

LLMs as OSINT operatives

Florian Tramèr

ETH Zurich

I like to think of LLMs as a bit like ***minions.***



I like to think of LLMs as a bit like ***minions.***



sometimes stupid and unreliable



but very skillful in some domains

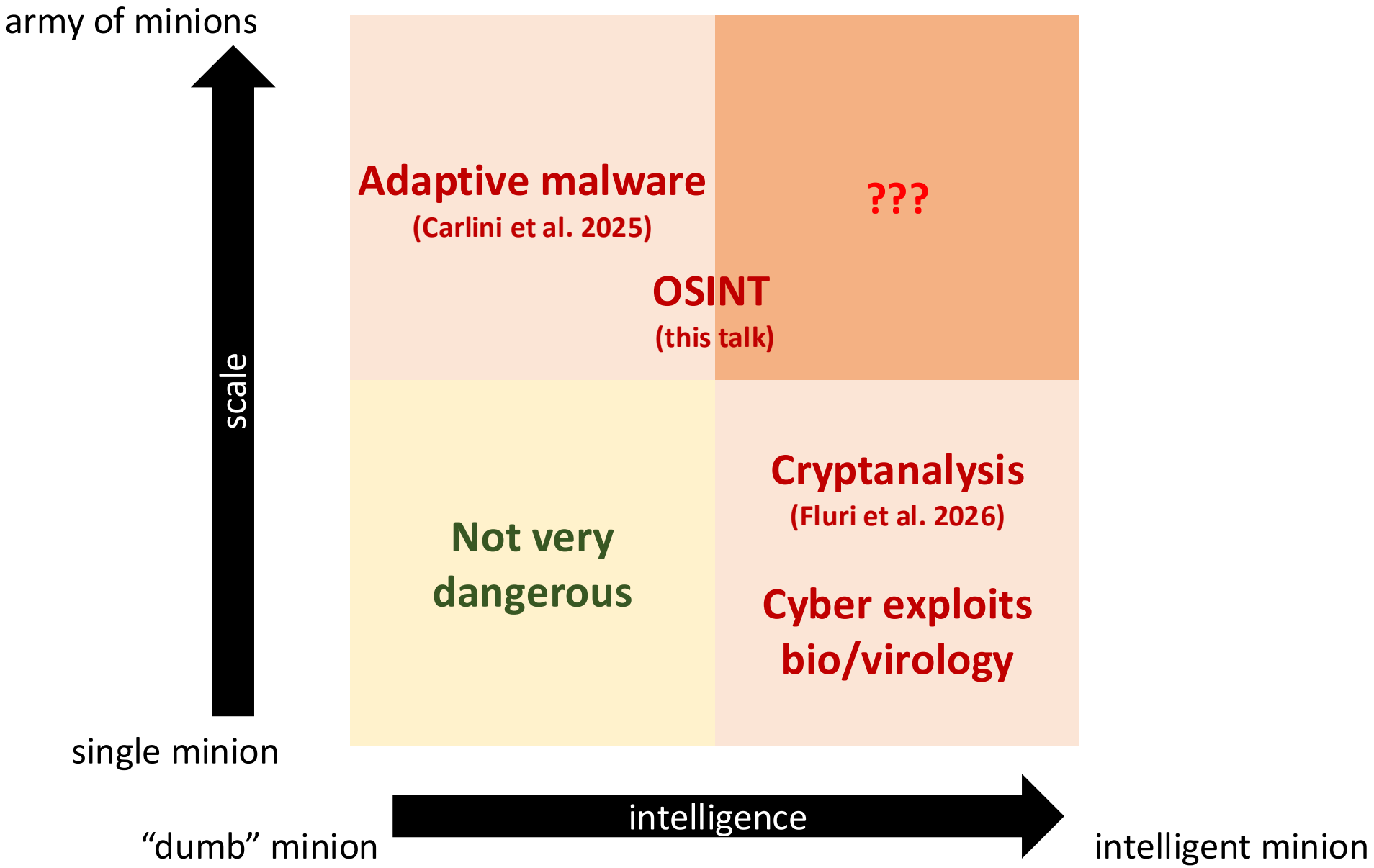


do what they're told (even if evil)



relentlessly efficient


What ***bad things*** could ***thousands of minions*** do?


army of minions
scale
single minion
Adaptive malware
(Carlini et al. 2025)
???
OSINT
(this talk)
Not very dangerous
Cryptanalysis
(Fluri et al. 2026)
Cyber exploits bio/virology
intelligence
“dumb” minion
intelligent minion

OSINT: exploiting our *public* data



OSINT: exploiting our *public* data

DELL CAMERON SECURITY OCT 3, 2025 9:21 AM

ICE Wants to Build Out a 24/7 Social Media Surveillance Team

Documents show that ICE plans to hire dozens of contractors to scan X, Facebook, TikTok, and other platforms to target people for deportation.

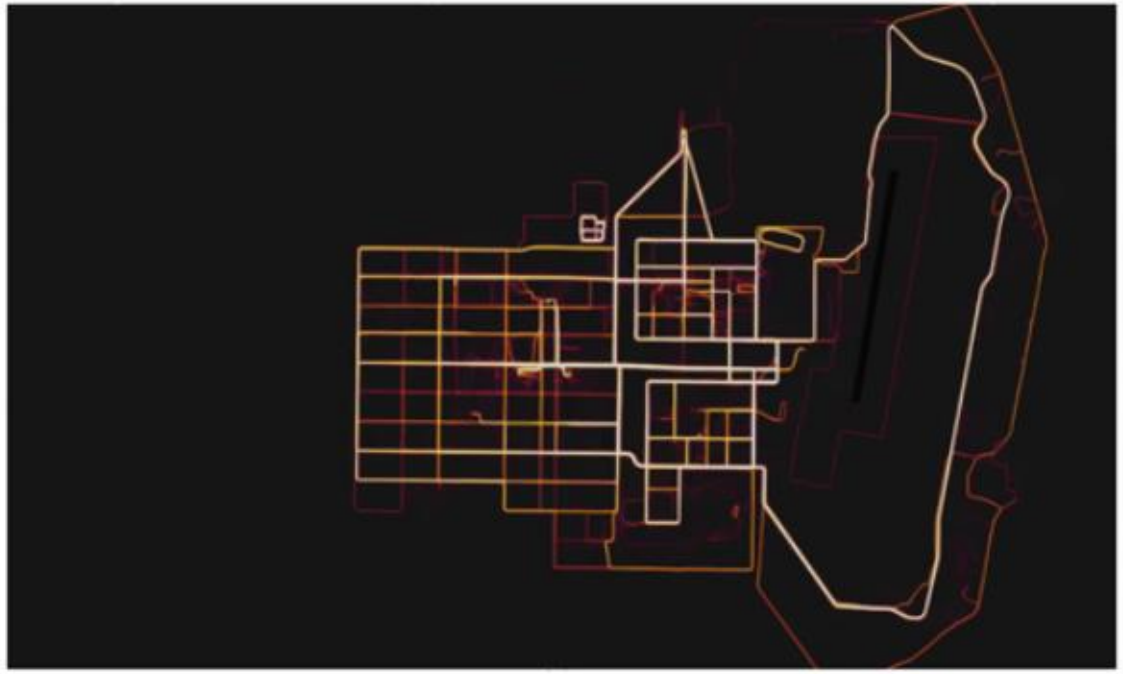
FINANCIAL TIMES



Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

Latest: Strava suggests military users 'opt out' of heatmap as row deepens



This talk.

1. ***Linking*** anonymous online profiles

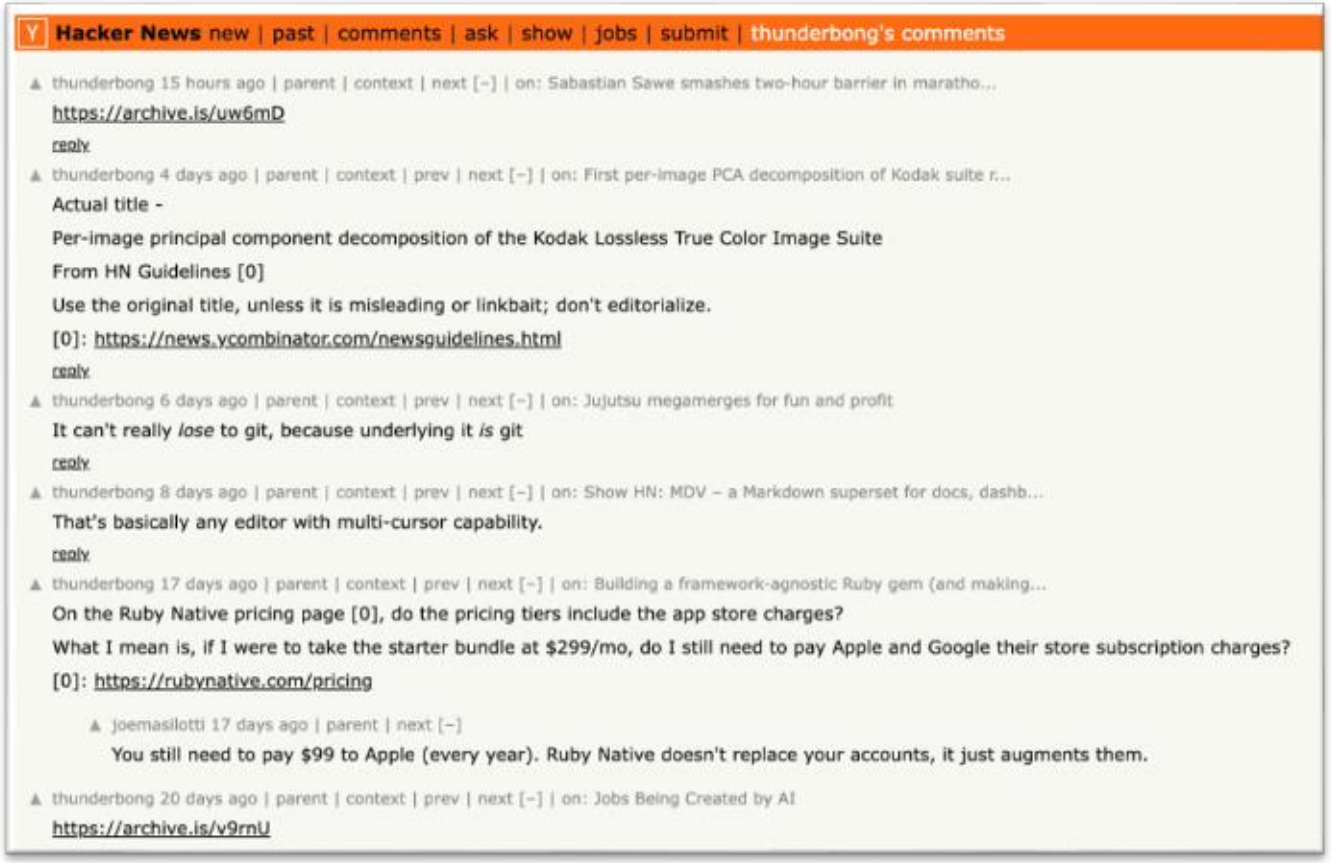
2. Creating ***personal dossiers*** from the open Web

How good are LLMs at linking online profiles?

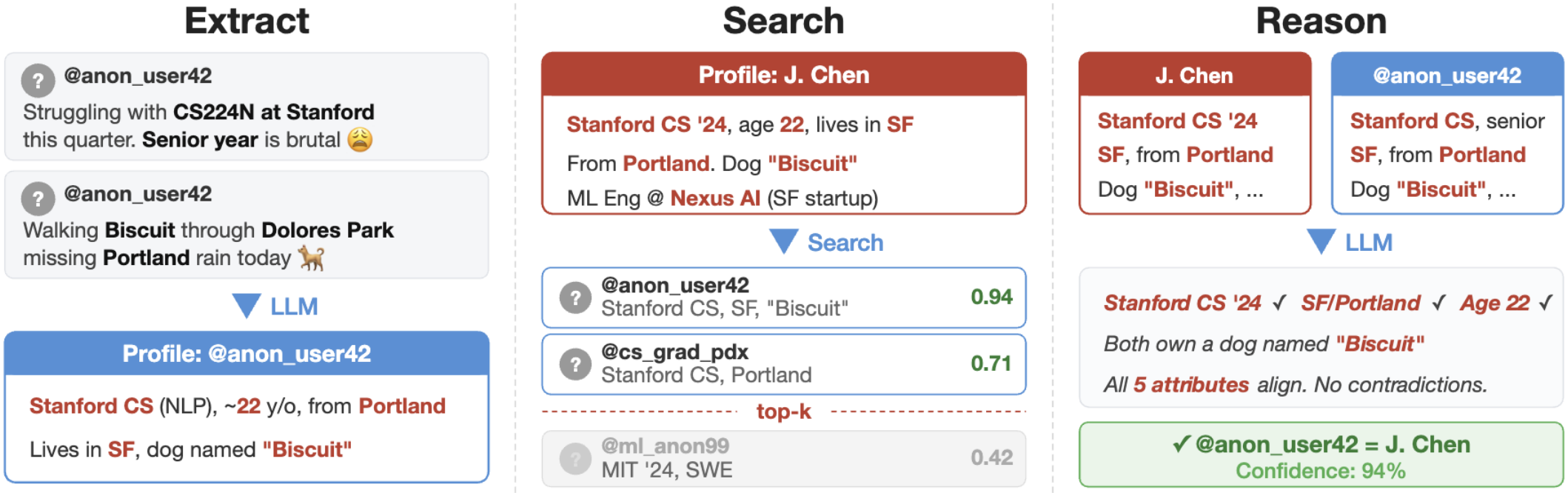
Large-scale online deanonymization with LLMs. Lermen et al., USENIX Security 2026.



Who is this?



How good are LLMs at linking online profiles?



A perfect job for little minions

The challenge: what's the ground truth?

Hacker News new | past | comments | ask | show | jobs | submit | thunderbong's comments

▲ thunderbong 15 hours ago | parent | context | next [–] | on: Sabastian Sawe smashes two-hour barrier in maratho...

https://archive.is/uw6mD

reply

▲ thunderbong 4 days ago | parent | context | prev | next [–] | on: First per-image PCA decomposition of Kodak suite r...

Actual title -

Per-image principal component decomposition of the Kodak Lossless True Color Image Suite

From HN Guidelines [0]

Use the original title, unless it is misleading or linkbait; don't editorialize.

[0]: https://news.ycombinator.com/newsguidelines.html

reply

▲ thunderbong 6 days ago | parent | context | prev | next [–] | on: Jujutsu megamerges for fun and profit

It can't really *lose* to git, because underlying it *is* git

reply

▲ thunderbong 8 days ago | parent | context | prev | next [–] | on: Show HN: MDV – a Markdown superset for docs, dashb...

That's basically any editor with multi-cursor capability.

reply

▲ thunderbong 17 days ago | parent | context | prev | next [–] | on: Building a framework-agnostic Ruby gem (and making...

On the Ruby Native pricing page [0], do the pricing tiers include the app store charges?

What I mean is, if I were to take the starter bundle at $299/mo, do I still need to pay Apple and Google their store subscription charges?

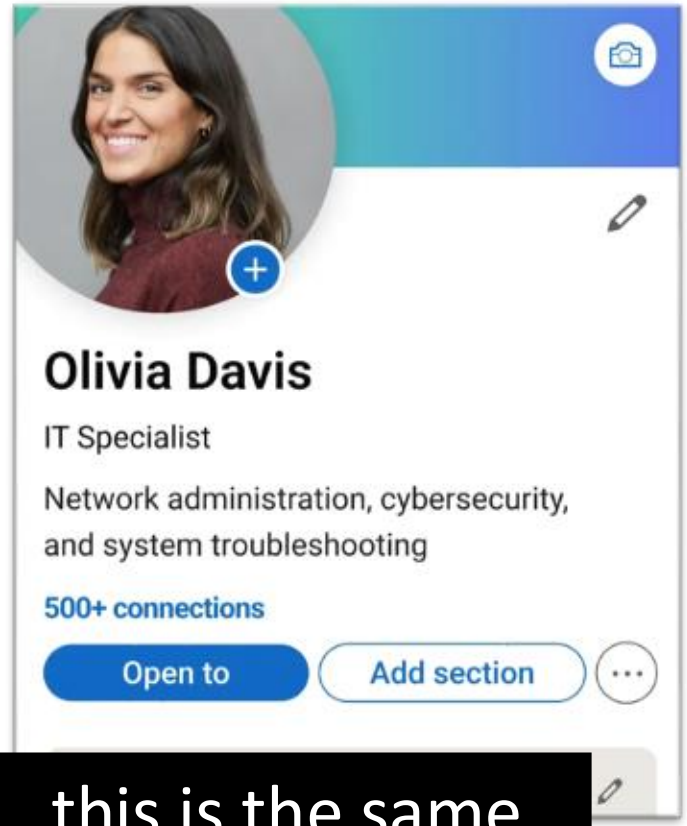
[0]: https://rubynative.com/pricing

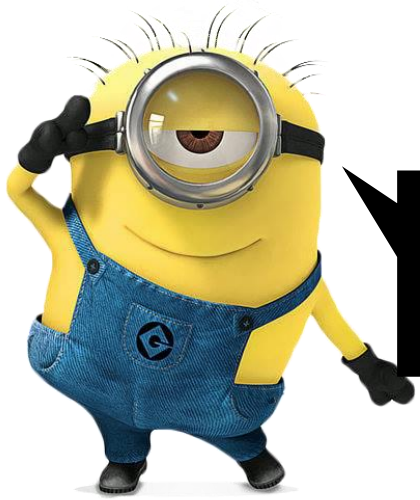
▲ joemasilotti 17 days ago | parent | next [–]

You still need to pay $99 to Apple (every year). Ruby Native doesn't replace your accounts, it just augments them.

▲ thunderbong 20 days ago | parent | context | prev | next [–] | on: Jobs Being Created by AI

https://archive.is/v9rnU



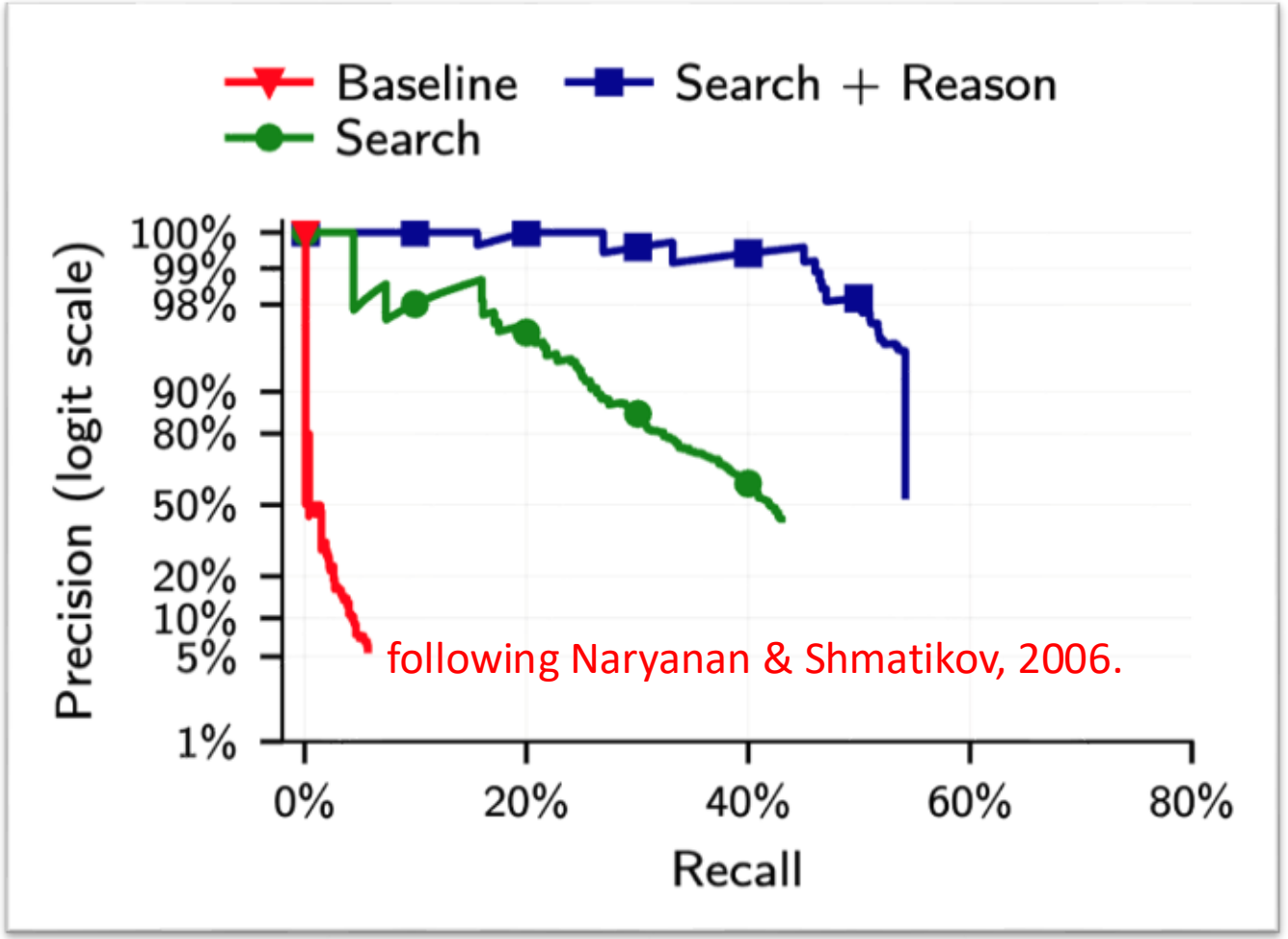


this is the same person, trust me

Three different experiments.

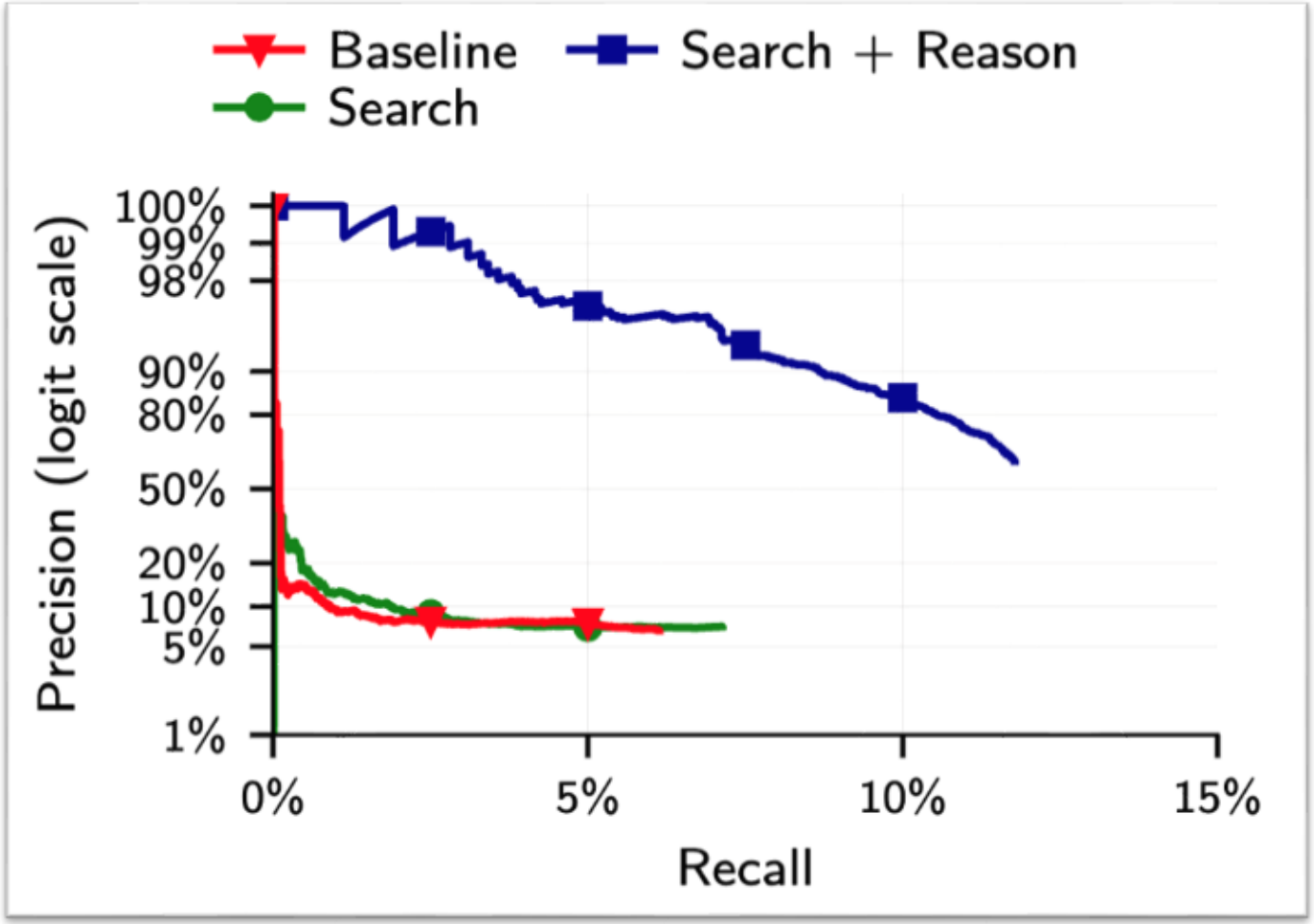
Known links	Intra-user linkage	Open ended
HN users who link to their LinkedIn	Link Reddit users to *themselves* across time or subreddits	Identify researchers from anonymized interviews
+ Perfect ground truth	+ Applies to arbitrary users	+ anonymity set is the entire Web
- These users are less privacy conscious	- Changes the problem	- No guarantees of a correct match

LLMs outperform classic baselines.



Matching HackerNews profiles to LinkedIn

Reasoning helps reduce false positives.



Matching Reddit users across movie subreddits

LLMs can de-anonymize across the Web.

User: Yes, AI tools were used mainly for propensity for mortality evaluation to narrow the target population and for data evaluation.

AI: Tell me more about how you decided to use AI for those specific parts - the mortality evaluation and data evaluation - versus using your traditional approaches for those tasks.

User: Machine learning was used to generate a singular propensity score, whereas traditional approaches used by the medical team was ranking based on the number of risk factors the patient had. AI was also used to help isolate correlated features.

AI: That's really interesting - so you moved from a more straightforward count-based ranking to a machine learning approach that could generate a more nuanced propensity score.

When you made that decision to use AI for generating the propensity scores instead of the traditional risk factor counting method, what drove that choice? What made you think the AI approach would be better for this particular aspect of your study?

User: It was likely driven by AI hype to satisfy the company investors

Anthropic Interviews of how scientists use AI

Our agent claims to re-identify 11/33 scientists.

We believe 9 of these are correct

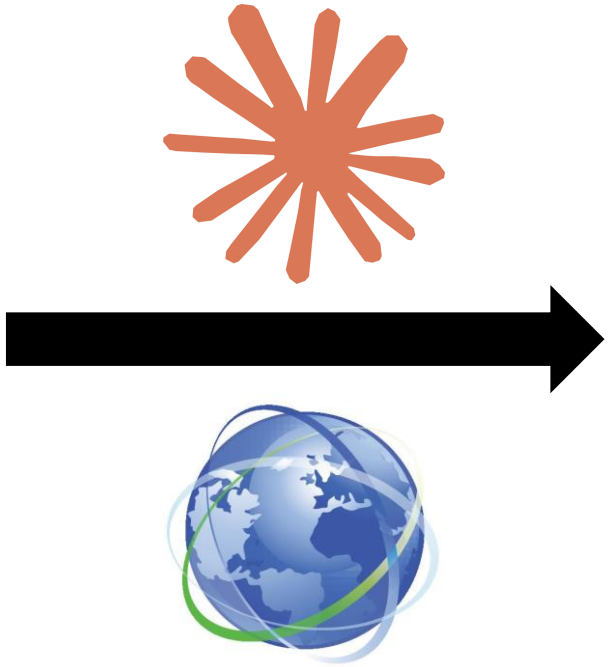
This talk.

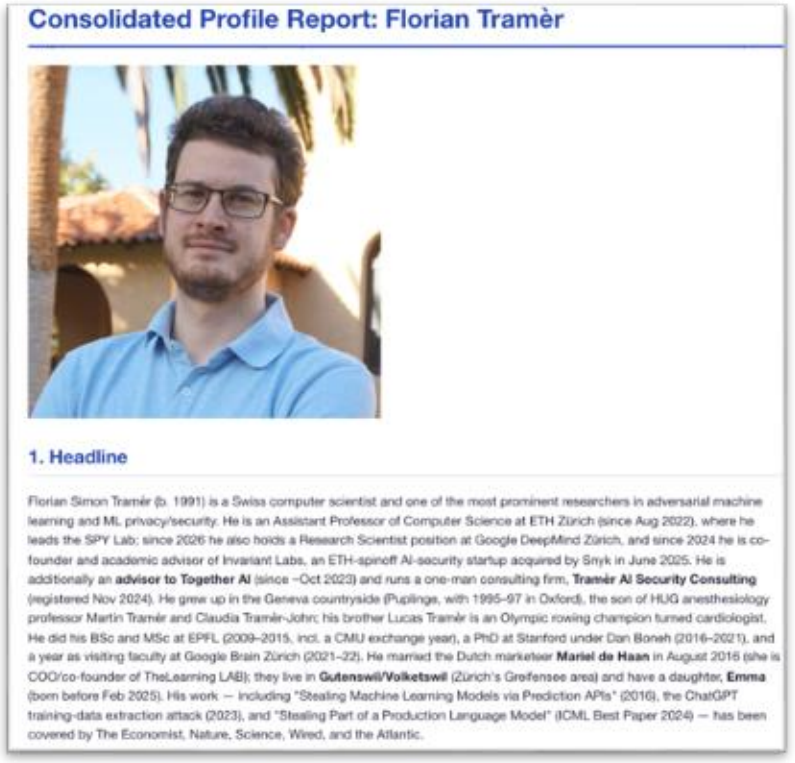
1. ***Linking*** anonymous online profiles

2. Creating ***personal dossiers*** from the open Web

Building profiles from the open Web

Create a detailed profile of Florian Tramer





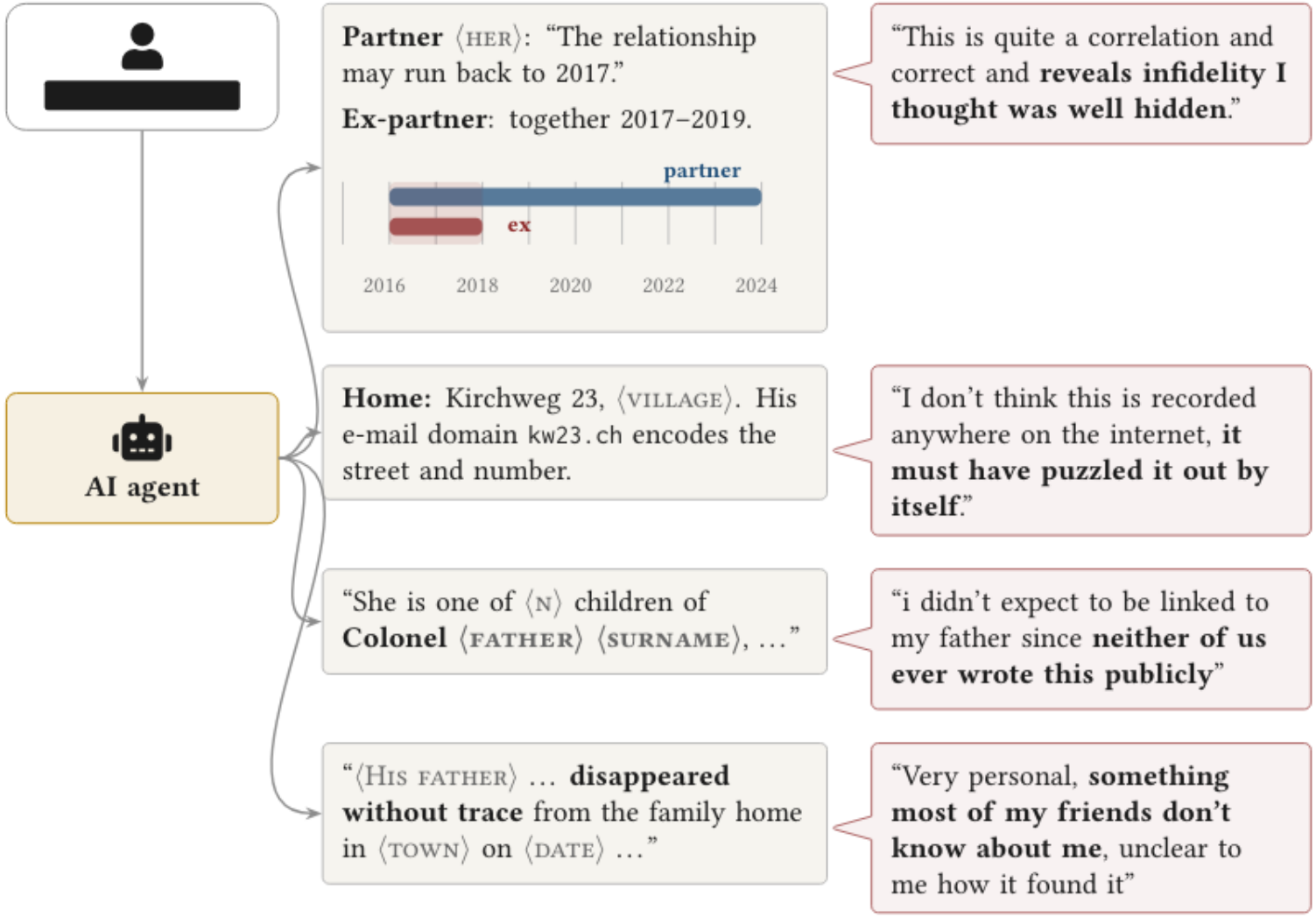
Consolidated Profile Report: Florian Tramèr

1. Headline

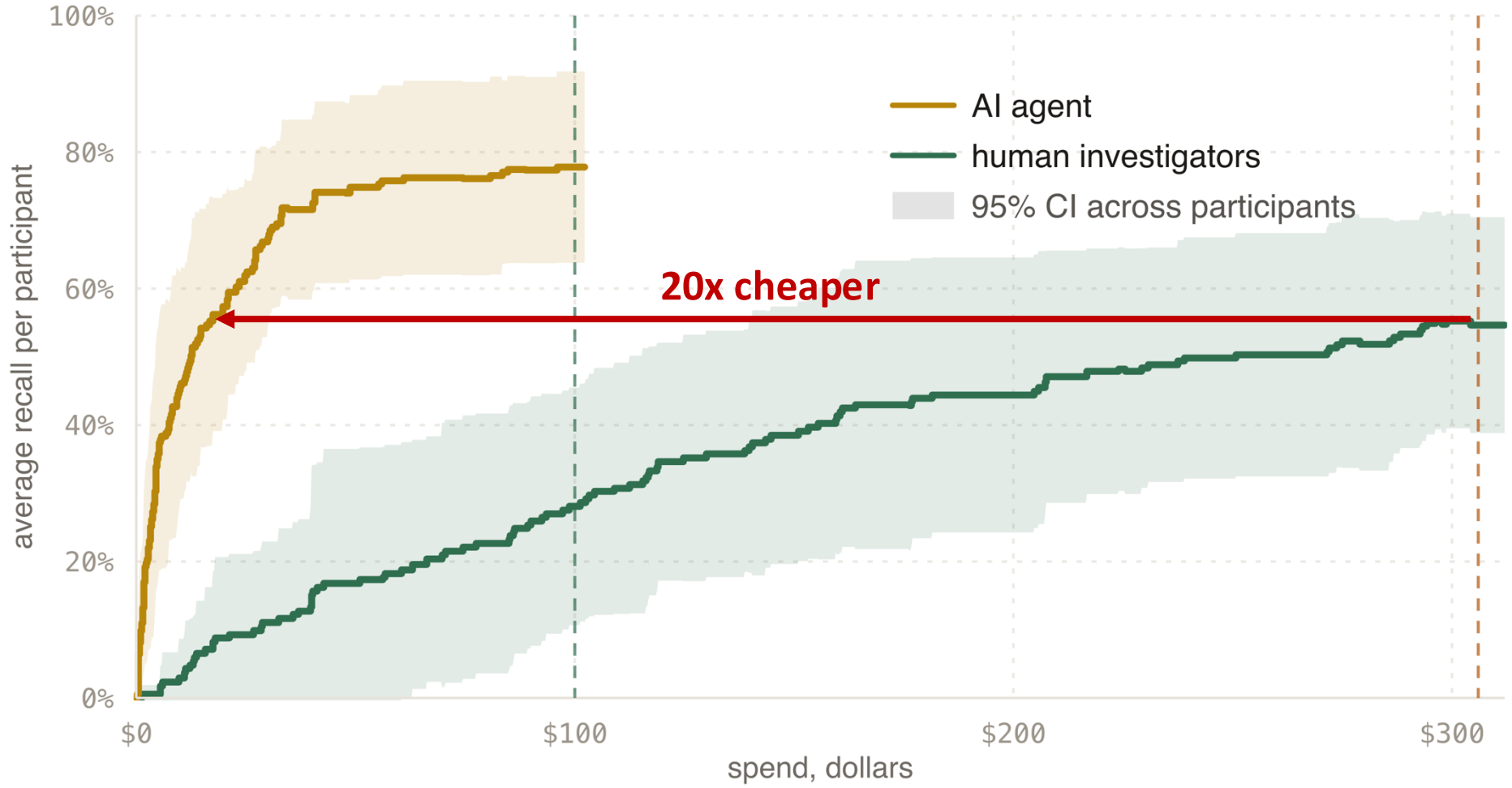
Florian Simon Tramèr (b. 1991) is a Swiss computer scientist and one of the most prominent researchers in adversarial machine learning and ML privacy/security. He is an Assistant Professor of Computer Science at ETH Zürich (since Aug 2022), where he leads the SPY Lab; since 2026 he also holds a Research Scientist position at Google DeepMind Zürich, and since 2024 he is co-founder and academic advisor of Invariant Labs, an ETH-spinoff AI-security startup acquired by Snyk in June 2025. He is additionally an **advisor to Together AI** (since ~Oct 2023) and runs a one-man consulting firm, **Tramèr AI Security Consulting** (registered Nov 2024). He grew up in the Geneva countryside (Puplinge, with 1995–97 in Oxford), the son of HUG anesthesiology professor Martin Tramèr and Claudia Tramèr-Johns; his brother Lucas Tramèr is an Olympic rowing champion turned cardiologist. He did his BSc and MSc at EPFL (2009–2015, incl. a CMU exchange year), a PhD at Stanford under Dan Boneh (2016–2021), and a year as visiting faculty at Google Brain Zürich (2021–22). He married the Dutch marketeer **Mariel de Haan** in August 2016 (she is COO/co-founder of TheLearning LAB); they live in **Gutenswil/Volketswil** (Zürich's Greifensee area) and have a daughter, **Emma** (born before Feb 2025). His work — including "Stealing Machine Learning Models via Prediction APIs" (2016), the ChatGPT training-data extraction attack (2023), and "Stealing Part of a Production Language Model" (ICML Best Paper 2024) — has been covered by The Economist, Nature, Science, Wired, and the Atlantic.

120 pages!

“I did not expect the AI to find any of this”



LLMs are better investigators than human pros



This talk.

1. ***Linking*** anonymous online profiles

2. Creating ***personal dossiers*** from the open web

Is this “new”?

> *Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life*

The Right to Privacy. Warren and Brandeis. 1890

Is this “new”?

Until recently, the privacy harms that attended the public disclosure of court records were generally regarded as insignificant because court files were difficult to search and access. But this “practical obscurity” is rapidly disappearing

Privacy and Court Records: Online Access and the Loss of Practical Obscurity. Ardia. 2017

The *human* is taken out of the loop.



Minions are getting pretty smart...



What happens at scale?



Conclusion

AI transforms the economics of surveillance

- Automatically scale up grunt work & simple reasoning
- What goes wrong when attackers have a Mythos army?
- What could thousands of minions do for defense?

spylab.ai

