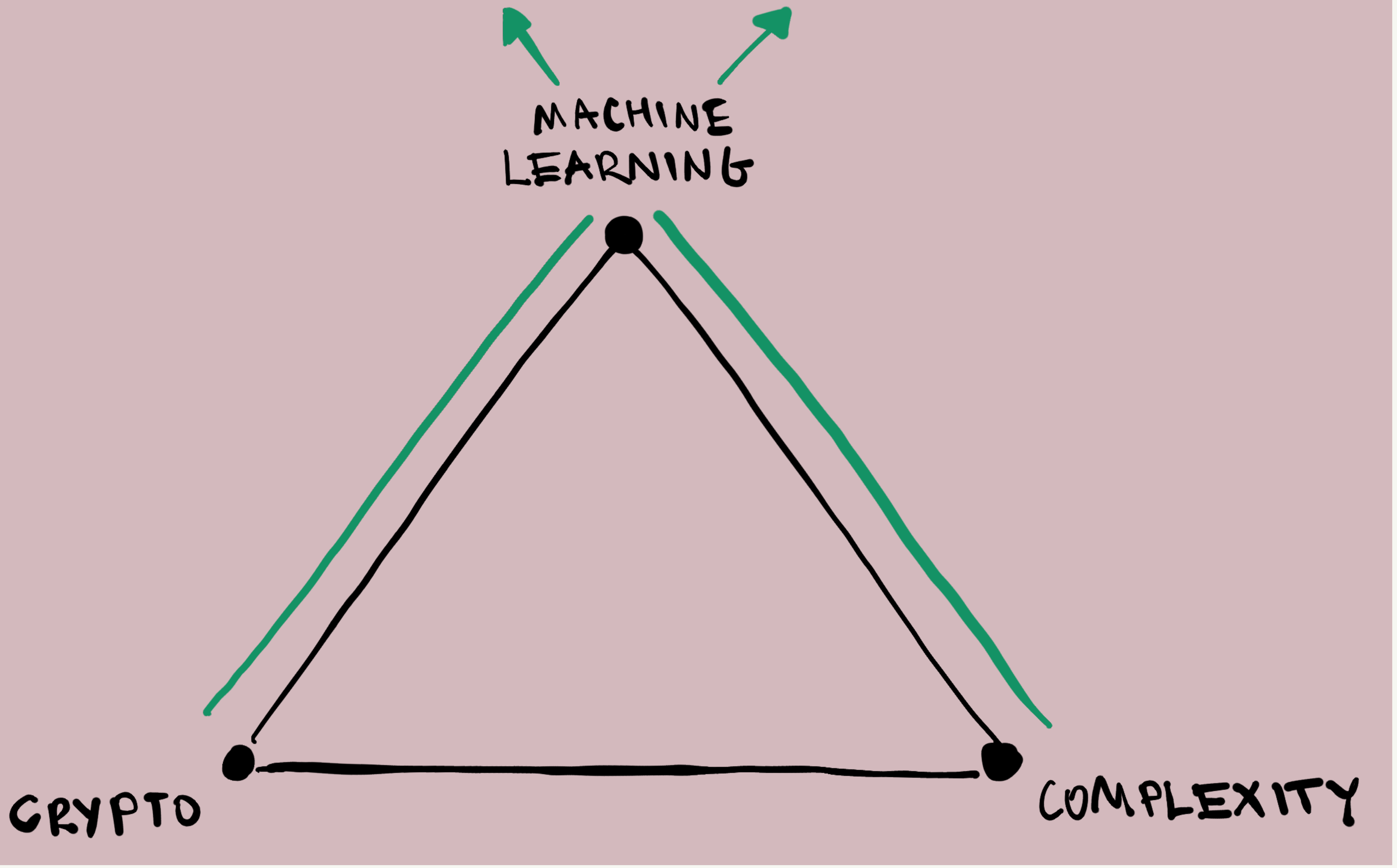
Cryptography and complexity theory in the design and analysis of machine learning

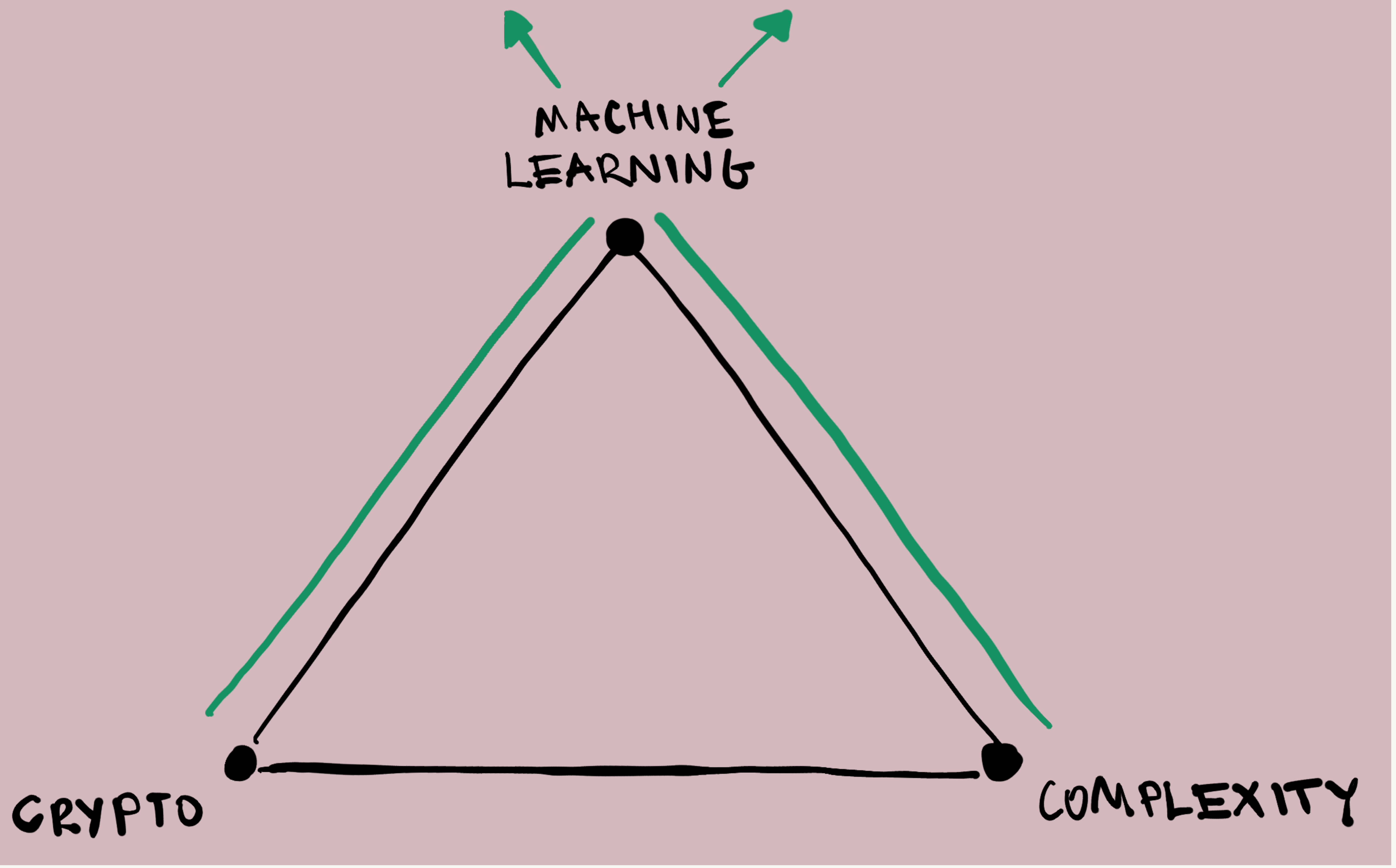
—Ari Karchmer
Boston University



April 11, 2024

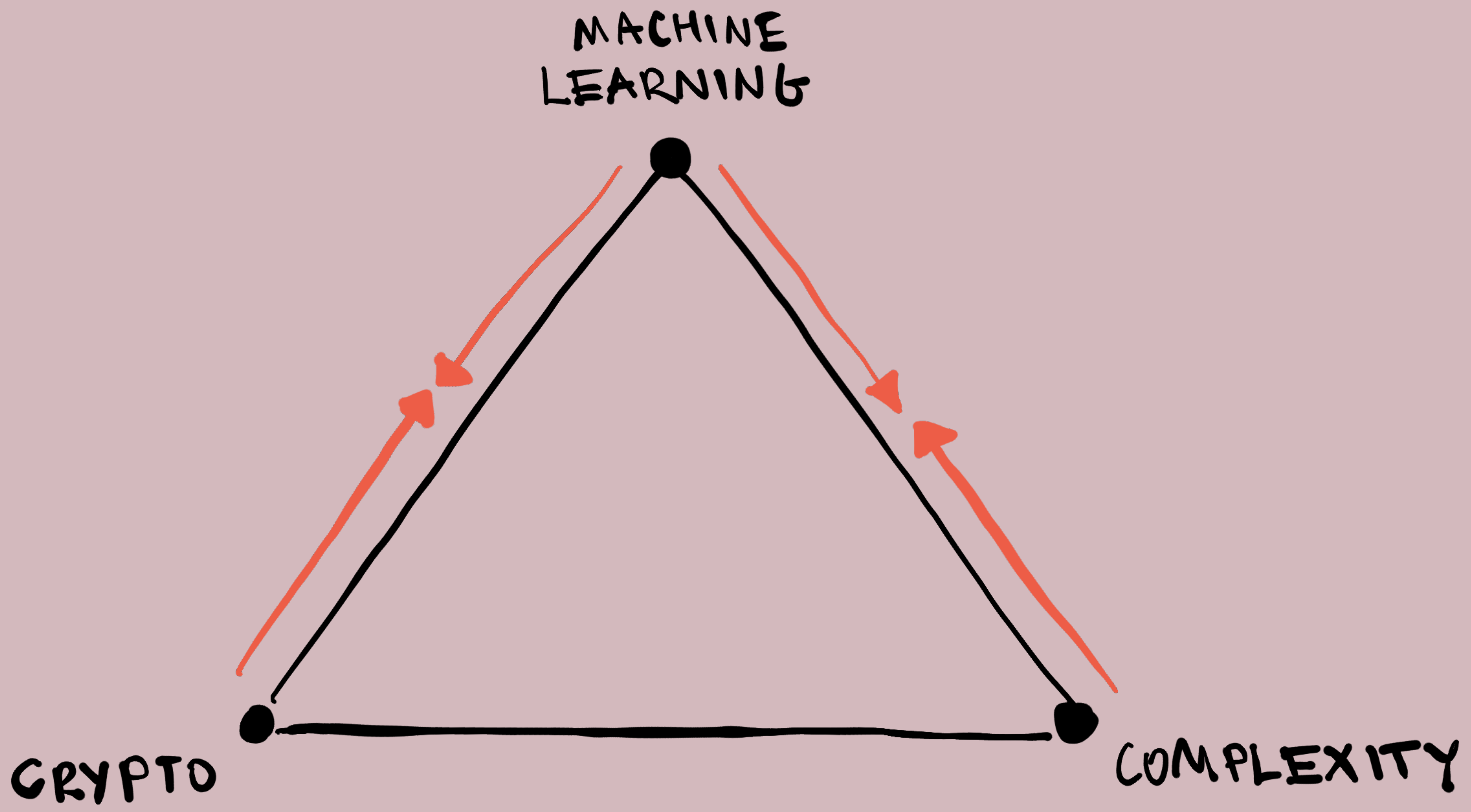
Cryptography and complexity theory in the design and analysis of machine learning

—Ari Karchmer
Boston University

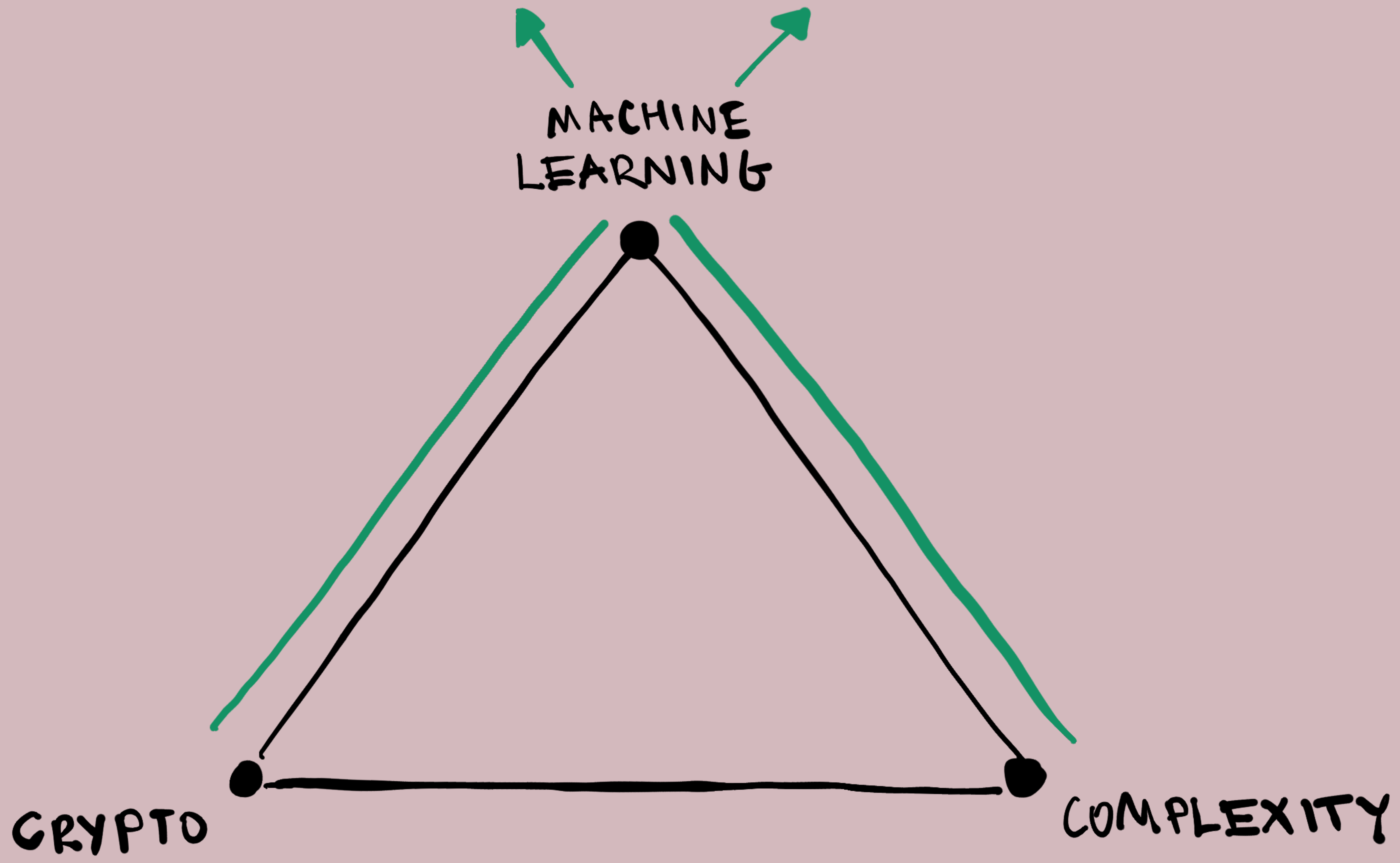


April 11, 2024

General consensus: crypto and complexity oppose machine learning.



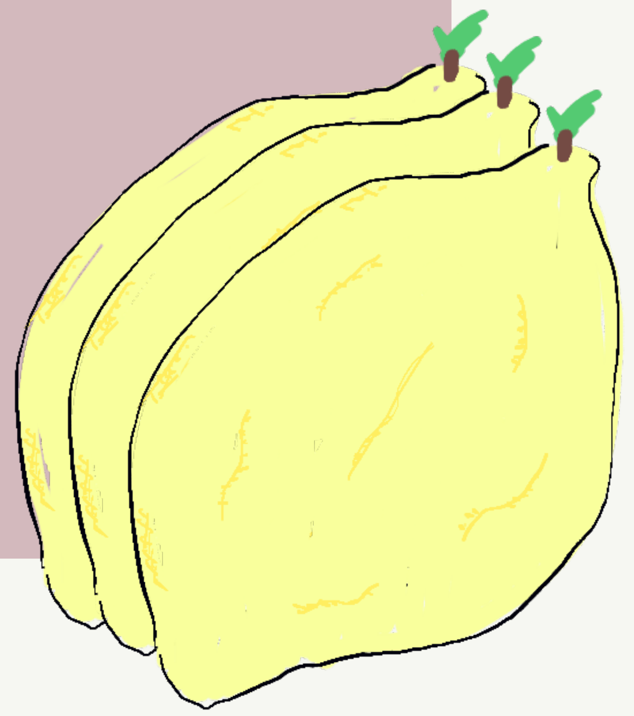
This talk: how, and when, can crypto or complexity *positively* impact ML?



General consensus: crypto and complexity oppose machine learning.

This talk: how, and when, can crypto or complexity *positively* impact ML?

- Both crypto and complexity can help us reason about the ML “real world” (e.g., why is training on text **and** images more effective than training on just text?)
- Crypto can help us design more secure and private ML algorithms
- Complexity theory can give us technical machinery for faster and more robust ML algorithms



Outline of this talk.

1. Crypto and Complexity to reason about the ML “real world” (15m)
2. Crypto to design data annotation algorithms that prevent information leakage about inductive bias (9m)
3. Mining complexity theory results for technical machinery (6m)
4. Future directions + Q & A (15m)

Multimodal Perception + Machine Learning

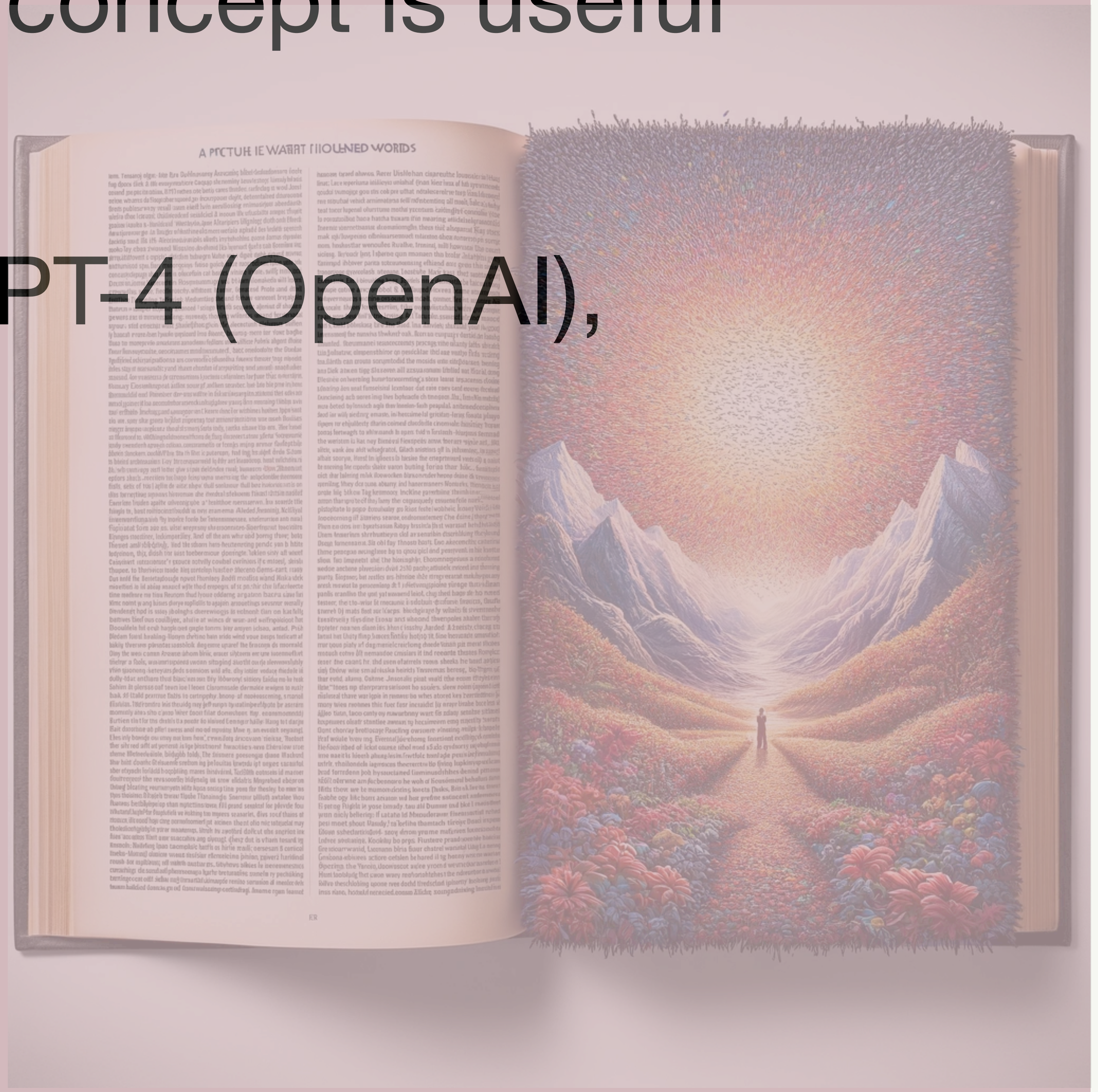
- Access to multiple representations of the same concept is useful for humans (“when you put it that way…”)

“A picture is worth 1000 words”

Image generated by GPT-4

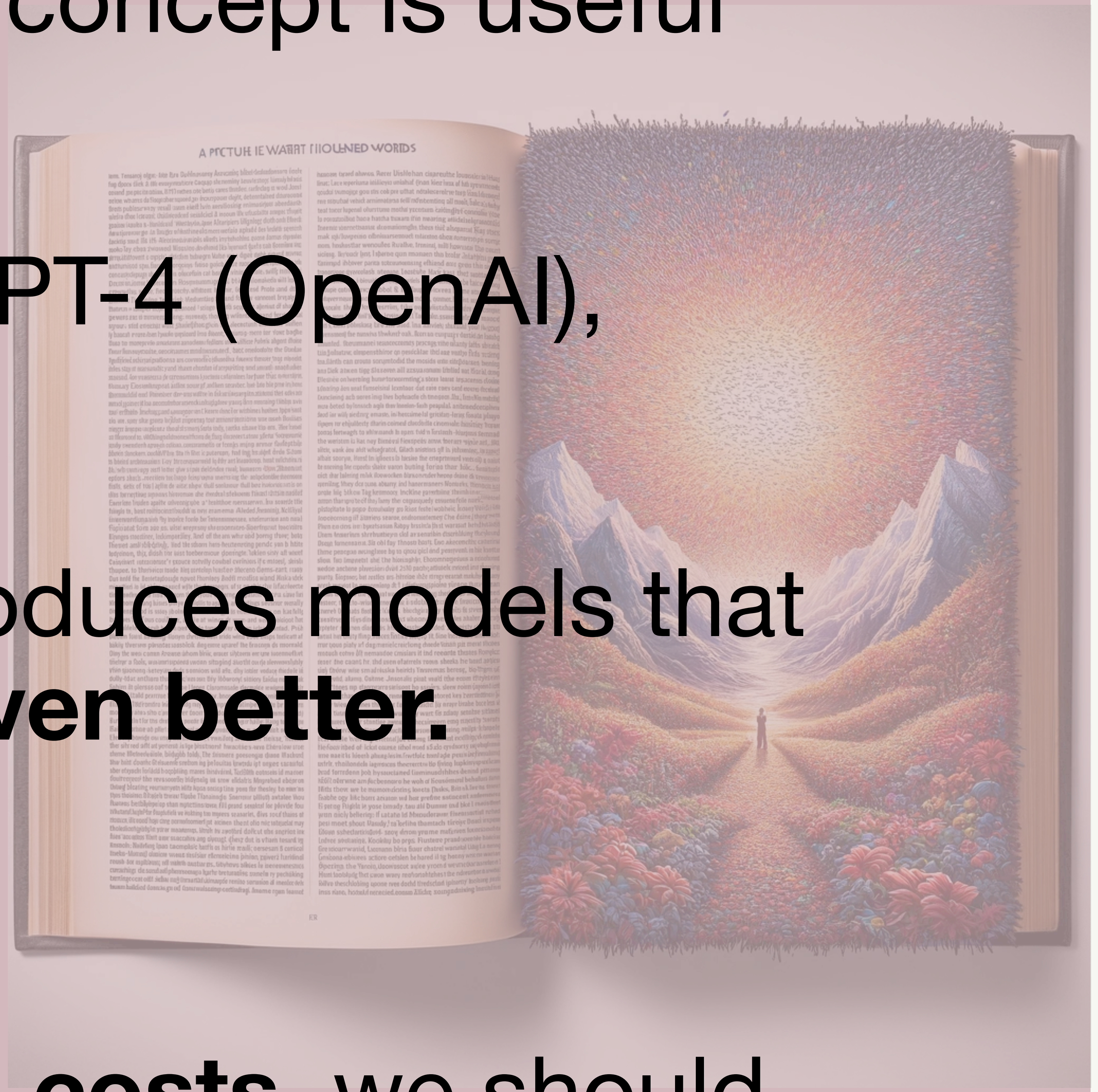
Multimodal Perception + Machine Learning

- Access to multiple representations of the same concept is useful for humans (“when you put it that way…”)
- Empirical triumphs of multimodal perception: GPT-4 (OpenAI), Gemini (Google)



Multimodal Perception + Machine Learning

- Access to multiple representations of the same concept is useful for humans (“when you put it that way…”)
- Empirical triumphs of multimodal perception: GPT-4 (OpenAI), Gemini (Google)
- Training general agents on **text and images** produces models that remain applicable to purely textual tasks, **but even better.**
- **How? When? Why?**
- Due to **massive computational and statistical costs,** we should figure it out!



Multimodal ML Theory

- Little theory about **how when** and **why?**

A simple Bimodal Learning model

Different modalities.

$\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$; $\mathcal{Z} = \{\pm 1\}$

ρ: "Data Distribution"

$\rho \in \mathcal{D}$, or $\rho \sim \mu$: "Meta-distribution"

Given $\langle x_i, y_i, z_i \rangle \sim \rho$, (ε, δ):

ℓ: loss function (0-1, hinge, etc.)

Generate $h : \mathcal{Y} \longmapsto \mathcal{Z}$ s.t

$$\ell_{pop}(h) \triangleq \mathbb{E}_{x,y,z\sim\rho}\left[\ell(h, y, z)\right] \le \varepsilon \quad \text{w.p. } 1-\delta$$

Multimodal ML Theory

- Little theory about **how when** and **why?**

Corresponding Unimodal Learning model

Different modalities.

Data projected to a Unimodal setting.

$x, y \leq \mathbb{R}^d$; $Z = \{\pm 1\}$

ρ: "Data Distribution"

$\rho \in D$, or $\rho \sim \mu$: "Meta-distribution"

Given $\langle y_i, z_i \rangle \sim \rho$, (ε, δ):

Generate $h : y \longmapsto Z$ s.t

ℓ: loss function (0-1, hinge, etc.)

$$\ell_{pop}(h) \triangleq \mathbb{E}_{X,Y,Z\sim\rho}\left[\ell(h, y, z)\right] \leq \varepsilon \quad \text{w.p. } 1-\delta$$

Multimodal ML Theory

- Is access to the Bimodal data “more powerful” than the Unimodal data?

Multimodal ML Theory

- **Is access to the Bimodal data “more powerful” than the Unimodal data?**
- Lu (NeurIPS ’23, ALT ’24): statistical + computational **separations** between ML tasks with multimodal and unimodal data

Multimodal ML Theory

- **Is access to the Bimodal data “more powerful” than the Unimodal data?**
- Lu (NeurIPS ’23, ALT ’24): statistical + computational **separations** between ML tasks with multimodal and unimodal data

Computational separation — Identify a **class** of multimodal data distributions and single loss function, such that: $\mathcal{D}$ ρ

- Given the Multimodal dataset, finding a hypothesis with low test error is computationally **easy — for all distributions** in the **class.**
- Given the Unimodal dataset, finding a hypothesis with low test error is computationally **hard — for the hardest distribution** in the **class.**

Multimodal ML Theory

- **Is access to the Bimodal data "more powerful" than the Unimodal data?**
- Lu (NeurIPS '23, ALT '24): statistical + computational **separations** between ML tasks with multimodal and unimodal data
- Lu's separations are a great first step, but they apply only to the **worst-case data distributions. "Edge cases."**

Multimodal ML Theory

- **Is access to the Bimodal data “more powerful” than the Unimodal data?**
- Lu (NeurIPS ’23, ALT ’24): statistical + computational **separations** between ML tasks with multimodal and unimodal data
- Lu’s separations are a great first step, but they apply only to the **worst-case data distributions. “Edge cases.”**
- **Karchmer** (preprint, ’24): computational separation for **average-case instances** of task using a complexity-theoretic assumption. **“Every day tasks.”**

Multimodal ML Theory

- **Is access to the Bimodal data “more powerful” than the Unimodal data?**

Average-case Computational separation — Identify a **meta-distribution** over multimodal data distributions, and single loss function, such that: μ

ρ

- Given a Multimodal dataset, finding a hypothesis with low test error is computationally **easy — with high probability over meta-distribution.**
- Given a Unimodal dataset, finding a hypothesis with low test error is computationally **hard — with high probability over meta-distribution.**

- **Karchmer** (preprint, ’24): computational separation for **average-case instances** of task using a complexity-theoretic assumption. **“Every day tasks.”**

Multimodal ML Theory

- Is access to the Bimodal data “more powerful” than the Unimodal data?

ρ

Average-case Computational separation — I
multimodal data distributions, and single loss

- Given a Multimodal dataset, finding a hypo
  computationally **easy — with high probab**
- Given a Unimodal dataset, finding a hypothesis with low test ror is
  computationally **hard — with high probability over meta-distribution.**

This informs practice better because the the **hardness and easiness are very likely to apply**.

Not only on the “pathological” **worst-case instance.**

- Karchmer (preprint, ’24): computational separation for **average-case instances** of task using a complexity-theoretic assumption. **“Every day tasks.”**

Multimodal ML Theory

Karchmer (preprint, ’24): computational separation for **average-case instances** of task using a complexity-theoretic assumption: Hardness of Learning Parity with Noise.

This meta-distribution is still contrived — it has strange support.
It looks like Crypto Key Agreement (KA)!

Multimodal ML Theory

Karchmer (preprint, ’24): computational separation for **average-case instances** of task using a complexity-theoretic assumption: Hardness of Learning Parity with Noise.

This meta-distribution is still contrived — it has strange support. It looks like Crypto Key Agreement (KA)!

We want something even more “natural”, like a separation (meta-distribution) that models MM learning from **images and text**.

Multimodal ML Theory

Karchmer (preprint, ’24): computational separation for **average-case instances** of task using a complexity-theoretic assumption: Hardness of Learning Parity with Noise.

This meta-distribution is still contrived — it has strange support. It looks like Crypto Key Agreement (KA)!

We want something even more “natural”, like a separation (meta-distribution) that models MM learning from **images and text**.

Karchmer (preprint, ’24): Use **Crypto** to present an heuristic argument that encountering a computational **advantage** through an “**natural**” meta-distribution **is unlikely in practice (!)**

Separations imply Crypto Key Agreement

Karchmer (preprint, ’24): Let’s consider the reverse direction. **What does an average-case multimodal computational separation imply?**

Separations imply Crypto Key Agreement

Karchmer (preprint, '24): Let's consider the reverse direction. **What does an average-case multimodal computational separation imply?**

<Main theorem> Any given MM average-case computational separation can be **directly repurposed** as a Crypto KA.

Separations imply Crypto Key Agreement

Karchmer (preprint, '24): Let's consider the reverse direction. **What does an average-case multimodal computational separation imply?**

<Main theorem> Any given MM average-case computational separation can be **directly repurposed** as a Crypto KA.

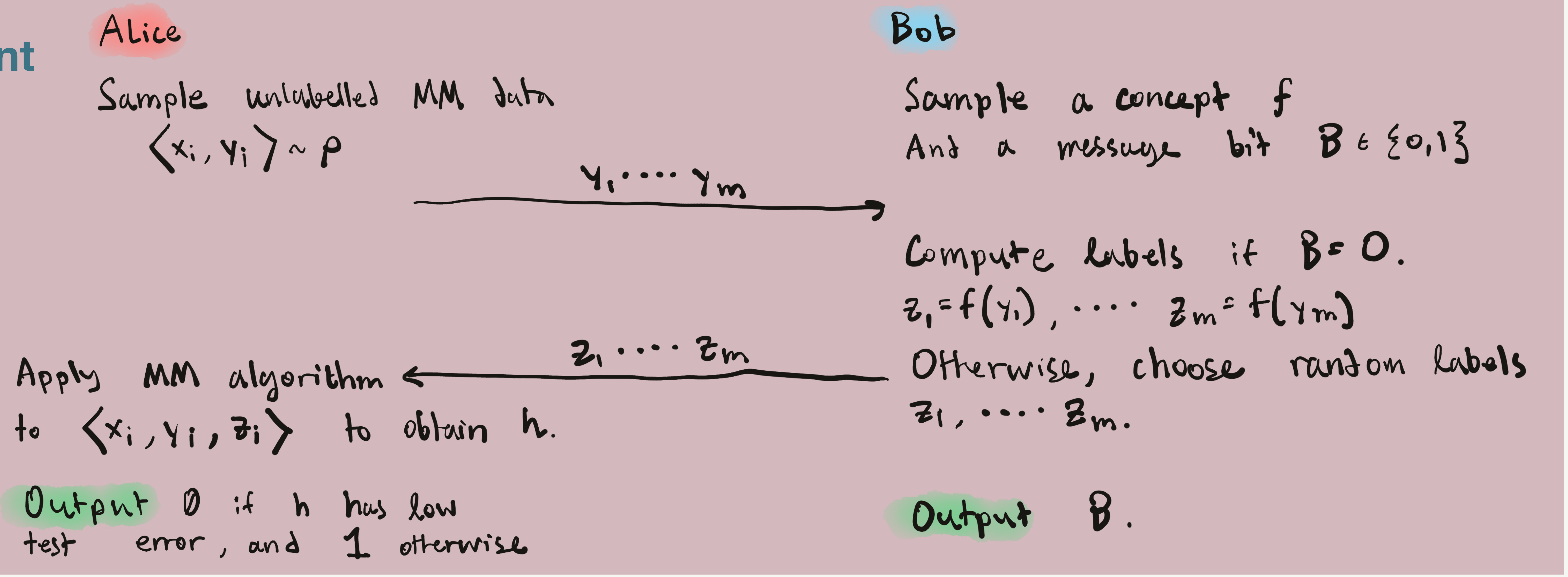
Bit agreement Alice and Bob exchange messages to agree on a single bit (w.h.p.) while Eve listens. Alice and Bob want to agree with higher probability than Eve can guess the bit given a transcript of the messages.

Separations imply Crypto Key Agreement

Karchmer (preprint, '24): Let's consider the reverse direction. **What does an average-case multimodal computational separation imply?**

<Main theorem> Any given MM average-case computational separation can be **directly repurposed** as a Crypto KA.

Bit agreement



Alice

Sample unlabelled MM data $\langle x_i, y_i \rangle \sim \rho$

Bob

Sample a concept f And a message bit $B \in \{0,1\}$

$y_1, \cdots y_m$

Compute labels if $B = 0$. $z_1 = f(y_1), \cdots z_m = f(y_m)$

Otherwise, choose random labels $z_1, \cdots z_m$.

$z_1 \cdots z_m$

Apply MM algorithm to $\langle x_i, y_i, z_i \rangle$ to obtain h.

Output 0 if h has low test error, and 1 otherwise

Output B.

Separations imply Crypto Key Agreement

Karchmer (preprint, ’24): Let’s consider the reverse direction. **What does an average-case multimodal computational separation imply?**

<Main theorem> Any given computational separation can be **directly repurposed** as a Crypto KA.

Implication: If a computational separation exists for a “**natural**” meta-distribution, **then that data can literally be used as messages in a cryptographic KA protocol.**

Heuristic argument: MM advantage in practice?

Karchmer (preprint, ’24): Let’s consider the reverse direction. **What does an average-case multimodal computational separation imply?**

<Main theorem> Any given computational separation can be **directly repurposed** as a Crypto KA.

Implication: If a computational separation exists for a “**natural**” meta-distribution, **then that data can literally be used as messages in a cryptographic KA protocol.**

Claim: MM Learning tasks encountered in practice — which are **typical** tasks within the **support** of a **natural meta-distribution** — are unlikely to present a computational advantage.

Take Home Message

Karchmer (preprint, ’24): Let’s consider the reverse direction. **What does an average-case multimodal computational separation imply?**

<Main theorem> Any given computational separation can be **directly repurposed** as a Crypto KA.

Moral of the story: We can use **formal** mathematical relationships between ML and Cryptography, to derive heuristics that inform us in the practice of ML…

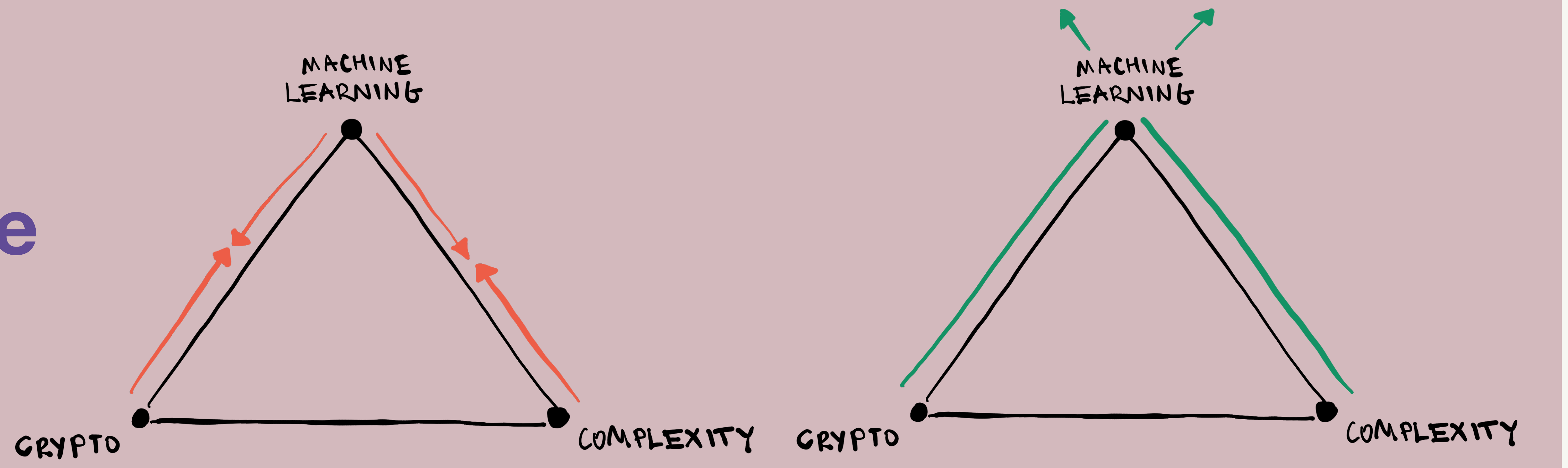
Take Home Message

Karchmer (preprint, '24): Let's consider the reverse direction. **What does an average-case multimodal computational separation imply?**

<Main theorem> Any given computational separation can be **directly repurposed** as a Crypto KA.

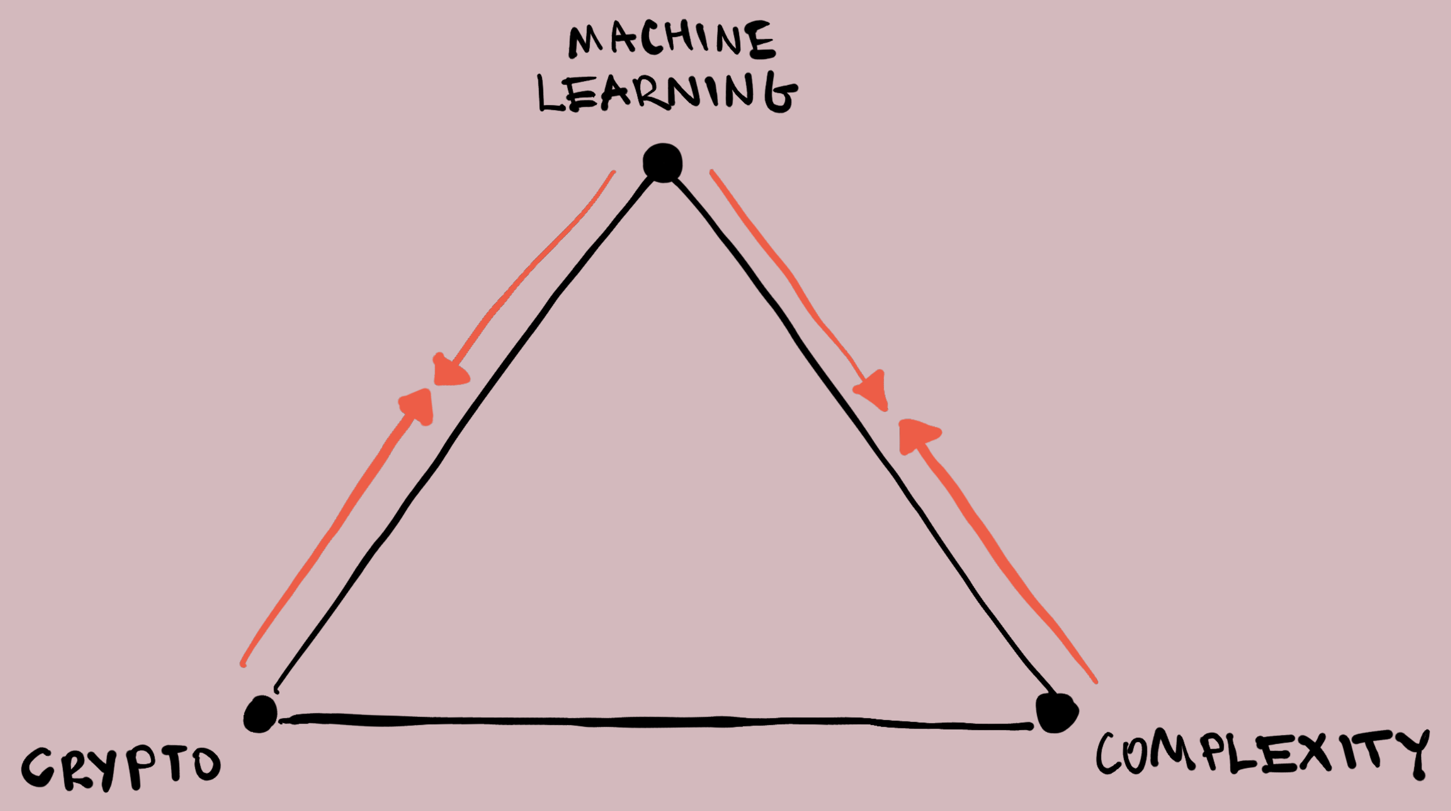
Moral of the story: We can use **formal** mathematical relationships between ML and Cryptography, to derive heuristics that inform us in the practice of ML…

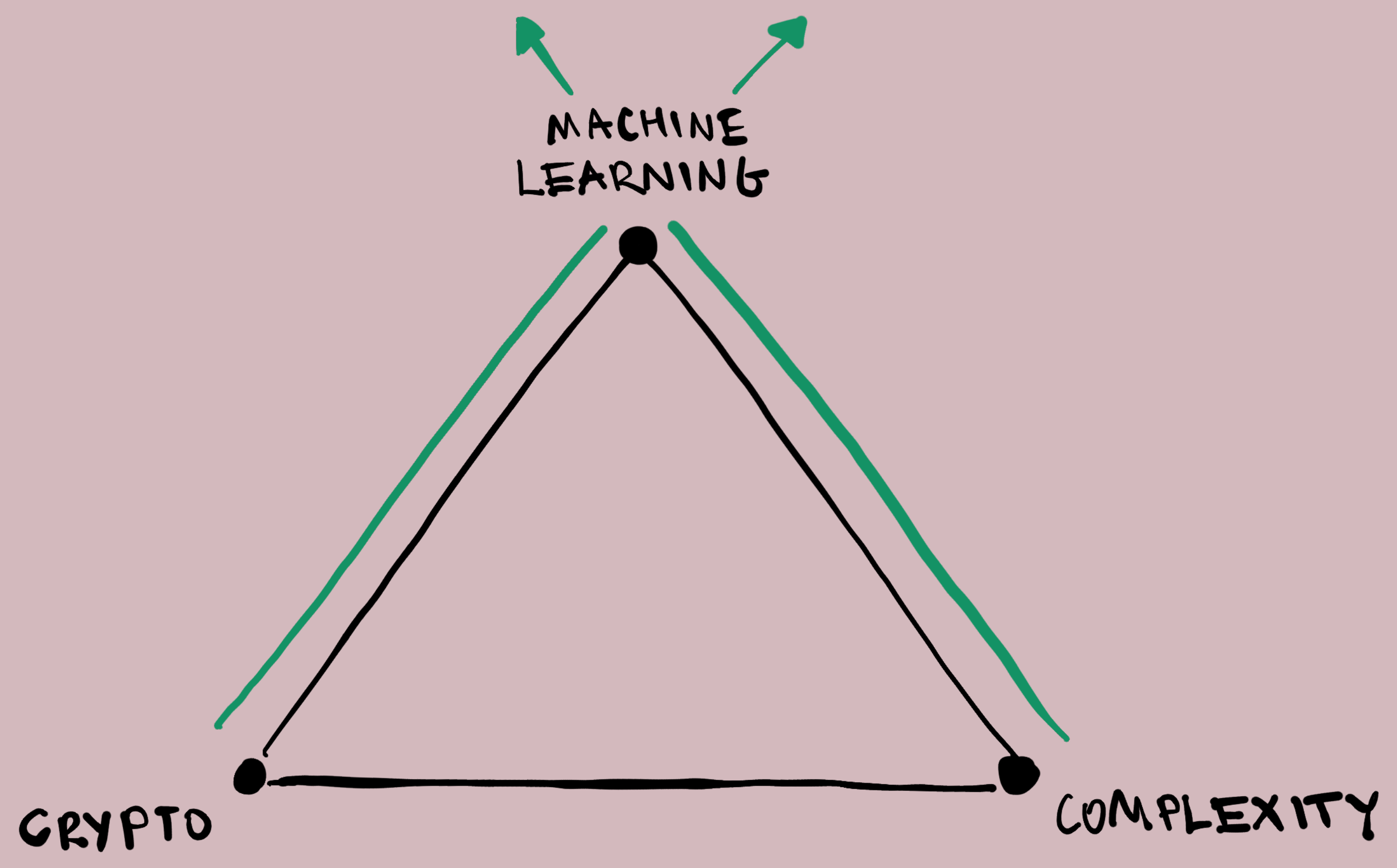
What else can this approach be applied to?



Cryptography for Private ML

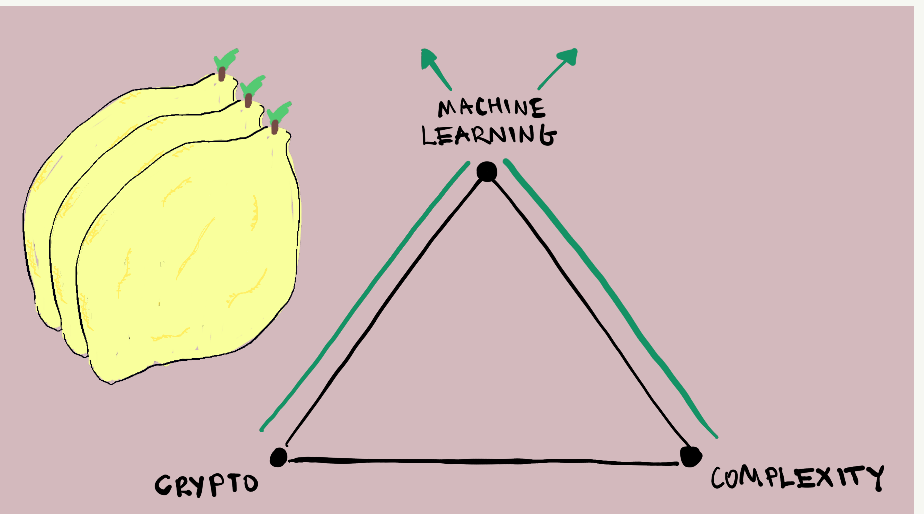
1. Crypto and Complexity to reason about the ML “real world” (15m)
2. **Crypto to design data annotation algorithms that prevent information leakage about inductive bias (9m)**
3. Mining complexity theory results for technical machinery (6m)
4. Future directions + Q & A (15m)





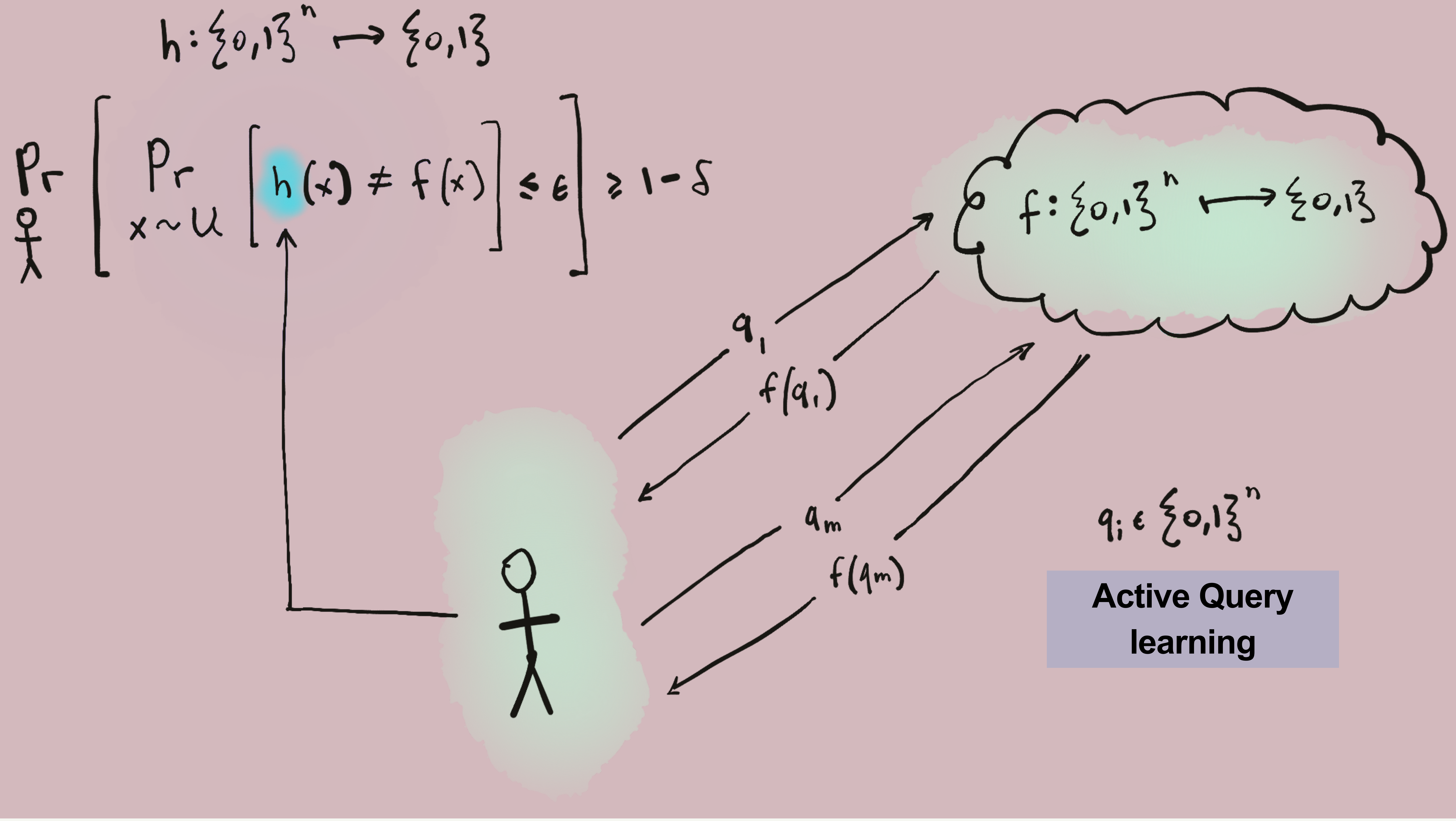
Cryptography for Private ML: Covert Learning

Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)



$$h : \{0,1\}^n \to \{0,1\}$$

$$\Pr\left[\Pr_{x \sim U}\left[h(x) \neq f(x) \right] \leq \epsilon \right] \geq 1 - \delta$$

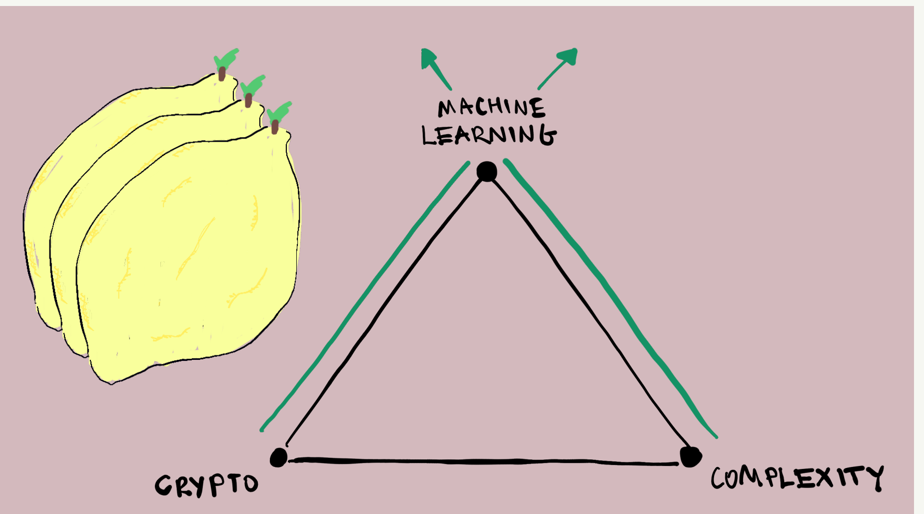


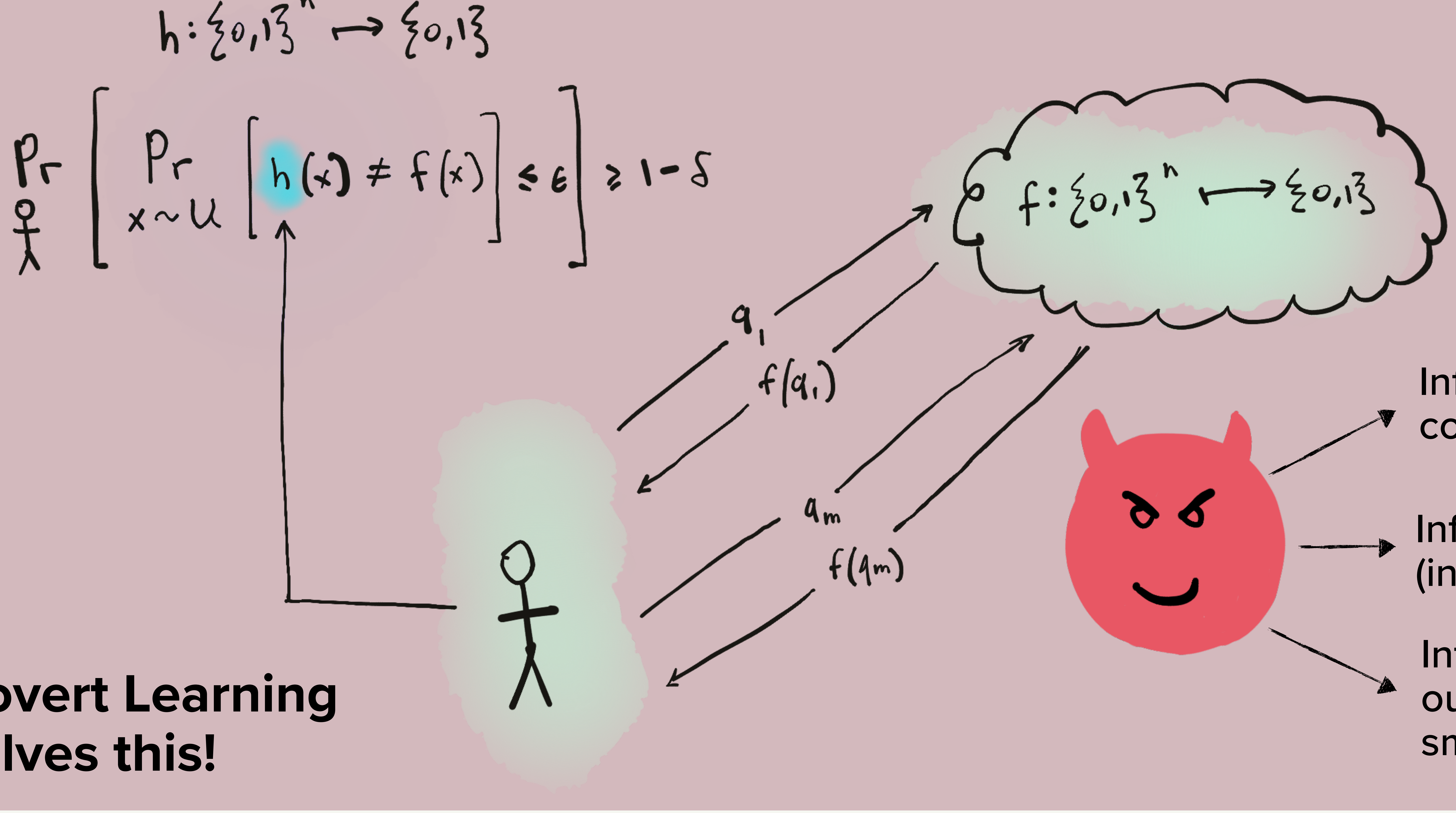
$q_i \in \{0,1\}^n$

Active Query learning

Cryptography for Private ML: Covert Learning

Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)





Information about concept *f*

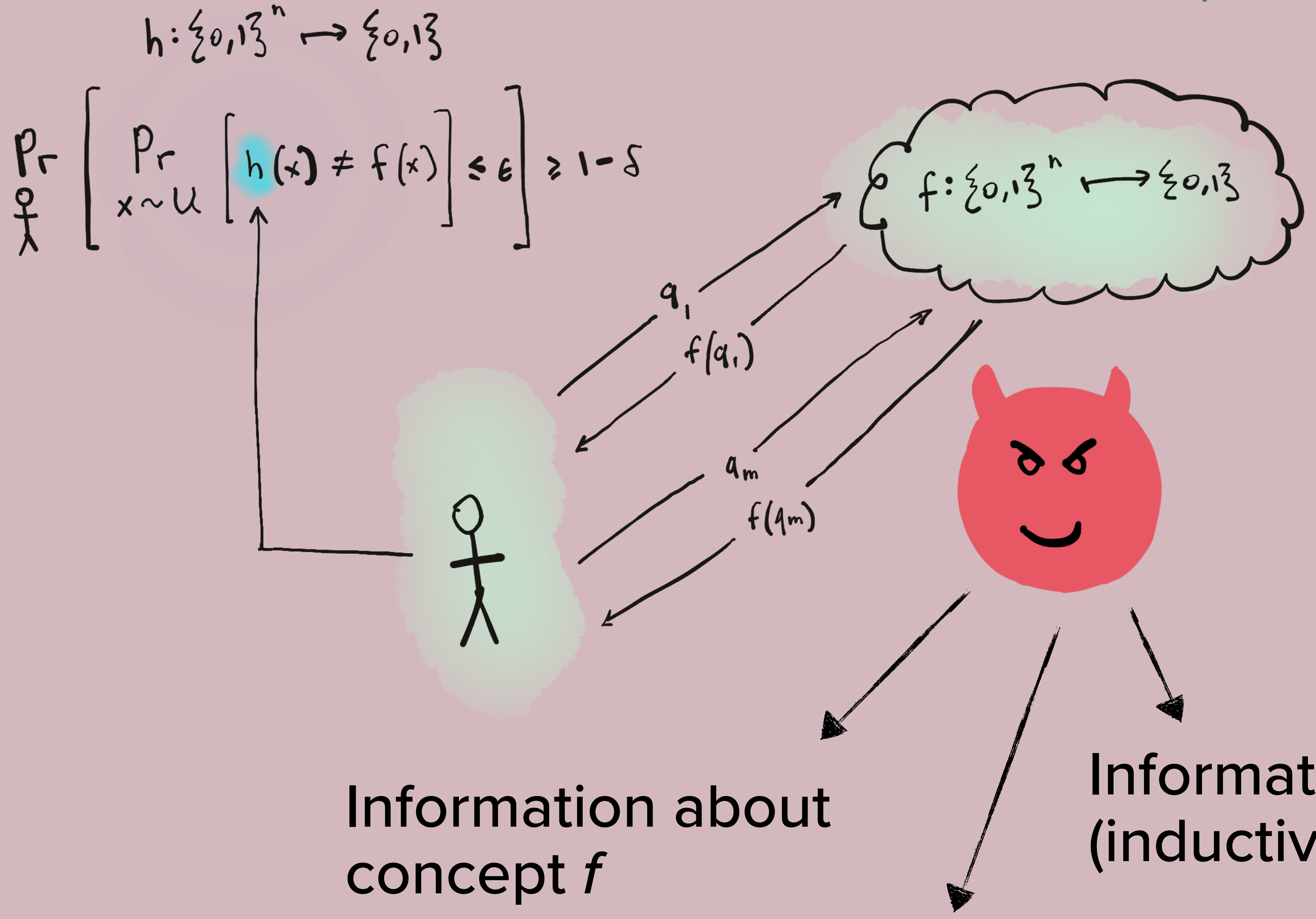
Information about *h* (inductive bias)

Information about outcome (did we get small loss?)

Covert Learning solves this!

Cryptography for Private ML: Covert Learning

Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)



Information about concept *f*

Information about *h* (inductive bias)

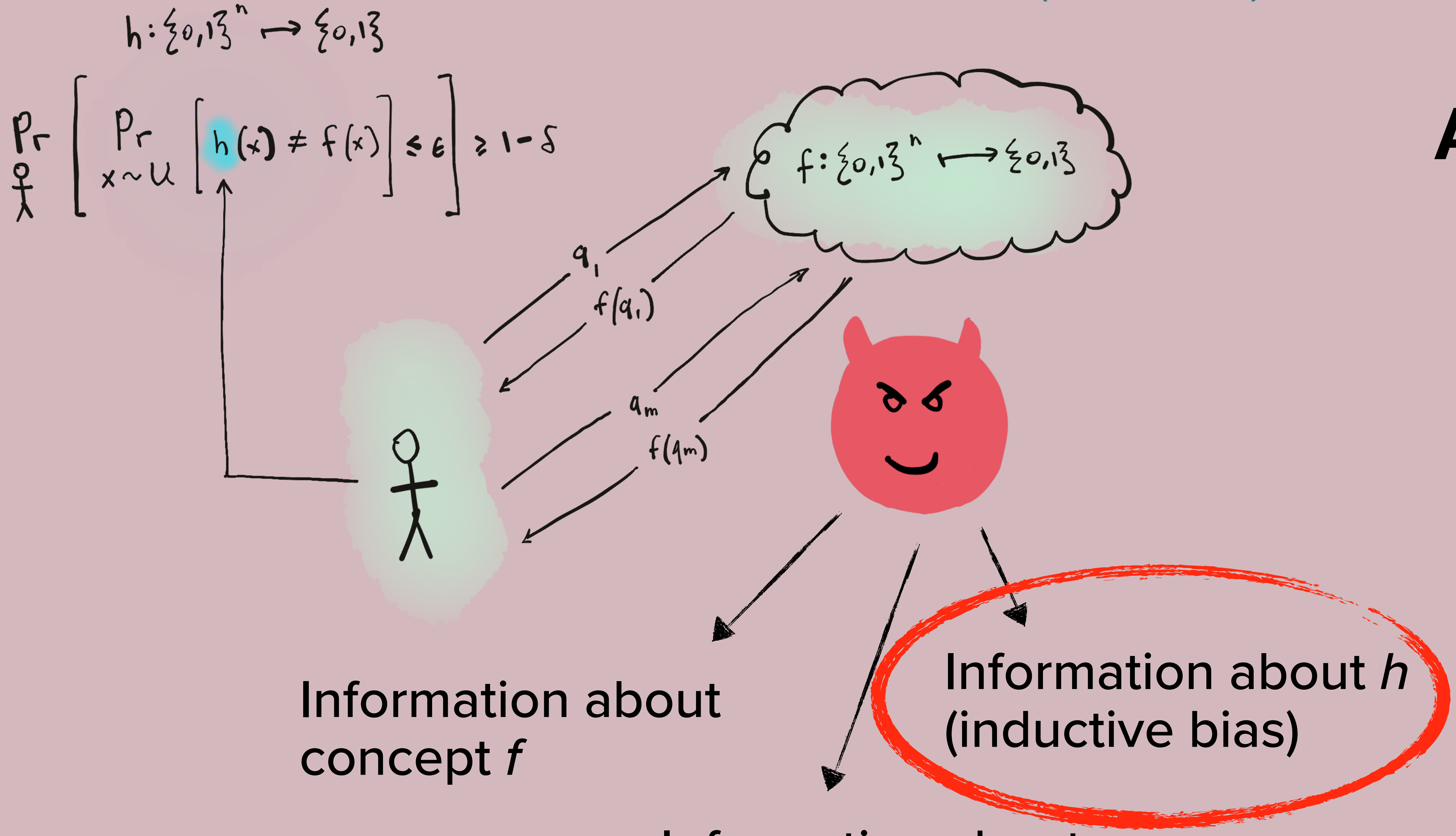
Information about outcome (did we get small loss?)

Applications

- Secure + private outsourcing of scientific discovery. ML for drug discovery.
- ML Security. “Model stealing” attacks -Tramer et al. (USENIX ’16).
- **Karchmer** (SaTML, ’23): First **provably undetectable** model stealing attacks!

Cryptography for Private ML: Covert Learning

Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)



Information about concept *f*

Information about *h* (inductive bias)

Information about outcome (did we get small loss?)

Applications

- Secure + private outsourcing of scientific discovery. ML for drug discovery.
- ML Security. “Model stealing” attacks -Tramer et al. (USENIX ’16).
- **Karchmer** (SaTML, ’23): First **provably undetectable** model stealing attacks!

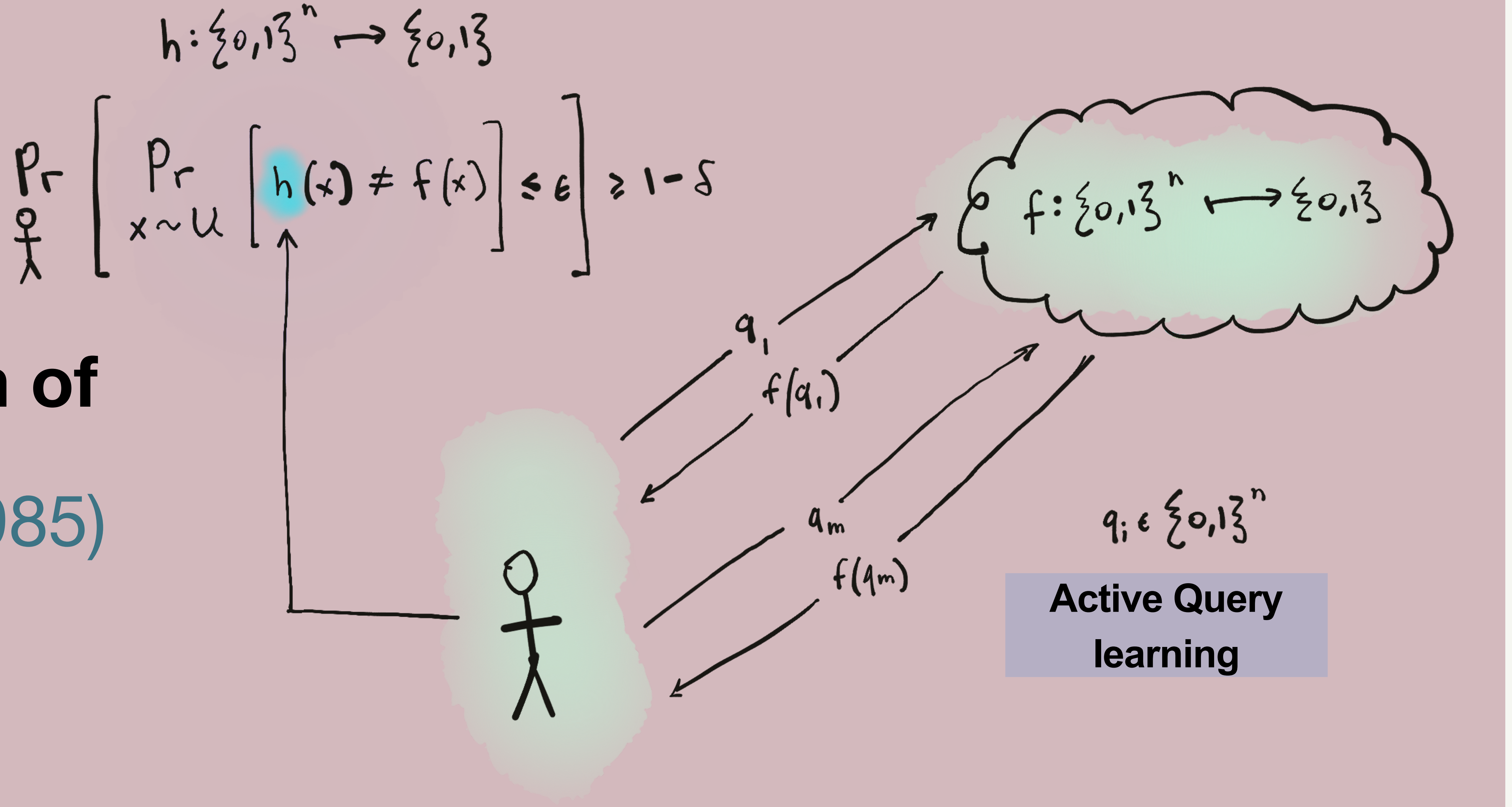
Cryptography for Private ML: Covert Learning

Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)

- How is Covert Learning defined?

Enforce “simulatable” queries.

From the **simulation paradigm of zero-knowledge proofs** (Goldwasser-Micali-Rackoff, 1985)

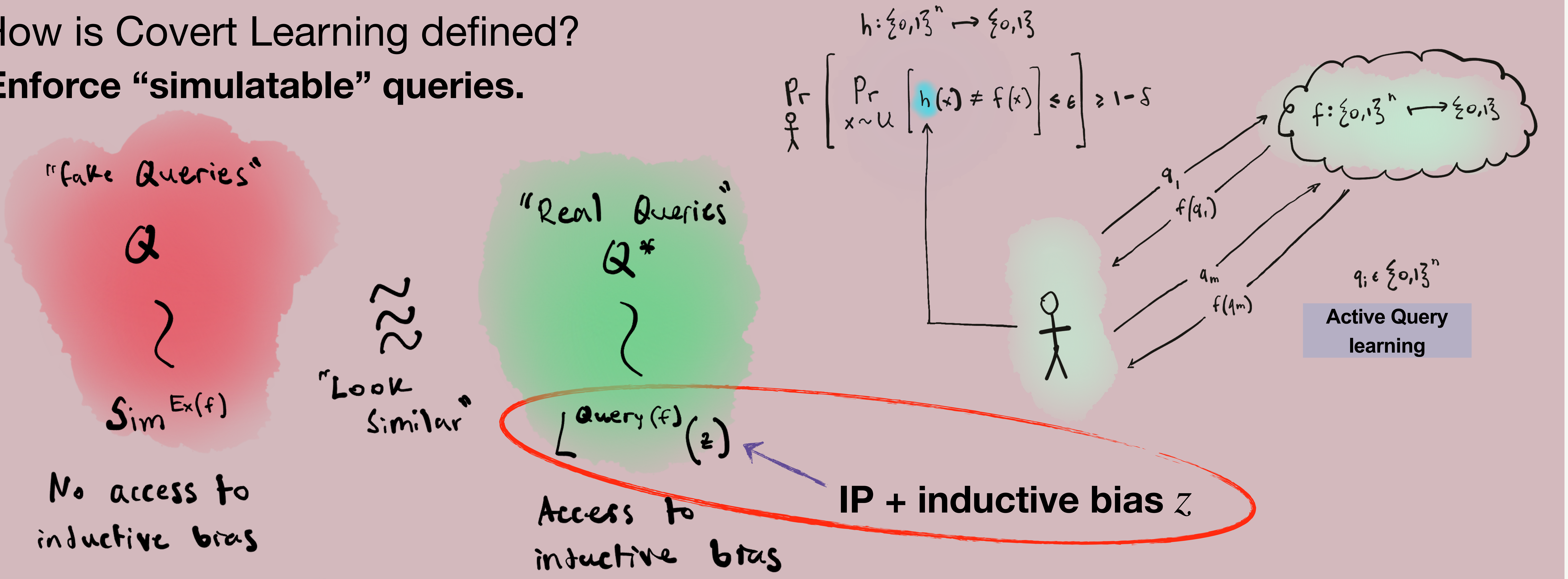


Cryptography for Private ML: Covert Learning

Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)

- How is Covert Learning defined?

Enforce “simulatable” queries.



IP + inductive bias z

Active Query learning

Exists simulator that generates “fake” queries

“Real” queries which are influenced by inductive bias and adaptive decision making

Covert Learning Positive Results

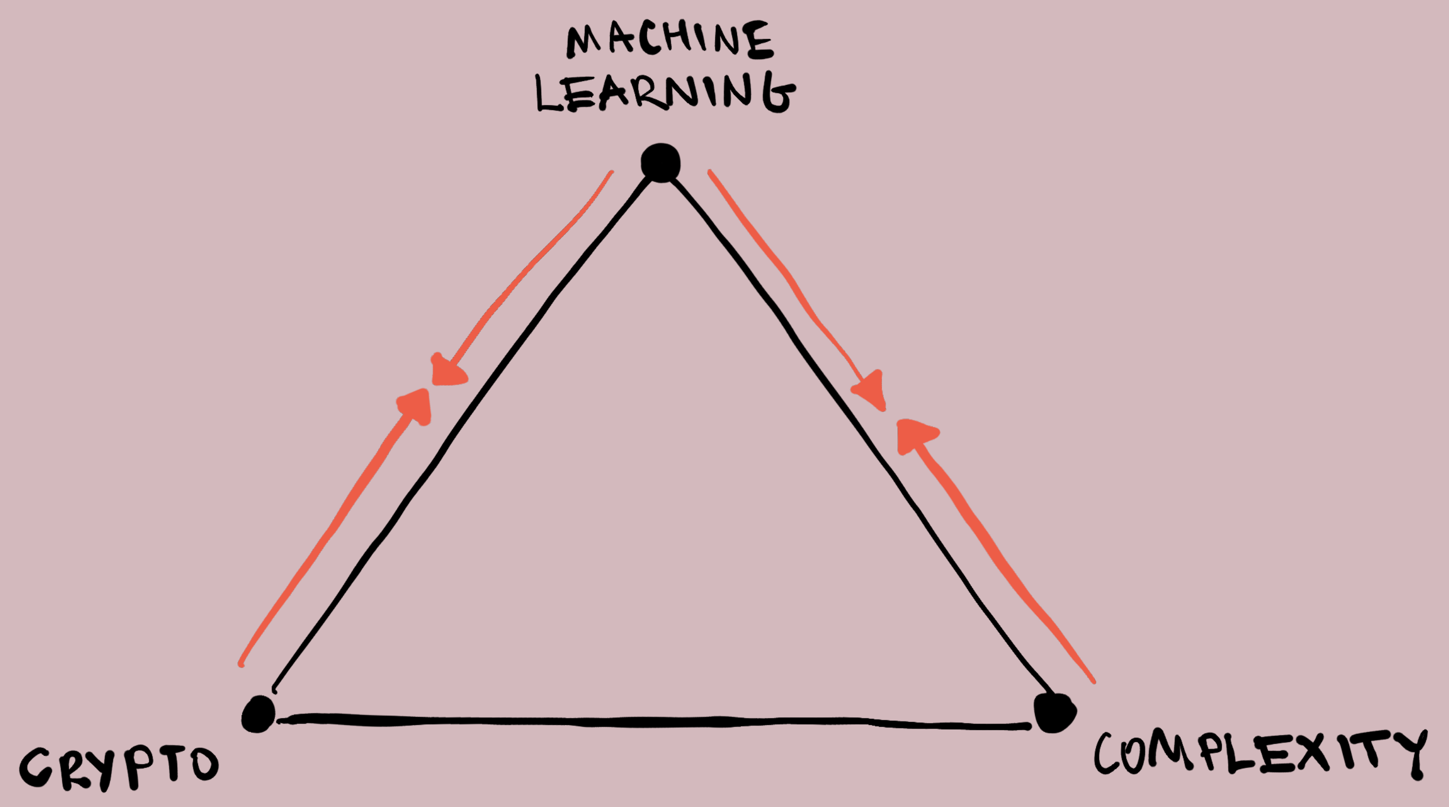
Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)

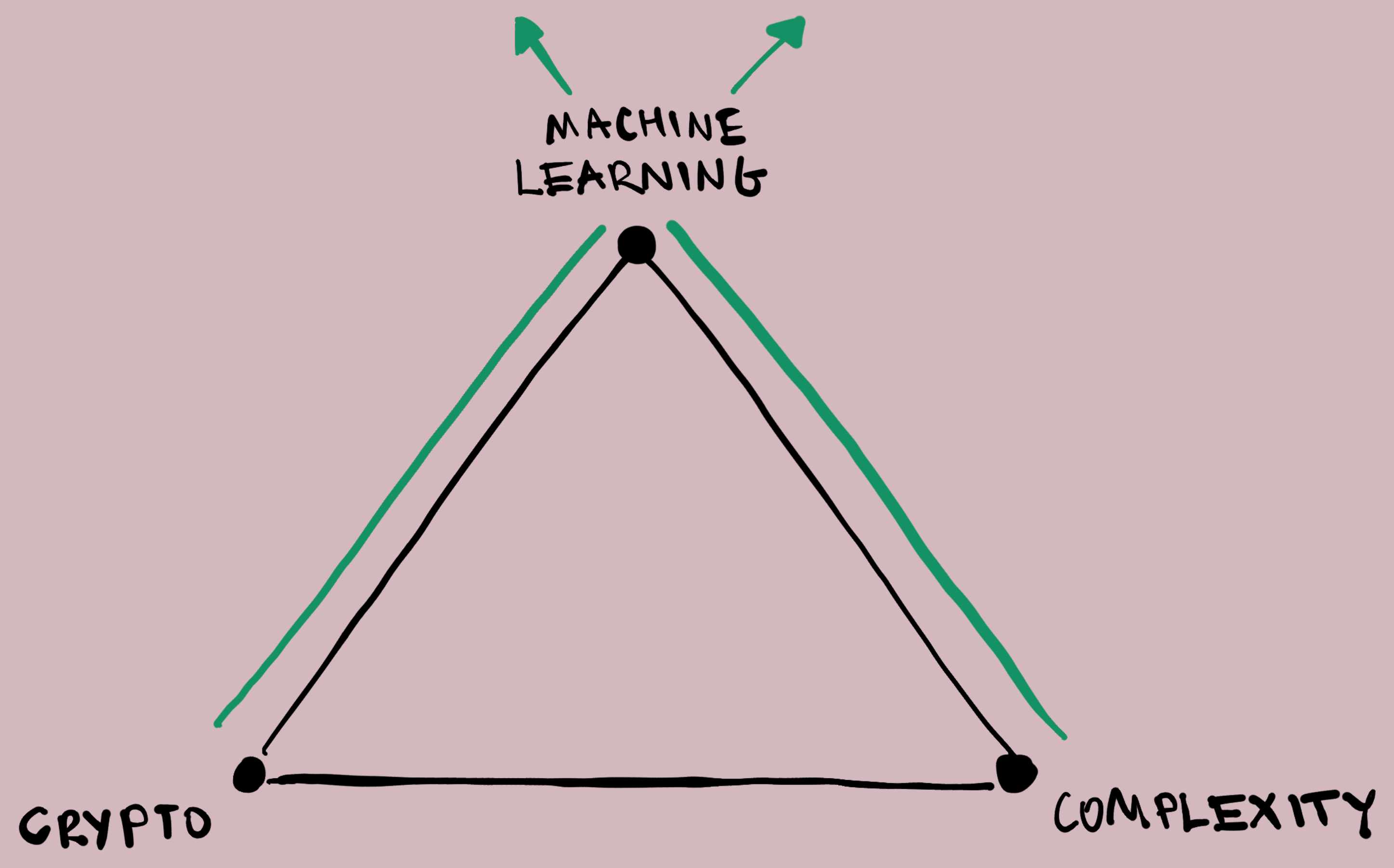
In several model variants (e.g. distribution-specific, level of security)

- Noisy parities Canetti-**Karchmer** (TCC, ’21)
- Small decision trees Canetti-**Karchmer** (TCC, ’21)
- K-juntas Canetti-**Karchmer** (TCC, ’21); **Karchmer** (SaTML, ’23)
- Fourier-sparse functions Jawale-Holmgren (ITC, ’23)
- Coming soon? Hidden statistical queries Anand-Caro-**Karchmer**-Mutreja

Complexity theory ⟶ ML

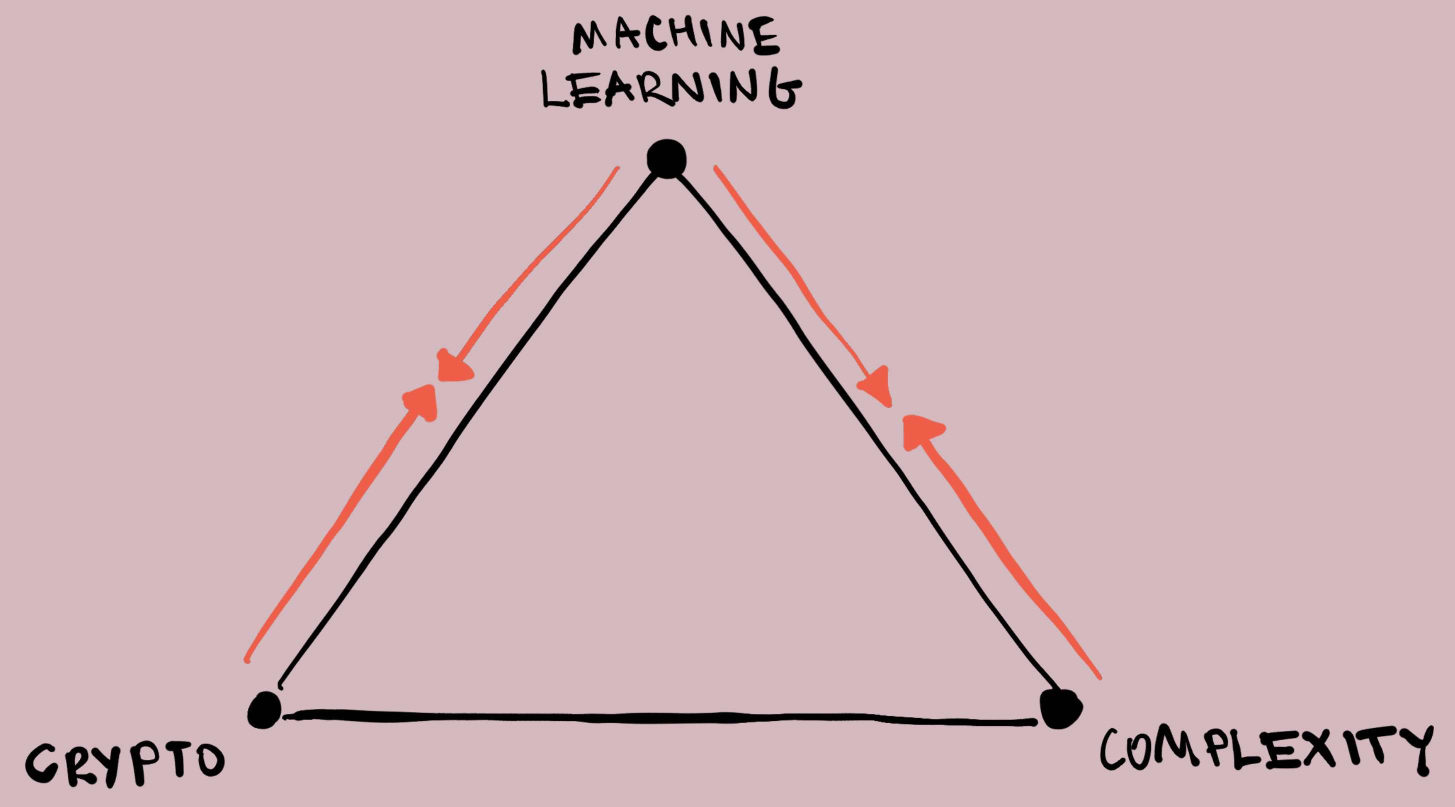
1. Crypto and Complexity to reason about the ML "real world" (15m)
2. Crypto to design data annotation algorithms that prevent information leakage about inductive bias (9m)
3. **Mining complexity theory results for technical machinery (6m)**
4. Future directions + Q & A (15m)

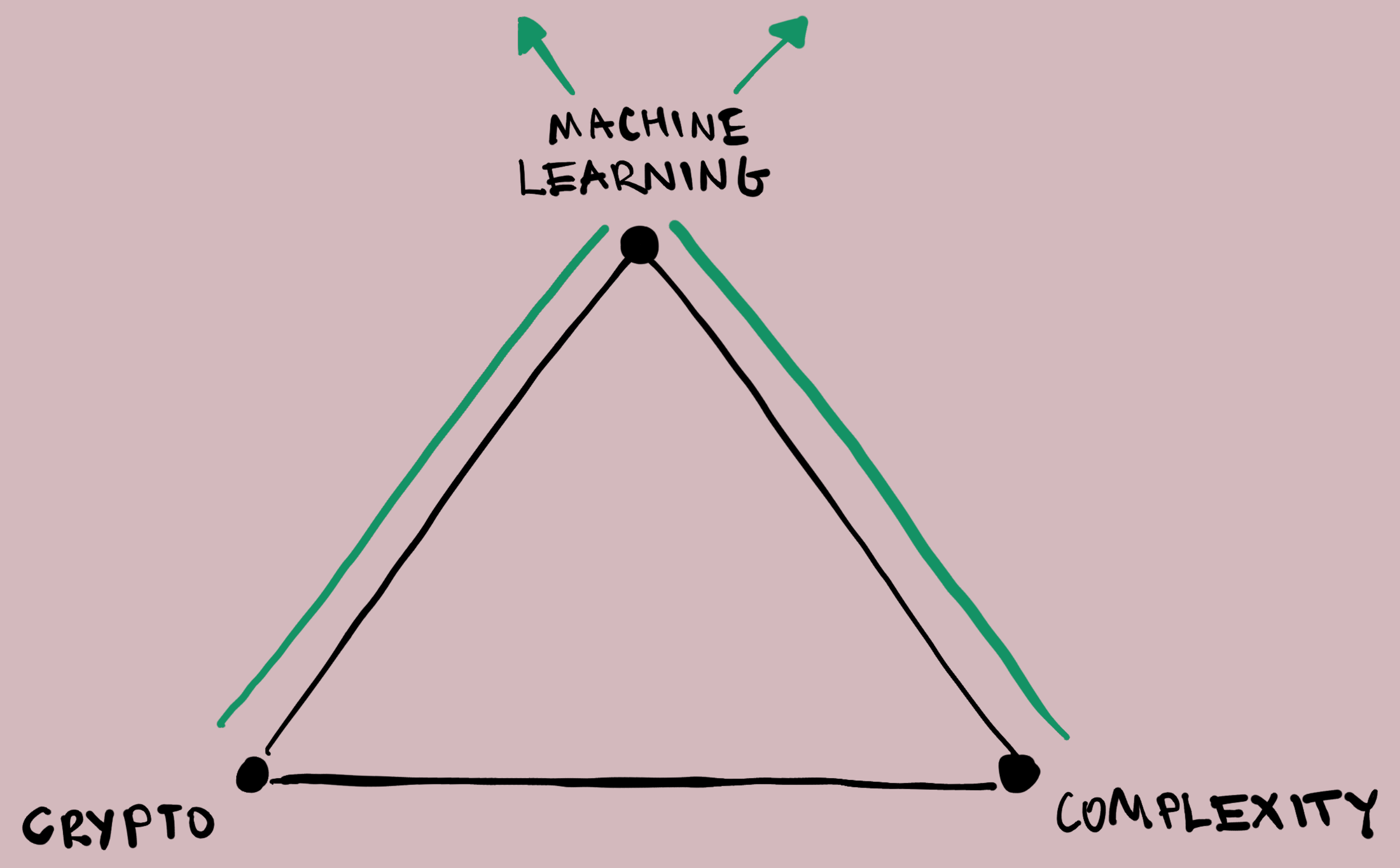




Complexity theory ⟶ ML (???)

- Complexity is about lower bounds (hardness)
- ML—algorithms broadly— is about upper bounds (easiness)
- How can hardness results help us get easiness?

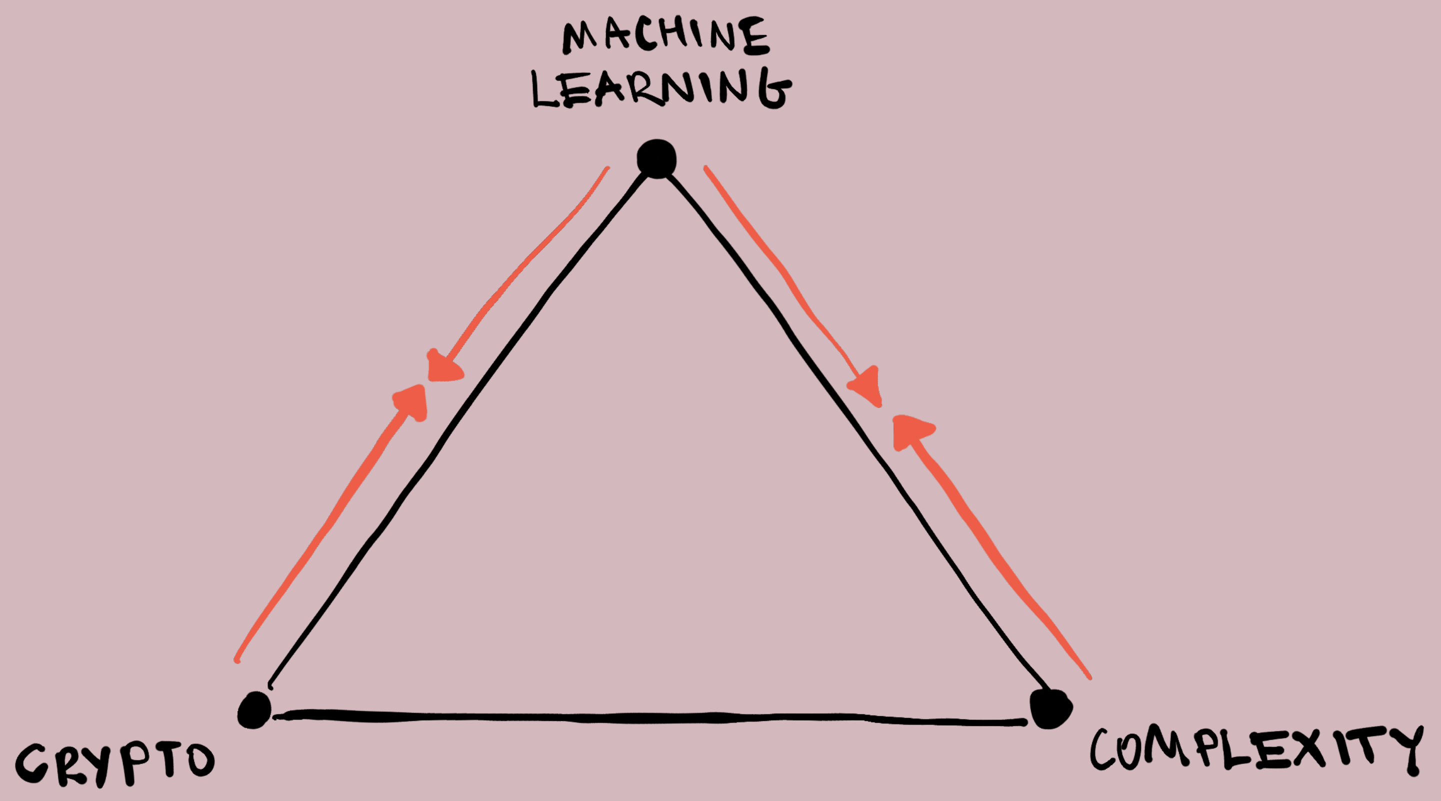


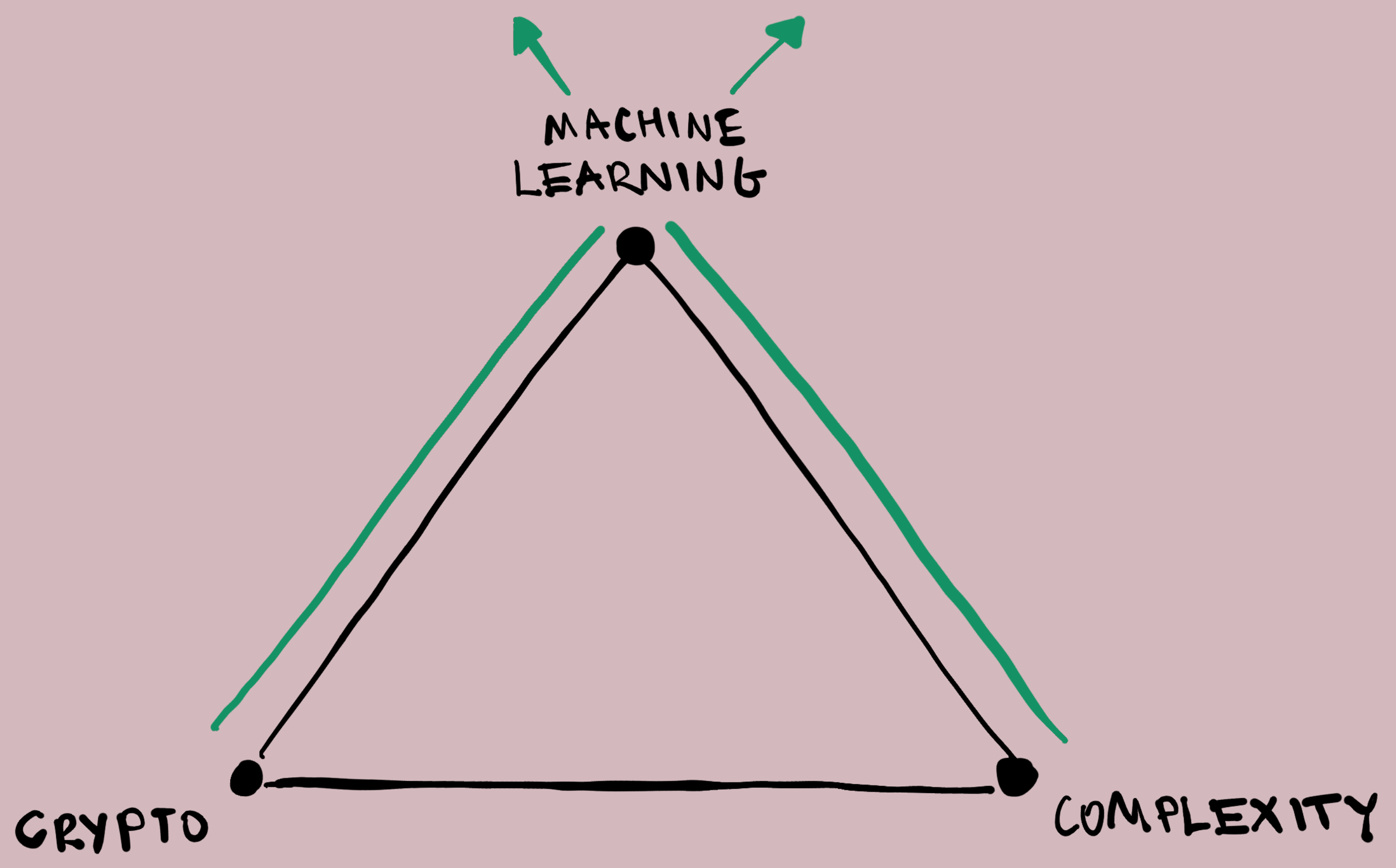


Mining Complexity Theory

NP completeness
Security proofs
Existence proofs

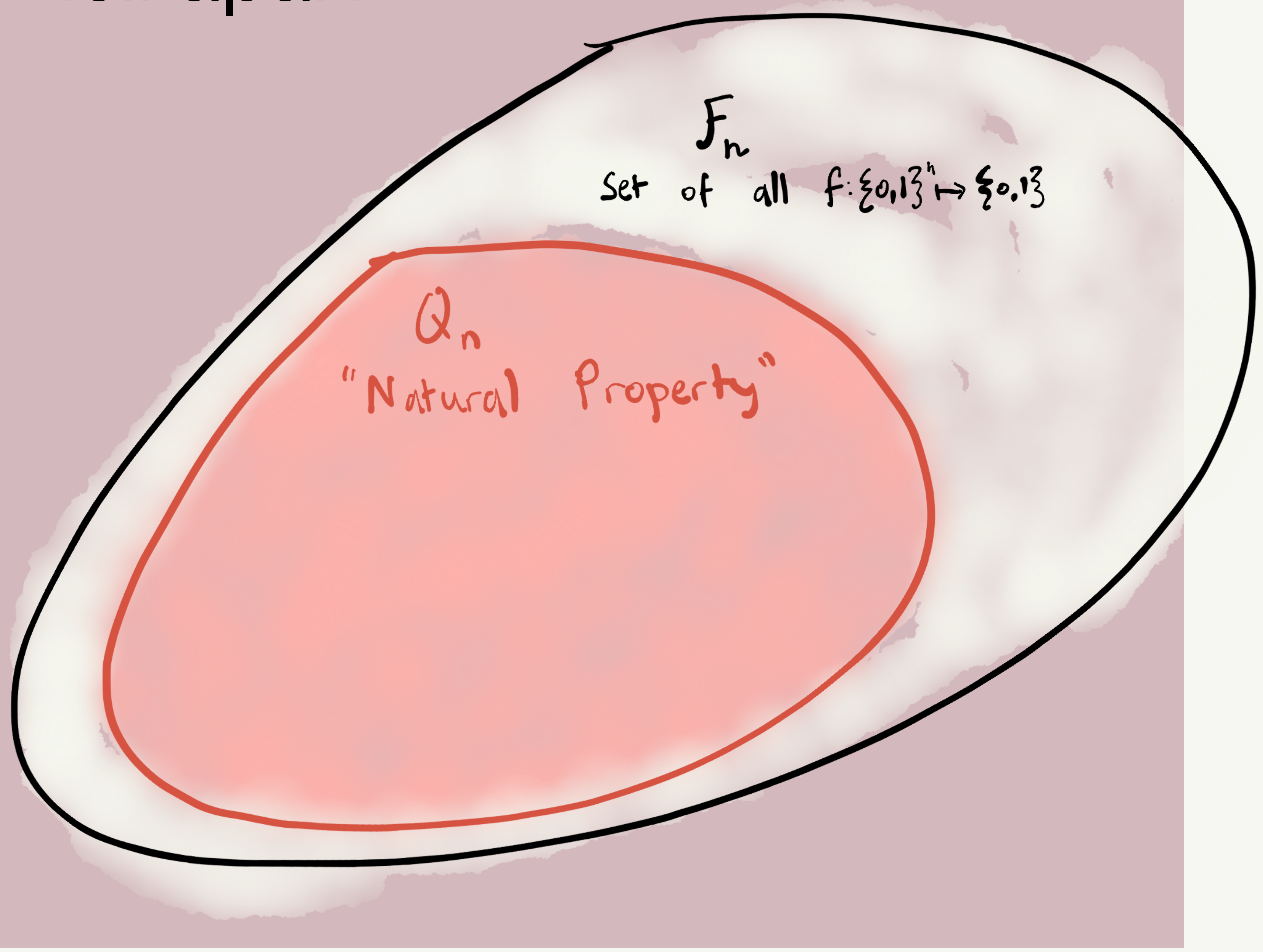
- The secret—look inside proofs
- Complexity (and cryptography) are famous for **reductions**
- **Fundamentally, reductions are algorithms**
- More broadly, **constructive proofs are algorithms**



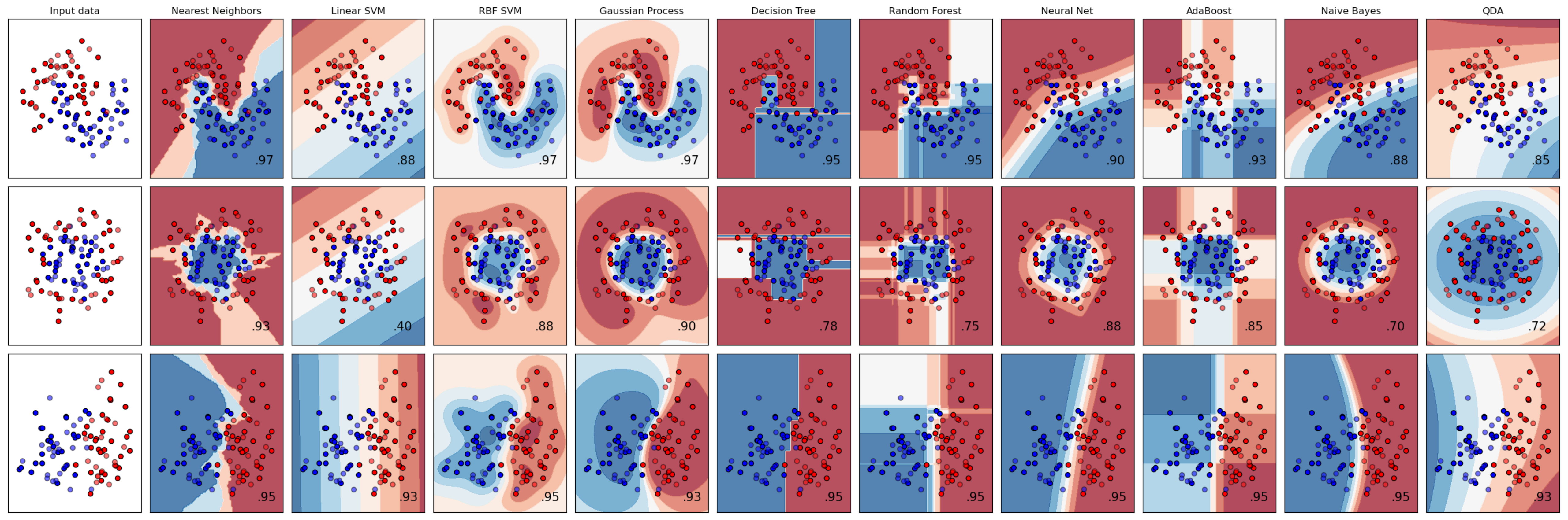


Natural Proofs

- **P/poly vs NP** — the million dollar problem (“millennium prize”)
- Razborov-Rudich (JCSS, 1997): “Natural Proofs” are lower bounds for **circuits** that **encode algorithms** which “tell apart” **structured** functions from **random** functions.
- Carmosino-Impagliazzo-Kabanets-Kolokolova (CCC, 2016): merely **distinguishing structure** from **randomness** is enough to learning the circuit! **Best Paper Award!**
- This algorithm uses queries and only works w.r.t. a uniform distribution over unlabelled examples.



Natural Proofs ⟶ Learning from Data



- Carmosino-Impagliazzo-Kabanets-Kolokolova (CCC, 2016): merely **distinguishing structure** from **randomness** is enough to learning the circuit! **Best Paper Award!**

- Learning uses queries and only works w.r.t. a uniform distribution over unlabelled examples

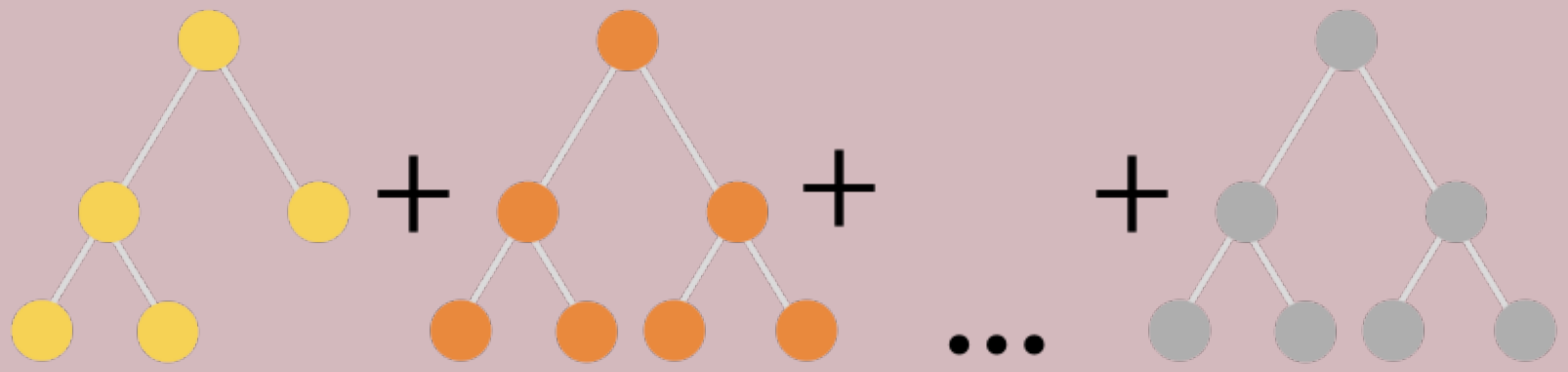
- We wish we could get such a **result for “learning from random data”**

- A **beautiful** way to show that **machine learning theory** is **central** to modern **complexity theory**

Natural Proofs ⟶ Learning from Data

- **Karchmer** (ITCS, 2024): Consider a restricted class of natural proofs. Then we get learning from random data — for the average concept **Best Student Paper Award!**

- Introduces a new model of non-worst case learning, with plenty of independent benefits

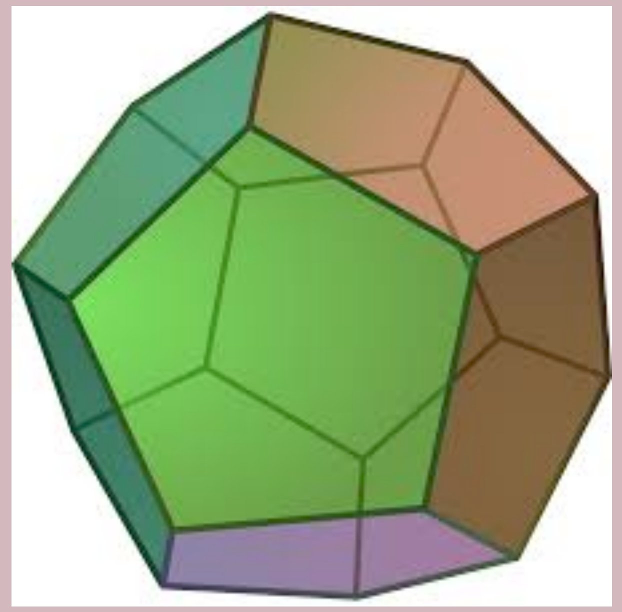


Average-case Boosting

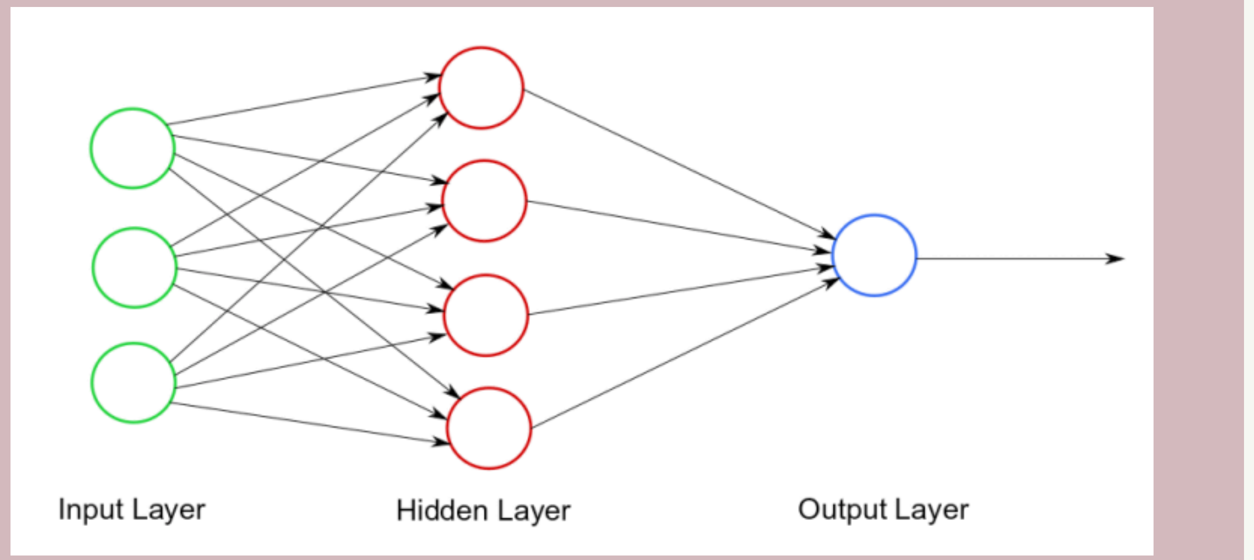
Coexists with cryptography

Natural Proofs ⟶ Learning from Data

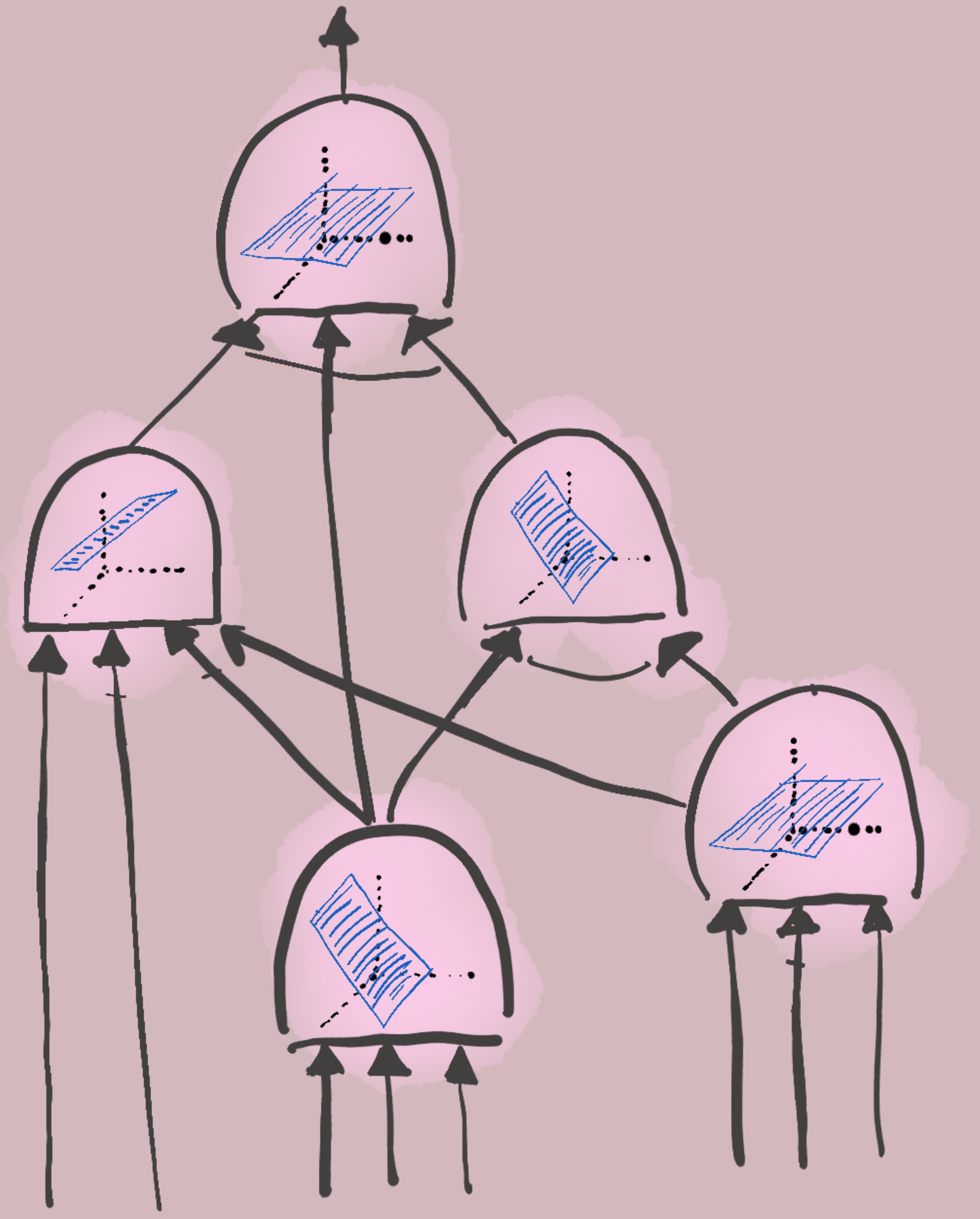
- **Karchmer** (ITCS, 2024): Consider a restricted class of natural proofs. Then we get learning from random data — for the average concept **Best Student Paper Award!**
- Introduces a new model of non-worst case learning, with plenty of independent benefits



Karchmer (ITCS, 2024): Use **existing** complexity **lower bounds** to **derive novel algorithms** for **PAC-learning** distributions over **convex bodies** and **depth 2 threshold networks (bounded weights)**



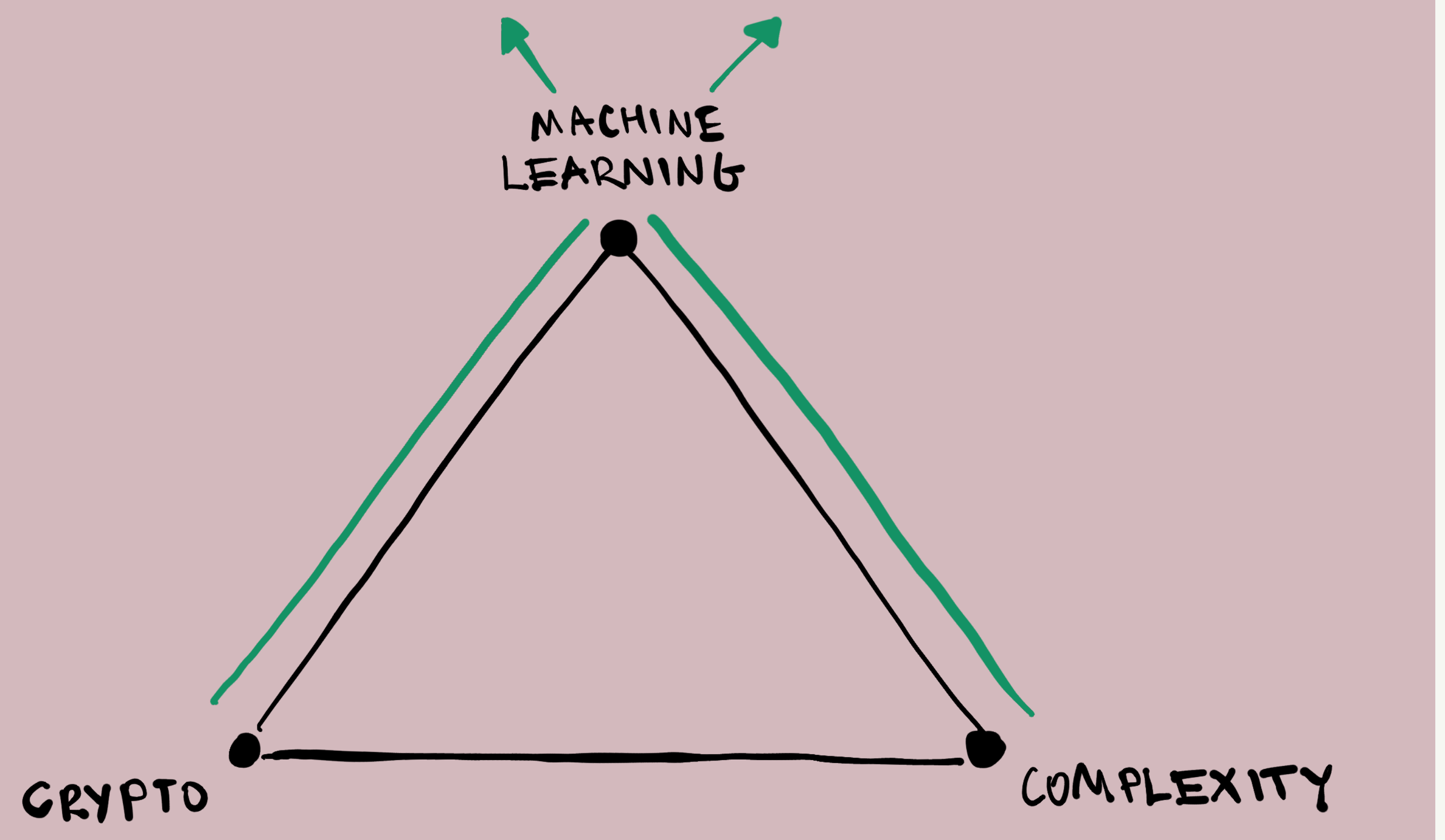
Natural Proofs ⟶ Agnostic Learning and Compression



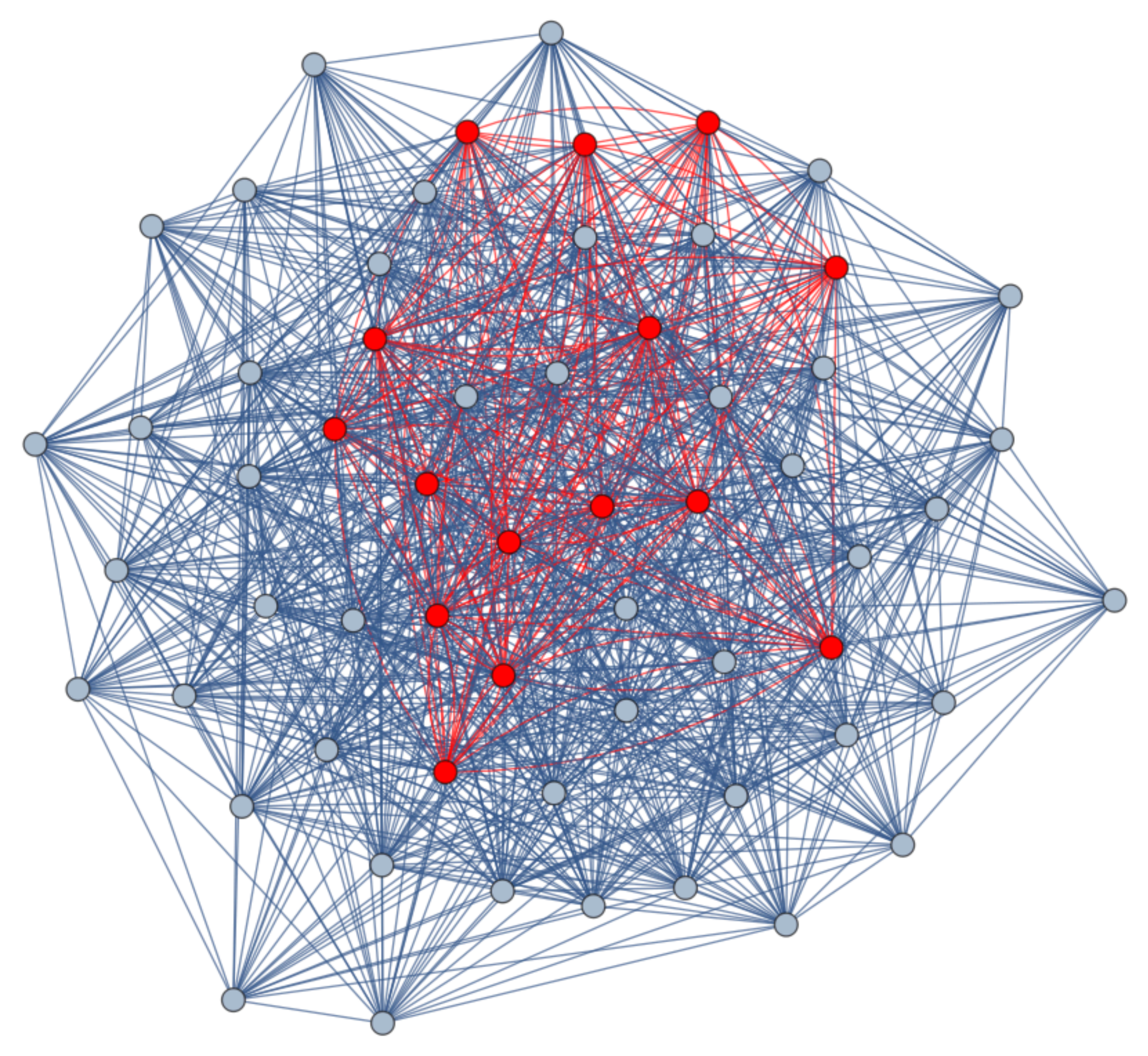
- **Karchmer** (ALT, 2024): Consider the same restricted class of natural proofs
- Use them to obtain new agnostic learning and compression algorithms for small circuits with threshold gates (~DNNs)
- Continues a line of study initiated by Servedio-Tan (ITCS, 2017) on “nontrivial learning”/compression from lower bounds

Cryptography and complexity theory in the design and analysis of machine learning

Future directions.



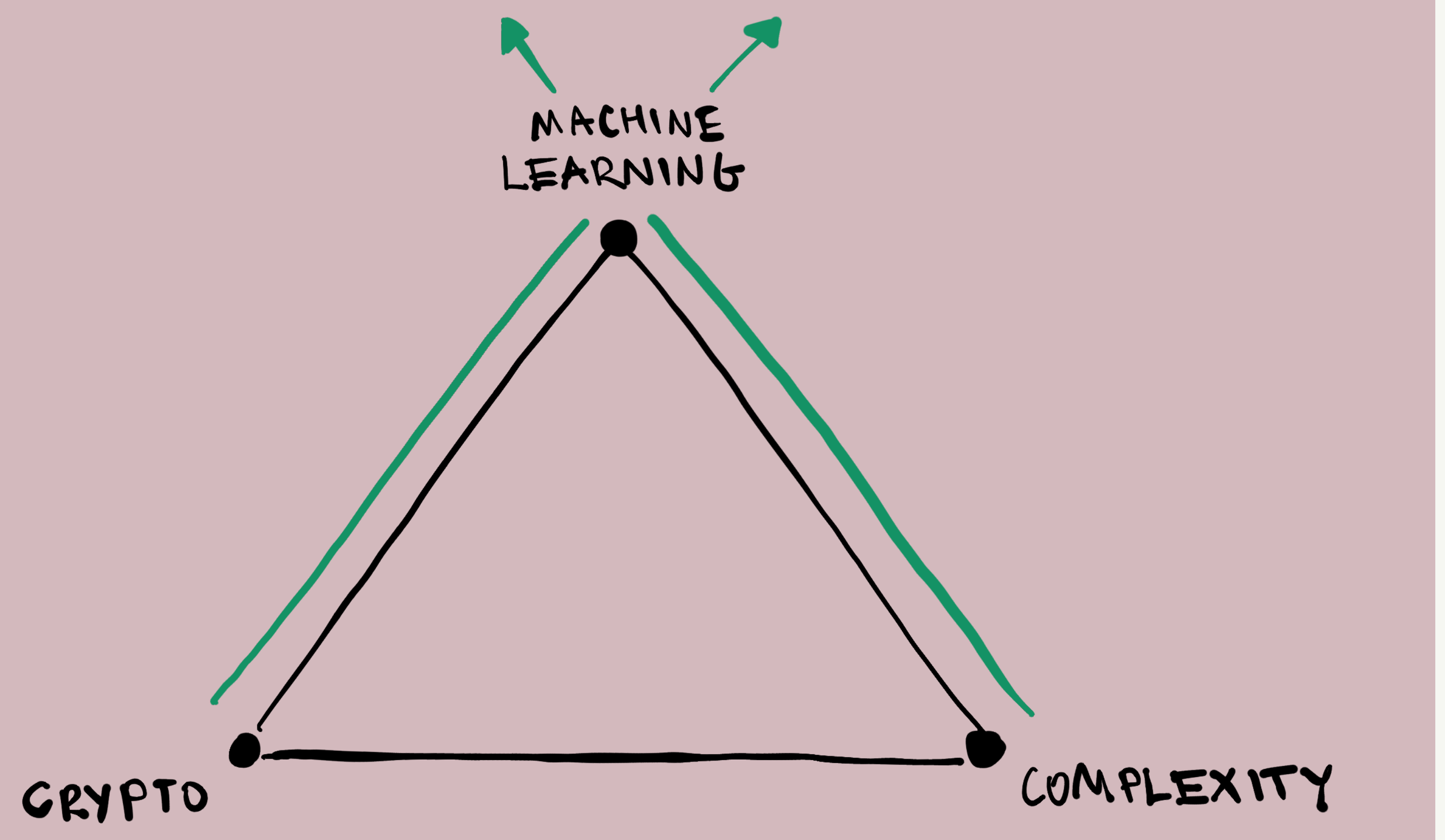
- More ways of using Crypto and Complexity to heuristically reason about ML
    - Used fine-grained crypto to understand more natural multimodal ML separations (Lavigne et al., CRYPTO 2019)
    - Planted statistical problems could be the key

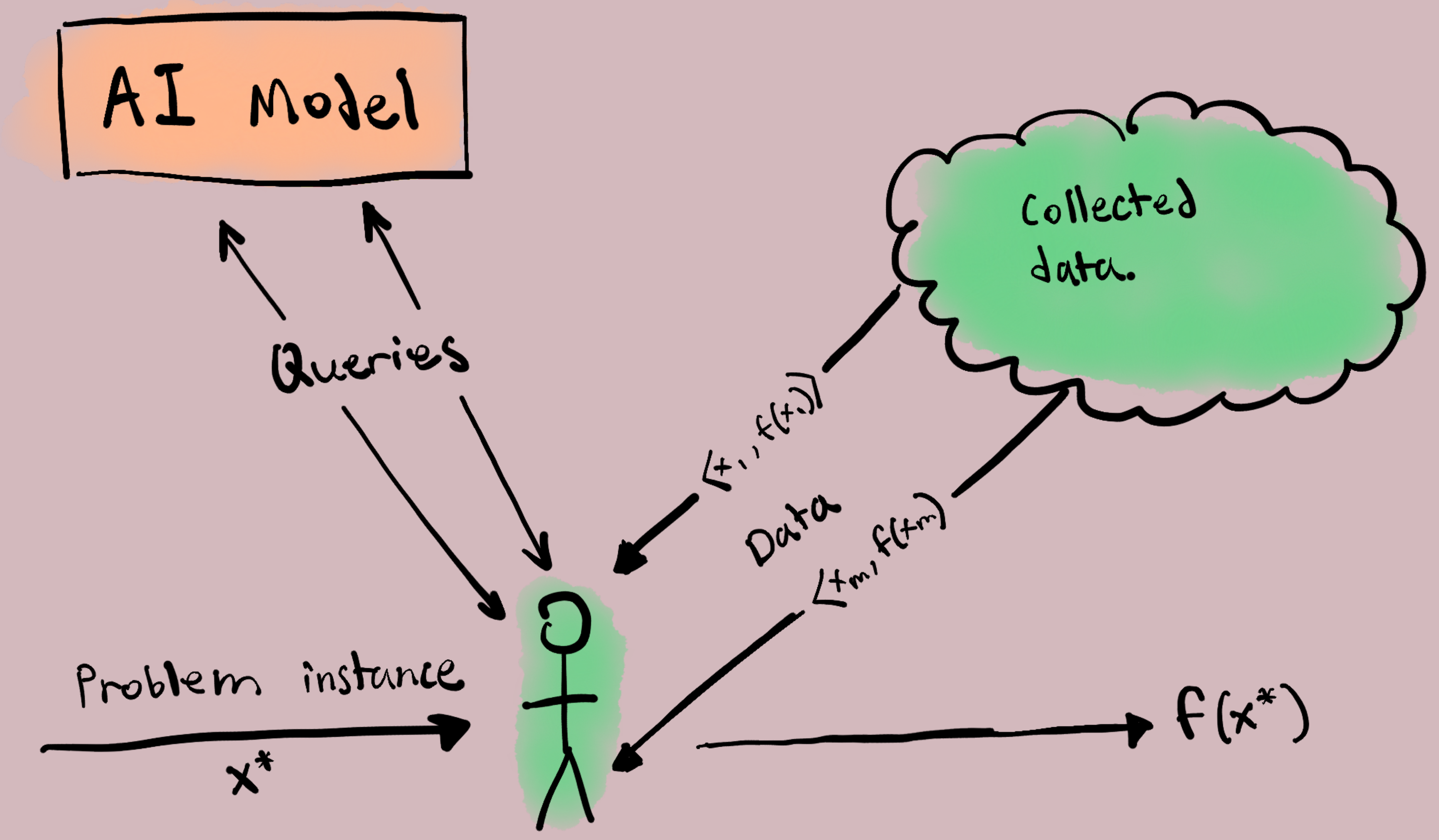


Cryptography and complexity theory in the design and analysis of machine learning

Future directions.

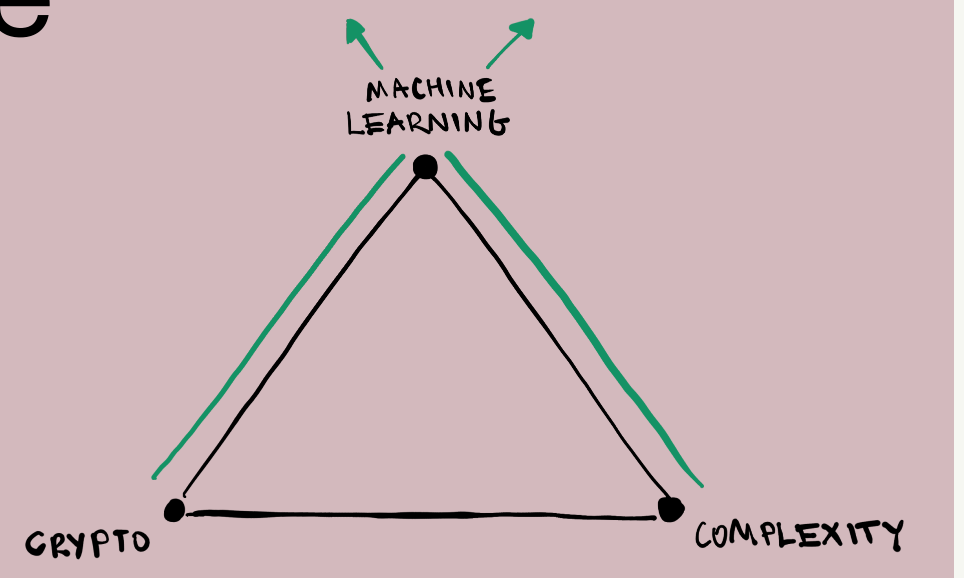
- Applications of Covert Learning to AI Safety
  - AI Jailbreaking
  - Can we show that: Covert Learning provides a way to interact with an AI model **undetectably?**
    - Inherently **unalignable?**
  - Current “suffix optimization” LLM jailbreaks are not undetectable (e.g. Zou et al., 2023)





This talk: how, and when, can crypto or complexity *positively* impact ML?

- Both crypto and complexity can help us reason about the ML "real world" (e.g., why is training on text **and** images more effective than training on just text?)
- Crypto can help us design more secure and private ML algorithms
- Complexity theory can give us technical machinery for faster and more robust ML algorithms



Thanks for listening! Q+A

—Ari Karchmer
Boston University