MARC SCHNEIDERMAN

Email: ai.architect@nteligence.com

ROLES:
Cognitive Computing
Information Security
Chatbots
Intelligent Virtual Assistants
Electronic Advisors
Future State Architecture
Symbolic Reasoning
Rule Engines
VoiceBots
Deep Learning
Intelligent Process Automation

INDUSTRY APPLICATION:
Financial Services
Insurance
High Technology
Healthcare
Pharmaceutical
Government
Retail
Transportation
Telecom / Cable TV
Media
Travel / Leisure
Energy

EDUCATION:
Benjamin Cardozo HS

ADDITIONAL TRAINING:
Certified in predictive modeling and data analytics by the SAS Institute

PATENT:
Automatic Thread Migration. A means by which to automatically relocate a software robotic process, completely intact, across a network of computers based upon changing environmental conditions.

EXPERIENCE SUMMARY
AI architect and data scientist with over twenty years of experience. Expert in the use of open source software infrastructure, development tools, frameworks, models, and libraries. Subject matter expert (SME) in artificial intelligence, decision science, and call center automation. Including NLU, Large Language Models (LLM), generative AI (text, audio, code, etc.) deep learning, automated reasoning, logic programming and intelligent process automation. Track record of creating highly successful software products and services. Extremely broad range of skills that have consistently created new revenue streams, reduced costs, increased profitability, and provided clients with a competitive edge in the markets they serve.

- Logical and Physical Software Architecture
- Future State Architecture and Roadmaps to Adoption
- Chatbots, Voicebots, and Intelligent Virtual Assistants
- LLMs (GPT-J, T5, Falcon, Mistral, BERT, Llama 2, etc.)
- Voice Models (VITS, Hifigan, Tacotron2, FastPitch.)
- Conversation Automation
- Retrieval Augmented Generation (RAG)

SKILLS INVENTORY:
Logical and Physical Software Architectures, **Spark**, Ignite, FAISS, AWS, EC2, S3, Azure, **Generative AI,** NLU, Business Rule Engines, Prolog, Speech Recognition, Speech Synthesis, **Retrieval Augmented Generation,** RAG, Chatbots, RASA, Intelligent Virtual Assistants, Voicebots, Hugging Face, **Transformers**, langchain, Llama Index, **GPT-J, Llama 2**, **Mistral**, **Falcon**, VITS, Stable Diffusion, Starcoder, **T5**, **BERT**, NLP, **QLoRA, PEFT, ChromaDB,** Direct Preference Optimization, **DPO,** accelerate, bitsandbytes, Whisper, Nemo, Asterisk, RPyC, Thrift, Ollama, vLLM, **Ray**, Cognitive Computing, Cryptography, Intelligent Process Automation, **Scikit-Learn**, Pandas, NumPy, SciPy, Deep Learning, **PyTorch,** Tensorflow**,** Keras, **RLHF**, Future State Software Architecture and Roadmaps to Adoption, PKI, BLEU, BERT Score, ROUGE, METEOR

PROFESSIONAL EXPERIENCE

***n*Teligence Corporation (West Windsor, NJ)**
January 2011 – Present

Generative AI Architect, Data Scientist, and Engineer
Custom Hardware Appliance for Voicebots
R & D Lab

- Designed and built a Linux based hardware appliance, using a 64 core Threadripper AMD CPU, and multiple NVIDEA Ampere GPUs. Machine utilized liquid cooling technology
- Developed the logical and physical architecture for a conversational AI environment for voicebots
- Identified, tested, installed and configured the necessary infrastructure components, development tools, Python libraries, and programming languages
- Wrote the “glue” code needed to integrate all of the core infrastructure subsystems, including a digital PBX complete with call center, and a state of the art deep learning environment, into a cohesive self-contained platform
- Developed a voice based gateway that would integrate the digital PBX and call center components with the AI environment
- Implemented voice biometrics as part of multi factor authentication protocol for incoming calls
- Created speaker specific embeddings from sample audio data, which was then used to finetune a multi-speaker VITS text-to-speech (TTS) model in order to clone a human voice
- Fine-tuned generic equivalent of GPT 3.5/4.x, using industry and domain specific data, in order to automatically generate voicebot responses to customer’s most Frequently Asked Questions (FAQs)
- Finetuned multiple open source Large Language Models (LLMs) on one or more GPUs using PEFT, QLoRA, and 4-bit quantization
- Deployed finetuned LLMs for low latency inference using industry standard open source tools and techniques
- Developed several proof of concepts, centered around Retrieval Augmented Generation (RAG)to satisfy customer service, product support, and sales use cases.
- Utilized Reinforcement Learning based on Human Feedback (RLHF) to improve the performance of a finetuned LLM, using a methodology called Direct Preference Optimization (DPO)
- Evaluated the performance of finetuned generative textual models using a combination of both manual methods as well as industry standard algorithms. To determine the similarity and quality of LLM generated text as compared to a reference body
- Utilized industry standard python libraries to store conversational history in memory, in conjunction with prompt engineering techniques and multiple LLMs, in order to ensure coreference resolution across multiple turns in a discourse between a digital agent and a caller

AI Architect
Voicebots and Chatbots
Prudential

- Developed the logical architecture that would be needed to integrate a cloud based digital PBX, with a conversational AI backend.
- The architecture defined all of the required subsystems and components, their areas of responsibility, separation of concerns, set of interactions, messages types and structure, protocols, as well as the flow of audio and textual data through the system

- This included a voice gateway, that would serve as an adaptor to bridge the SIP based telephony world with that of a robust natural language processing (NLP) and understanding (NLU) backend
- Defined a client side Application Programming Interface (API) that would enable a digital PBX platform to seamlessly interact with a conversational AI back end
- The API eliminated all existing dependencies on VoiceXML
- Developed protocol for handling both voice and text through bi-directional channels for streaming audio and document based message passing
- Identified gaps between the proposed architecture and existing Genesys and Amazon AWS Lex v2 plug-in and integration components.

Conversational AI Architect and Data Scientist
ADP

- Created a strategic long term vision, and logical architecture, for conversational AI within the client's organization, covering both chatbots, voicebots, and intelligent virtual assistants.
- Created a logical architecture described major subsystems, and their individual components, their responsibilities, and how they would interact with one another
- Architecture defined the use of generative deep learning models to automate responses to customer questions
- Using Tensorflow and its bundled Keras library, built a custom text based deep learning model, using an inception architecture, that would classify chatbot user utterances into one of ninety-seven different potential intents.
- This custom built model had an accuracy rate of about 92%, compared to IBM Watson Asssitant's out of the box accuracy of ~84%
- Performed initial exploratory work in the area of multi language deep learning models for the client's international business unit.
- Prototyped two different multi lingual deep learning models using a multi-lingual BERT checkpoint, and the second from a pre-built XLM-Roberta model.
- The second multi-lingual model achieved an accuracy of 94%, from English language validation data, and generalized well across other languages, such as French and Spanish
- Explored the possibility of using a Generative Pre-Trained transformer model (GPT) to provide dynamically created, automated, chatbot responses

Senior AI Architect
Intelligent Process Automation, Intelligent Virtual Assistants
Conifer Healthcare

- Developed the logical and physical architectures for an intelligent virtual assistant (IVA) that would work right alongside a clinical review nurse.
- Conceived an enterprise level reference architecture for artificial intelligence, that seamlessly integrated machine learning (ML), robotic process automation (RPA), natural language understanding (NLU), and operational decision support (ODM)
- Developed the logical architecture in support of the overall vision of the platform, identifying all of the required components and subsystems, including their individual responsibilities and areas of concern
- Drilled down into the physical architecture for the distributed real-time computing subsystem using the Python Ray package

- Worked with IS business partners, and lines of business leaders, to identify, document, and rank potential use cases
- Designed and built a proof of concept that combined optical character recognition, along with deep learning based text classification, to automatically categorize written correspondence, in PDF and .TIFF formats, received from payors.
- Developed and deployed production level RandomForest model, built using the scikit-learn machine learning library, which would predict the likelihood of a payor (medical insurer) rejecting a claim due to a lack of financial responsibility.
- Wrote Impala queries, and data mangling routines using Pandas, to transform information stored within the data lake into a format suitable for modeling

AI and Information Security Architect
AmeriHealth Caritas

- Researched and assessed leading approaches to building deep learning models that would analyze medical images in order to make a diagnosis. Using a sample x-ray dataset provided by the National Institute of Health, adapted readily available open source deep learning models in order to maximize their recall rate, to an impressive 96%, for the classification of pneumonia and malaria images
- Developed a detailed business case and reference architecture, for the use of machine learning, natural language understanding, and human like heuristics to improve member population health. The goal was to delay the onset, as well as reduce the mortality and morbidity rates normally associated with strokes and diabetes. This would be accomplished through the identification of high risk groups and the subsequent creation of intervention strategies targeted towards affecting (delaying if possible) the medical outcome.
- Developed a Natural Language Processing/Understanding (NLP/NLU) front end for an intelligent virtual assistant that could assess the risk of stroke. Utilized Prolog's Definite Clause Grammar (DCG) to perform constituency parsing, ensuring verb conjugation, plurality of nouns, and subject verb agreement. Through the use of lambda calculus expressions, incrementally built a knowledge based representation of the meaning of the sentences
- Used scikit-learn, learn-imbalanced, and pandas, to build a highly accurate machine learning model that correctly identified 5 out 6 plan members who were likely to experience a stroke at some time in the future.
- Constructed data pipeline used to manipulate raw training and test data. Dropped rows having large numbers of missing values, imputed mean, and performed one hot encoding of nominal values.
- Built a "No Cost" Robotic Process Automation (RPA) environment from scratch, using common off the shelf infrastructure components. This included the JADE mobile agent platform, which provided a highly scalable, container based system for executing, managing, and monitoring the entire lifecycle robotic software processes. Also incorporated the jBPM business process manager, in support of externalization of both business logic and robotic workflows.
- Developed a lightweight RPA framework (set of API's), that "wrapped" the above mentioned open source infrastructure components. This simplified the coding effort required, and reduced the complexity of writing, any type of RPA application.
- Developed an enterprise wide reference architecture for a new Public Key Infrastructure (PKI) initiative within the organization. Definition of components and services included certificate issuance, lifecycle management, policy management, registration, software defined security module (SDSM), cryptographic component, certificate discovery, certificate repository, and validation authority

- Defined the technical criteria that would be used to evaluate certificate lifecycle management products from Venafi and Keyfactor. Established criteria were included within a formal RFP sent out to the above vendors. RFP technical section covered the complete range of core product capabilities including system interfaces, connectors and gateways, automation of core lifecycle tasks, certificate and key discovery processes, automated installation of certificates within existing infrastructure, generation of certificate signing requests (CSR), establishing policies, devops uses, policy management, and workflow management.
- Defined technical criteria for evaluating the capabilities of hardware security modules (HSM) from both Thales/Gemalto, and nCipher
- Using the Java Cryptography Architecture/Cryptography Extensions developed a system in support of "secure digital identities" for use by internal employees and contractors, as well as plan members. The system generated root, sub-CA, and end entity certificates, as well as their corresponding private RSA 2048 keys, and stored them on low cost USB based Yubikeys. The Yubikeys ensured that the key material was "unextractable" but could still be used for signing and identity verification. Utilized both PKCS#11 and PIVA industry standard programming interfaces within the system

Artificial Intelligence Architect
Chatbots and Intelligent Virtual Assistants
Bristol Myers Squib

- Reviewed potential AI use cases, across business lines, including drug discovery, translational medicine, medical, clinical trials, legal, regulatory, manufacturing, and commercial, in order to identify those which had the greatest chance of being successfully implemented using currently existing technologies.
- Developed an AI Platform Reference Architecture, that effectively combined learning, reasoning, natural language understanding, machine vision, and speech recognition. The architecture was created to support an extremely wide range of Robotic Process Automation (RPA), operational decision management (ODM), machine learning (ML), natural language processing and understanding (NLP/NLU) use cases, on a single integrated platform.
- Authored a strategic roadmap for the use of artificial intelligence across the enterprise, looking outwards three to five years. The roadmap covers the phased adoption of the functional capabilities described with the AI Platform Reference Architecture.
- Began work on the first phase of the physical implementation of the AI Platform Reference Architecture, using Amazon EC2 instances and S3 storage
- Built deep learning models that automated the previously manual process of identifying adverse drug events. Cleansed, embellished, and joined text extracted from live chat conversations, as well as incident reports taken directly from health care providers (HCPs). Using Tensorflow, and the Keras framework, built and tested binary classification models. The deep learning models predicted the likelihood of an adverse event occurring, solely from the text of a chatbot conversation. The deep learning models effectively combined both Long Short Term Memory (LSTM) networks along with One Dimensional Convolutional Neural Networks (CNN), to achieve a level of accuracy comparable to humans.
- Utilized Amazon AWS EC2 instances, and S3 storage to build a working prototype of the Adverse Event Detection System (AEDS).
- Assessed the potential of using Prolog based definite clause grammars (DCG's) for natural language understanding (NLU), in order to successfully address the issue of semantic, anaphoric, and pragmatic ambiguity that naturally exist within utterances, making them difficult to accurately classify using deep neural networks
- Provided technical oversight on multiple ongoing conversational AI initiatives, including those being built using AWS Lex, Microsoft Bot framework, and Verint's IVA platform

- Received extensive hands-on training for AWS cloud based service offerings, including Lex, Sagemaker, and Lambda.

Enterprise Data and Analytics Architect
Barclays Bank

Consulting member of global enterprise architecture team

- Developed the Future State Data and Analytics architecture for the US charge card division
- Defined the logical layers of the data lake architecture, including data storage, compute, ingestion, and AI/machine learning
- Defined the physical architecture, including the selection of all infrastructure, components, and tools, some of which were not included within the major Hadoop distributions
- Performed a gap analysis to determine what infrastructure and software components were missing from a machine learning stack perspective
- Developed a roadmap by which to migrate the charge card division's core marketing offer engine off a legacy SAS environment, and onto one based upon bid data and machine learning technology
- Prototyped a next generation intelligent offer engine, based upon cognitive computing technologies, including a fully conversant intelligent assistant
- Built custom Tensorflow deep learning neural networks as well as H2O models, that predicted the likelihood of a customer responding to a balance transfer offer, as well as determined the customer's propensity to spend on their charge card over the next 90 days
- Performed a complete assessment of the existing Hadoop data lake environment, including all projects currently in production, and made a set of concrete recommendations to Barclays on how to derive the most value from the environment
- Assessed the functional capabilities and technical design of the existing marketing offer engine, which was a batch oriented system built using SAS and a homegrown rules engine.
- Designed a new intelligent offer engine, using common off the shelf (COTS) AI infrastructure, tools, and libraries, that would provide personalized product recommendations across divisional lines, making the best possible offer based upon events that were taking place in real-time.
- Developed a prototype intelligent virtual assistant (IVA) that would enable the marketing team to author product eligibility rules via plain English conversation. In addition, the IVA would also provide advice to data scientists on how to improve their deep learning models.
- The prototype of the new AI based system processed cardholder 'life' events in real-time, allowing offer eligibility to change dynamically, and hence the recommendations being made.
- Authored eligibility and offer logic in Prolog, which was leveraged by the cognitive engine at application runtime