# **Yang** Bai

baiyang94@ufl.edu | 352 284 2346 | Portfolio | Linkedin



SUMMARY

Researcher designing **Machine Learning** algorithms for structured and unstructured data with applications in **Knowledge Graphs** and **Natural Language Processing**.


## EDUCATION

### UNIVERSITY OF FLORIDA

**Gainesville, FL, USA**
**2019 - 2023(Expected)**
PhD. Computer Science
GPA: 3.96/4.0

**2016-2018**
MSc. Electrical and Computer Eng.
GPA: 3.5/4.0

### SICHUAN UNIVERSITY

**2012-2016 | Chengdu, China**
BSc. Micro Electronics

## SKILLS

### LANGUAGES

| | |
|---|---|
| Python : | *Expert* |
| Java : | *Expert* |
| SQL/SPARQL : | *Expert* |
| C/C++ : | *Intermediate* |
| JavaScript : | *Intermediate* |
| F# : | *Intermediate* |

### TOOLS

| | |
|---|---|
| PyTorch | TensorFlow |
| Keras | Scikit-learn |
| Hugging Face | NLTK |
| SciPy | Pillow |
| OpenCV | OpenIE |
| Matlab | NumPy |
| Pandas | Oracle DB |
| REST API | Flask |
| Docker | Akka.NET |
| Git | Linux |
| Google Test | JUnit |

## COURSEWORKS

Elements of Machine Intelligence
Deep Learning for Computer-Graphics
Applied Machine Learning
Trustworthy Machine Learning
Distributed Operating System
Programming Language Principles
Database Management System
Database System Implementation
Analysis of Algorithms
Advanced Data Structures
Computer Networks

## WORK EXPERIENCE

### University of Florida — Graduate Student Researcher | Sep. 2019-present

- Active Interpretation of Disparate Alternatives
  - Individual contributor and team lead in the DARPA-sponsored project "Active Interpretation of Disparate Alternatives(AIDA)", an alternative hypotheses search engine over event-enteric knowledge graphs. **Our system achieves top performance at the NIST TAC SM-KBP2020 evaluation.**
  - Developed a two-level graph searching algorithm to explore knowledge graphs at both mention-level and cluster-level improving the final F1 score by **25%**.
  - Developed a graph clustering algorithm to differentiate alternative hypotheses by measuring both structural and semantic distance between candidates, which improves the original cluttering quality(v-measure) by **20%**.
- **Multi-answer open-domain question answering with controversial stance mining for query-based large-scale check-worthy claim detection**
  - Constructed a benchmark dataset using the Twitter API with three annotators.
  - Designed the new evaluation metrics, data schema, and annotation instructions.
  - Developed an annotation tool with a user-friendly UI.
  - Developed an end-to-end pipeline to evaluate how different modules along the pipeline (including information retrieval, machine reading comprehension, and distinct answer selection module) affect the final performance.

### Nokia Bell Labs — Machine Learning Intern | Jun. 2022-Aug. 2022

- Proposed and implemented a retrieval-based framework to ease ticket root cause analysis by retrieving the most relevant log lines from the attached log files (10-100M log lines/ticket) given ticket information.
  - Conducted data cleaning, processing, visualization, and analysis on massive time-series semistructured system-level log corpus.
  - Developed a dense log retrieval system that finetunes self-pretrained tickets and log encoders through an adaptive multi-model machine learning framework.
  - The best model outperforms a BM25 baseline model by **16.1%**.

## SELECTED PUBLICATIONS **Google Scholar**

More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering
**Yang Bai**,D.Wang **arXiv 2021**

GAIA at SM-KBP 2020 - A Dockerized Multi-media Multi-lingual Knowledge Extraction, Clustering, Temporal Tracking and Hypothesis Generation System
M.Li,...,**Yang Bai**,..., D.Wang **TAC 2020**

GAIA at SM-KBP 2019-A Multi-media Multi-lingual Knowledge...
M.Li,...,**Yang Bai**,..., D.Wang **TAC 2019**

# COURSEWORK

## GRADUATE

Advanced Machine Learning(A)
Network Sciences (A)
High Dimensional Data (A)
Numerical Linear Algebra (A)
Numerical Optimization (A)
Stochastic Methods (A)
Theoretical Statistics I & II (A,B)
Stat Decision Theory (A)
Spectral Estimation (A)
Image Processing (A)
Artificial Intelligence* (A)
Machine Learning* (A)
Deep Learning+ (A)
Neural Networks for Machine-Learning^ (A)
(*Stanford online course, +Udacity, ^Coursera by Geoffrey Hinton)

## UNDERGRADUATE

Numerical Optimization (grad)
Time Series Analysis (grad)
Linear Algebra
Communication Systems
General Mathematics
Digital Communication
Digital Signals and Systems
Digital Signal Processing
Algebra I and II
Topology
Mathematical Analysis
Wireless Communication

# PUBLICATIONS **Google Scholar**

### Hotel2vec: Learning Attribute-Aware Hotel Embeddings with Self-Supervision
**Ali Sadeghian**, S. Minaee, I. Partalas, D.Z. Wang, B. Cowan. To be submitted to KDD 2020.

### DRUM: End-To-End Differentiable Rule Mining On Knowledge Graphs
**Ali Sadeghian**, R. Armandpour, P. Ding, D.Z. Wang. NeurIPS 2019.

### Measuring Impact of Climate Change on Tree Species: analysis ...
Hyun Choi, Sergio Marconi, **Ali Sadeghian**, Ethan White, Daisy Zhe Wang. TCCML-NeurIPS19.

### SoPhie: An Attentive GAN for Predicting Paths Compliant to Social ...
A. Sadeghian, V. Kosaraju, **Ali Sadeghian**, N. Hirose, S. Savarese. CVPR 2019.

### Mining Rules Incrementally over Large Knowledge Bases
X. Zhou, **Ali Sadeghian**, D.Z. Wang. SDM19.

### Automatic semantic edge labeling over legal citation graphs
**Ali Sadeghian**, L. Sundaram, D. Wang, W. Hamilton, K. Branting, C. Pfeifer. Journal of A.I. and Law 2018.

### Machine learning approaches in GIS-based ecological modeling of ...
A. Mollalo, **Ali Sadeghian** and et. al. Journal of Acta Tropica 2018

### Temporal Reasoning Over Event Knowledge Graphs.
**Ali Sadeghian**, M. Rodriguez, D.Z. Wang, A. Colas. KBCOM 2018-best paper award.

### Automatic Target Recognition Using Discrimination Based On Optimal Transport.
**Ali Sadeghian**, D. Lim, J. Karlson, J. Li, ICASSP 2015.

### Energy-Aware Adaptive bi-Lipschitz Embeddings.
**Ali Sadeghian**, B. Bah, V. Cevher, SampTA 2013.

# ADDITIONAL EXPERIENCE

**LIONS at EPFL** **Summer 2012 | Switzerland**
| Intern Research Assistant, Professor Volkan Cevher
- We introduced a novel scalable learning algorithm and provided a rigorous estimation guarantee by leveraging game theoretic tools.

**Institute of Network Coding at CUHK** **Summer 2011 | Hong Kong**
| Intern Researcher, Professor Sidharth Jaggi
- Secure network coding: We designed an algorithm that achieves a higher throughput rate in networks with adversaries.

### Finding the Optimal Boundary in statistical models
- Analytical methods for finding the optimal boundary in statistical models. Providing a new approach based on image processing techniques.

### Automatic Instrument sound separation in polyphonic music
- We designed a Recursive Neural Network to convert a polyphonic recording of a music piece into separate recordings for each instrument.

### Yelp challenge: Modeling the impact of businesses on each other
- We used machine learning techniques to observe how opening of a new business affected customer ratings for older business in the same area.

### Software defined radio
- I implemented a two way FDD communication link with pulse shaping, AGC and ARQ using C++ on N2x0 USRP.

### Teaching Assistant
- Stochastic Processes, University of Florida. Fall 2017.
- Digital Communication course, Sharif U.T. spring 2012.

### Lecturer & Organizer: Iranian national Gold medalists
**2011 | Tehran, Iran**