Mathias Lecuyer

PhD candidate in Computer Science at Columbia University

Education

2013–Current **PhD candidate in Computer Science**, *Columbia University*, New York, *GPA 4.0*.
- Areas: Distributed Systems, Statistics, Machine Learning, Privacy, Transparency.
- Advisors: Roxana Geambasu and Augustin Chaintreau.

2011–2012 **Master of Science in Computer Science**, *Columbia University*, New York, *GPA: 3.90*.
- Graduate research project with professors Augustin Chaintreau and Roxana Geambasu.
- Classes focused on Systems and Machine Learning.

2008–2011 **Ingénieur diplômé**, *Ecole Polytechnique*, Paris France, *GPA: 3.90*.
France's top-ranking *Grande Ecole* for Science and Engineering.
- Multidisciplinary training in Computer Science, Economics, Physics and Biology.
- Major in Computer Networks.

2006–2008 **Classes préparatoires in Science**, *Lycée Louis-le-Grand*, Paris France, *GPA: 4.0*.
After high-school, intensive two-year preparatory course in mathematics and physics for competitive entrance to France's leading schools of Science and Engineering.

Work Experience

Summer 2017 **Research Intern**, *Microsoft Research*, NYC.
- Applied Reinforcement Learning (RL) to Systems, with a focus on off-policy evaluation.
- Build an nginx prototype to deploy RL policies [1].

2012–2014 **Teacher Assistant**, *Columbia University*.
- Teacher Assistant in Roxana Geambasu's Distributed Systems class.
- Teacher Assistant for python in Mark Hansen's Data Journalism class.
- Teacher Assistant in Augustin Chaintreau's Computer Networking class.

Summer 2013 & 2014 **Freelance**, *Milky & Floatingapps*.
- Built a rental data visualization web-app, with a Rails API backed by Cassandra and Elasticsearch, and a frontend in d3 and React.
- Built the second screen iOS application *Le Grand Journal* of a major French TV Channel Canal+, with a real time twitter feed for each program, and replays of previous shows.
- Built an iOS application for email visualization as timelines, in Objective-C.

2009–2012 **Cofounder**, *Pionid*, Paris, France.
Cofounder of a mobile and web startup. Built a group messaging app, and developed a game on iOS.

Apr–Jun 2011 **Junior Consultant**, *Atos*, Paris, France.
Junior IT consultant. Worked on an archiving solution as a service.

Awards

2012-2013 **Magic Grant**, *Dispatch*, Brown Institute.
Built and deployed a secure reporting tool for conflict areas. Worked in an interdisciplinary team advised by Augustin Chaintreau (Columbia CS), Susan McGregor (Columbia Journalism), and Chris Haseman (Tumblr).

2011 **Fellowship**, *Bourse Carnot*.
An excellency fellowship for French students in the Unites-States, focused on Research and Entrepreneurship.

549 Riverside Drive, #5A – New York, NY 10027
+1 (917) 355 4318 • mathias@cs.columbia.edu
http://mathias.lecuyer.me

Research Theme

Privacy in a Data-Driven World.

Data has become the principal asset of the Internet era. While this data offers unique opportunities to improve personal and business effectiveness, it also poses serious risks to users' privacy, and to organizations, by exposing extensive data stores to external and internal attacks. In my research, I build tools and design mechanisms that leverage statistics and machine learning to: increase the current Web's transparency by revealing how personal data is being used; and enable a more rigorous and selective approach to big data collection, access, and protection, to reap its benefits without imposing undue risks.

Research

Selectivity **Minimizing Data Exposure in Machine Learning Applications**.

Challenging the common practice in both private and public sectors of collecting vast quantities of personal information, I ask whether it is possible to build data-driven systems that are more selective with the data they collect. To explore this question I built Pyramid [2], a data management system that leverages training set minimization techniques to reduce data exposure in ML applications. More precisely Pyramid uses count-based featurization to summarize past data before it is archived in cold storage. The counts, kept differentially private, are used with a small amount of recent observations, called the hot data, to train ML models. Using this technique, as well as system mechanisms to reduce the impact of differentially private noise, Pyramid is within 4% of previous models' accuracy while training on, and thus exposing, less than 1% of the raw data. This way, ML based applications can reap the benefits of big data without undue risks.

Transparency **Data Use Transparency Infrastructure**.

To add transparency to data uses on the Web, I am building a series of scalable, generic, and reliable tools to detect data flows within and across web services. My initial system, XRay [6], offers a first system design and theoretical building blocks to detect the use of digital personal data for targeting and personalization. The key insight in XRay is to infer targeting by correlating user inputs (such as searches, emails, or locations) to service outputs (such as ads, recommendations, or prices) based on observations obtained from user profiles populated with different subsets of the inputs. My latest tool, Sunlight [4], leverages rigorous statistical methods to determine the *causes* of online targeting at great scale and based on solid statistical justification.

Impact of the Sharing Economy.

In a recent controversy about Airbnb's impact on cities, three reports from public institutions and lobbying groups arrived at opposite conclusions with seemingly contradictory facts about the occupancy distribution. To inform this debate I implemented a reliable way to estimate booking rates and revenues of the platform's hosts. I evaluated this method by comparing its results to metrics previously released by Airbnb or the NY Attorney General from subpoenaed data. With a complete view of the distribution of revenue, I found that previous claims that seemed at odds are all explained by a variant of the "inspection paradox" [3].

Synapse **Heterogeneous-database replication**.

Synapse [5] is an heterogeneous-database replication system, which lets programmers of complex, multi-service Web applications share data across services running on very distinct database engines, in real time, and with solid consistency semantics. We deployed Synapse at an NYC startup.

Dispatch **Secure and private reporting for citizen journalists**.

The rise of smartphones and social media has transformed journalism, allowing the public to skip the middleman and get news "straight from the source." These networks, however, are inherently fragile, consisting of easily targeted devices and relatively centralized systems that authorities may block or shut down. Our response to this pressing issue is Dispatch [7], which allows journalists and citizen reporters to publish their work securely,using an identity-based encryption scheme, and provides censorship-tolerant functionality with Bluetooth message passing when no other connection is available.

Talks and Visibility

March 2016 **Sunlight media coverage**.

My project Sunlight was mentioned in a special report from The Economist, "The data republic": `http://www.economist.com/news/special-report/21695195-safeguard-democracy-use-data-should-be-made-transparent-possible-data`

Summer 2015 **Interest from the FTC**.

My research has spawned interest from the Federal Trade Commission, who has invited my group to present our research to them.

549 Riverside Drive, #5A – New York, NY 10027
+1 (917) 355 4318 • mathias@cs.columbia.edu
http://mathias.lecuyer.me

October 2014 **Princeton Web Privacy and Transparency workshop**.
I was invited to the closed-section of the Princeton Web Privacy and Transparency workshop to discuss the future directions of the field.

August 2014 **XRay media coverage**.
XRay was featured in The New York Times Bits blog: `http://bits.blogs.nytimes.com/2014/08/18/xray-a-new-tool-for-tracking-the-use-of-personal-data-on-the-web/`.

July 2014 **UW MSR Summer Institute research perspective talk**.
Research perspective talk about "XRay: Enhancing the Web's Transparency with Differential Correlation" at the Summer Institute cosponsored by the University of Washington and Microsoft Research.

Spring 2012 **Talk at the Spring Journalism & Technology Breakfast**.
Talk about Dispatch at Columbia's Journalism School semi-annual Journalism & Technology Breakfast.

Publications

[1] Mathias Lécuyer, Joshua Lockerman, Lamont Nelson, Siddhartha Sen, Amit Sharma, and Aleksandrs Slivkins. Harvesting randomness to optimize distributed systems. In *HotNets*, 2017.

[2] Mathias Lécuyer, Riley B. Spahn, Roxana Geambasu, Tzu-Kuo Huang, and Siddhartha Sen. Pyramid: Enhancing selectivity in big data protection with count featurization. In *Proceedings of the IEEE Symposium on Security and Privacy (Oakland)*, 2017.

[3] Mathias Lécuyer, Max Tucker, and Augustin Chaintreau. Improving the transparency of the sharing economy. In *Proceedings of the 26th International World Wide Web Conference (WWW)*, 2017.

[4] Mathias Lécuyer, Riley B. Spahn, Giannis Spiliopoulos, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2015.

[5] Nicolas Viennot, Mathias Lécuyer, Jonathan Bell, Roxana Geambasu, and Jason Nieh. Synapse: New data integration abstractions for agile web application development. In *Proceedings of the European Conference on Computer Systems (EuroSys)*, 2015.

[6] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. XRay: Increasing the web's transparency with differential correlation. In *Proceedings of the USENIX Security Symposium*, 2014.

[7] Kanak Biscuitwala, Willem Bult, Mathias Lécuyer, T. J. Purtell, Madeline K. B. Ross, Augustin Chaintreau, Chris Haseman, Monica S. Lam, and Susan E. McGregor. Weaving a safe web of news. In *SNOW*, 2013.

549 Riverside Drive, #5A – New York, NY 10027
+1 (917) 355 4318 • mathias@cs.columbia.edu
http://mathias.lecuyer.me