-Camera Vehicle Re-Identification | Machine Learning** Jan 2020 - May 2021

*Working on AI cities to derive actionable insights from sensor data for improving the transportation system.*

* During AI city 2020, I Processed images from street cameras for feature extraction and training of deep learning models for object detection and classification using multi-stage algorithms and synthetic data generation.
* Working on AI city 2021 for vehicle Identification using Natural Language Processing and Computer Vision using large datasets consisting of text and video frames.

# >EDUCATION

**Ph.D. Electrical and Computer Engineering,** *University of Illinois at Urbana-Champaign* May 2019 **Graduate courses in Electrical Engineering,** *University of California Berkeley Concurrent Enrollment* Dec 2011 **M.Tech. Optoelectronics and Optical Communication,** *Indian Institute of Technology, Delhi* Jan 2003

# >RESEARCH EXPERIENCE

*Publication and Patents:*

Published state of the art results. Optical inspection of nanoscale structures using a novel machine learning based synthetic image generation algorithm, Sanyogita Purandare et al. Optics Exp. 27. 17743. 10.1364/OE.27.017743 (2019).

Provisional patent U.S. Patent Application No. 62649885: “Defect detection using machine learning”, L. Goddard and S. Purandare.

Provisional patent U.S. Patent Application No. 62842853: “Optical inspection of nanoscale structures using a novel machine learning based synthetic image generation algorithm”. L. Goddard, S. Purandare, A. Schwing.

*Ebook:*

S. Purandare, S. Qin, S.-V. Dhople, A. Dominguez-Garcia, R. Pilawa, and L. L. Goddard, Harnessing Energy, L. L. Goddard, editor. Creative Commons, 2012. <http://psl.mntl.illinois.edu/HarnessingEnergy.zip>.

# >LEADERSHIP + AWARDS

* Sponsored by Booz Allen Hamilton to attend Women's Early Career Accelerator program during GTC 2019 conference; received training using deep learning techniques such as CNN, LSTM, GAN, Autoencoder on NVIDIA tools/platforms (NVIDIA XGBoost, NVIDIA RAPIDS, NVIDIA TensorRT) alongside mentoring and 1 year platform access.
* Recommended for PhD Dissertation Completion Fellowship in 2018 by the UIUC-ECE
* Yang Research and Entrepreneurship Award Finalist in UIUC-ECE department 2017
* GATE scholarship for M.Tech. based on academic performance
* Recommended for DAAD-IIT Scholarship Program in IIT Delhi for academic performance

# >CERTIFICATIONS

**Security and Privacy of Big Data,** *Coursera* 2021 - Present

**Business Management for Engineers,** *University of Illinois at Urbana-Champaign* 2019 - Present