Aron Szanto

– New York, NY –
(+1) 914 815 5728 • aszanto9@gmail.com • aronszanto.com

Education

Harvard University
Cambridge, MA, Applied Mathematics, Computer Science and Economics *2014–2018*
Highest Honors, Phi Beta Kappa.
Fay Prize for Most Outstanding Work in Any Field, Hoopes Prize for Outstanding Thesis. GPA 3.9.
Completed courses for B.A. and M.S. degrees simultaneously. Graduate coursework in Machine Learning, Deep Learning, Artificial Intelligence, Natural Language Processing, Economics, Data Systems, Cryptography, and Security. Research in machine learning for fake news identification, computer vision, natural language processing, multi-agent systems, high-performance NoSQL databases, and privacy and security.

Research

Honors Thesis, Harvard University
Fake News Detection via Content-Blind Learning *2017–2018*
Developed novel graph kernel technique to identify fake news articles at 84% accuracy based solely on the pattern of travel through social networks. Awarded prizes for best senior thesis and for best work in any subject in 2018.

Department of Computer Science, Harvard University
Visual Question Answering *2018*
Introduced deep learning method to ascertain whether an artificial intelligence has learned both language and vision for a multimodal visual question answering task. Used new system to beat state-of-the-art VQA benchmark. Presented paper at KDD.

Data Privacy Lab, Harvard University
Airbnb Host Re-Identification *2017*
Developed statistical algorithm to re-identify private Airbnb hosts using public voter records. International news coverage of work, blog post syndicated by Hacker Noon and #1 on Y Combinator and Reddit. Paper in *Journal of Technology Science*, Q3 2018.

Research summaries, code links, and additional projects on following page

Work Experience

Machine Learning Engineer
***Kensho Technologies**, New York, NY* *July 2018–Present*

- Responsible for machine learning and natural language processing research related to healthcare, finance, and national security.

Product Engineering Manager
***MarketFactory**, New York, NY* *January 2016–April 2018*

- Manage research and development of next-generation products at high-growth NY Fintech SaaS startup
- Lead projects in machine learning, distributed data system architecture, and big data analysis
- Supervise algorithmic and engineering development for 1.2PB database compression and migration, resulting in compression ratios of >95% and performance speedup up to 100x

Technical Skills

Comfortable: Python (inc. Numpy, Scipy, Pandas), C++, C, Machine Learning, Neural Networks and Deep Learning (scikit-learn, pytorch), Java, Elasticsearch, NetworkX, Git, Linux, SQL, LaTeX

Familiar: TensorFlow, JavaScript, Cryptography, HTML/CSS, Go, Swift, Assembly

Languages and Interests

English and Hungarian: Fully bilingual. **Spanish:** Conversational

Orchestral cellist and competitive ultimate frisbee player.

Recent Research and Papers

Fake News Detection on Twitter Networks
A new method in artificial intelligence that detects fake news at 84% accuracy based on the paths that rumors take as they travel through Twitter. Crucially, false stories are identified not by considering linguistic or user-identifying content, but instead by analyzing the shape of the social networks surrounding news articles. This is the first application of predictive analytics to the largest collection of fake news stories and associated social networks ever assembled. With its high accuracy, this work represents the state of the art for fake news identification in this domain. This contribution also has a unique practical impact: while existing models can be deceived by a single fake news writer, attacking this system requires concerted action to trigger the reshaping of a social network. (PyTorch, NetworkX, scikit-learn, Python) Winner of Fay, Hoopes, and Blumberg Prizes. Profile in *The Harvard Crimson*: `https://goo.gl/CGiRiY`

Counterexamples in Visual Question Answering
Deep learning method to ascertain whether an AI has learned both language and vision in a visual question answering task. This work developed a plug-and-play technique to test any state-of-the-art VQA model by asking it to provide a counterexample image. For example, when asked what color a fire hydrant is in a picture, a standard VQA model would respond "red", but our extension challenges it also to display a picture of a black hydrant. We leveraged the new method to beat the existing record performance on the counterexample task linked to the seminal VQA 2.0 dataset. Paper accepted to Deep Learning Day at KDD, August 2018 in London, UK. (PyTorch, scikit-learn, Python)
Research Paper: `https://arxiv.org/pdf/1806.00857.pdf`
Code: `https://github.com/aronszanto/VQA-Counterexamples`

Artificial Intelligence and GitHub Collaboration
Information systems for large-scale collaboration on open source projects. Used LSTM autoencoders and hidden Markov models to analyze historical GitHub data, developing a model that could predict a user's future contributions to a repository as well as determine whether a given project will be successful. These findings lay the foundation for an AI product manager that assists collaboration by actively managing a project's contributors, organizing them into subteams, finding relevant users to bring into the project, and shaping the work that users do on the project to maximize its success. (TensorFlow, Keras, Python)
Paper: `https://goo.gl/7MNMDj`
Code: `https://github.com/aronszanto/DeepLearning-GitHub`

Airbnb Security Exploit
Airbnb claims that they protect the privacy of their hosts using a probabilistic location fuzzing mechanism. I developed a statistical algorithm to reidentify Airbnb hosts using public voter records, demonstrating that Airbnb's platform is not identity-secure. The model was able to reidentify over 40% of a random sample of Airbnb listings and present an analysis of risk factors for the reidentification of a given listing.
Blog post syndicated by Hacker Noon, reached #1 on Y Combinator and Reddit: `https://goo.gl/vD1NGi`
News coverage in International Business Times and El Diario (`https://goo.gl/PUiAu1`; `https://goo.gl/izyd3r`). Paper to appear in The Journal of Technology Science in Q3 2018. Code and data publication pending.

Skiplist-Based LSM Tree
A high-performance NoSQL data system written entirely in native C/C++. It is optimized for transactional workloads with high rates of writes, and supports up to 6 million updates per second. Leverages the probabilistic data structures skiplists and Bloom filters as well as cache- and disk-aware algorithmic techniques. (C, C++)
Paper: `https://goo.gl/bQict2`
Code: `https://github.com/aronszanto/sLSM-Tree`

Secure Execution in JavaScript over WebAssembly
Built first WebAssembly virtual machine that runs in native JavaScript and implemented a novel taint tracking system that allows a user to run untrusted WA code while monitoring the flow of sensitive data through the application. Proved mathematical security and correctness guarantees and demonstrated runtime efficiency at scale. (JavaScript, WebAssembly)
Paper: `https://goo.gl/b5uDZq`
Code: `https://github.com/aronszanto/wasm-taint-tracking`