

AnoMed Seminar



**ESFANDIAR MOHAMMADI
UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR IT-SECURITY
PRIVACY & SECURITY GROUP**

11 May 2023

AnoMed Seminar

- Joint seminar series of the AnoMed partners: every other Thursday, 11:00 - 11:50
 - 50 minutes, 30 - 40 minutes talk, 10 - 20 minutes discussion
- Each partner presents challenges (phase I) and existing / initial results (phase II)
 - Hopefully that sparks collaboration beyond the promised results
- We coordinate
 - requirements analysis (Anforderungsanalyse)
 - adjust project goals and timelines
 - plan congress / conferences
 - project reports
- Today:
 - Privacy challenges in machine learning & initial results (Esfandiar Mohammadi)
 - Initial results & paper:
defending against privacy attacks in video classification (Nele Brügge)

KI-Kongress

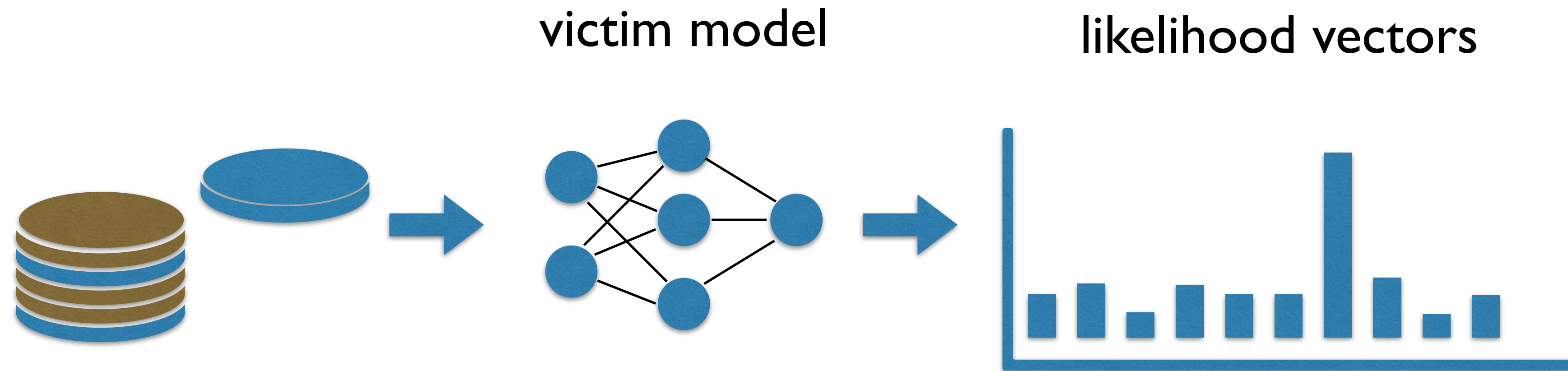
- Next big event
 - Together with KI-SIGS (which will take place at another day)
- Date: 12, 13, or 14 September 2023
- Plan: each partner presents initial results
- We also invited 4 associated anonymity projects (from the Forschungsnetzwerk Anonymität)

Roadmap

- ▶ **REPETITION: PRIVACY ATTACKS BASED ON CONFIDENCE**
- **REPETITION: PRIVACY ATTACKS BASED ON BOUNDARY DISTANCE**
- **REPETITION: LOSS-BASED ATTACK**
- **REPETITION: RECONSTRUCTING TRAINING DATA POINTS**
- **PRIVACY ATTACKS ON LARGE LANGUAGE MODELS**

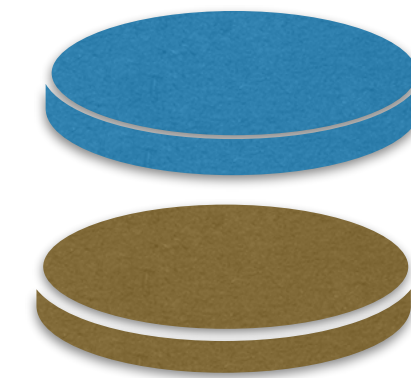


Membership Inference -- Model Confident on Training Data

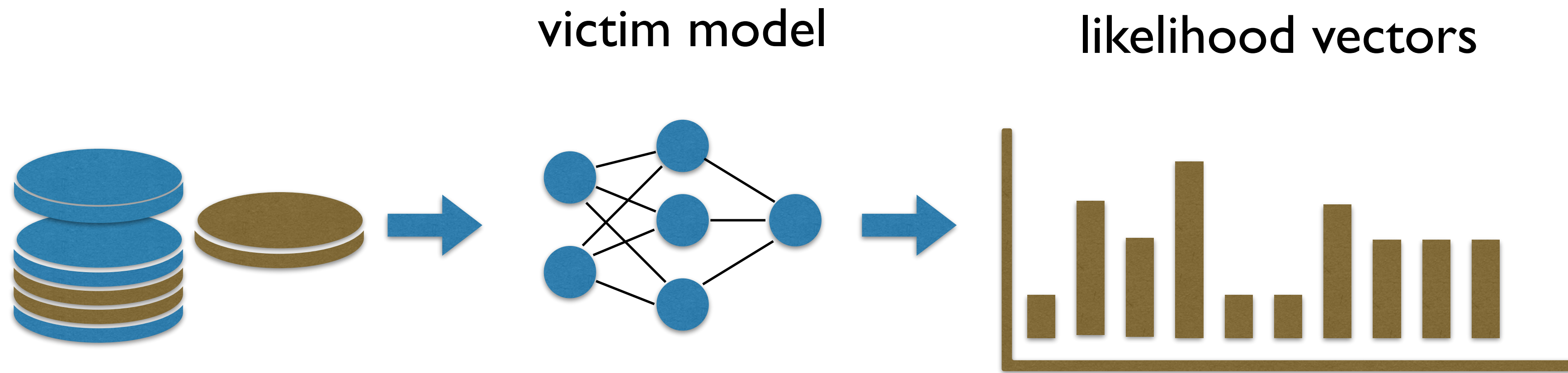


bars more pronounced for trainings data
(likelihood vector has low entropy)

- Data points used during training (member)
- Data points not used during training (non-member)

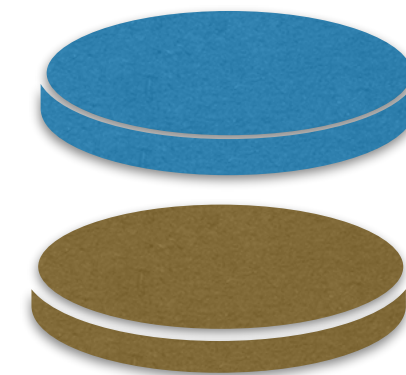


Less Confident on Non-Training Data

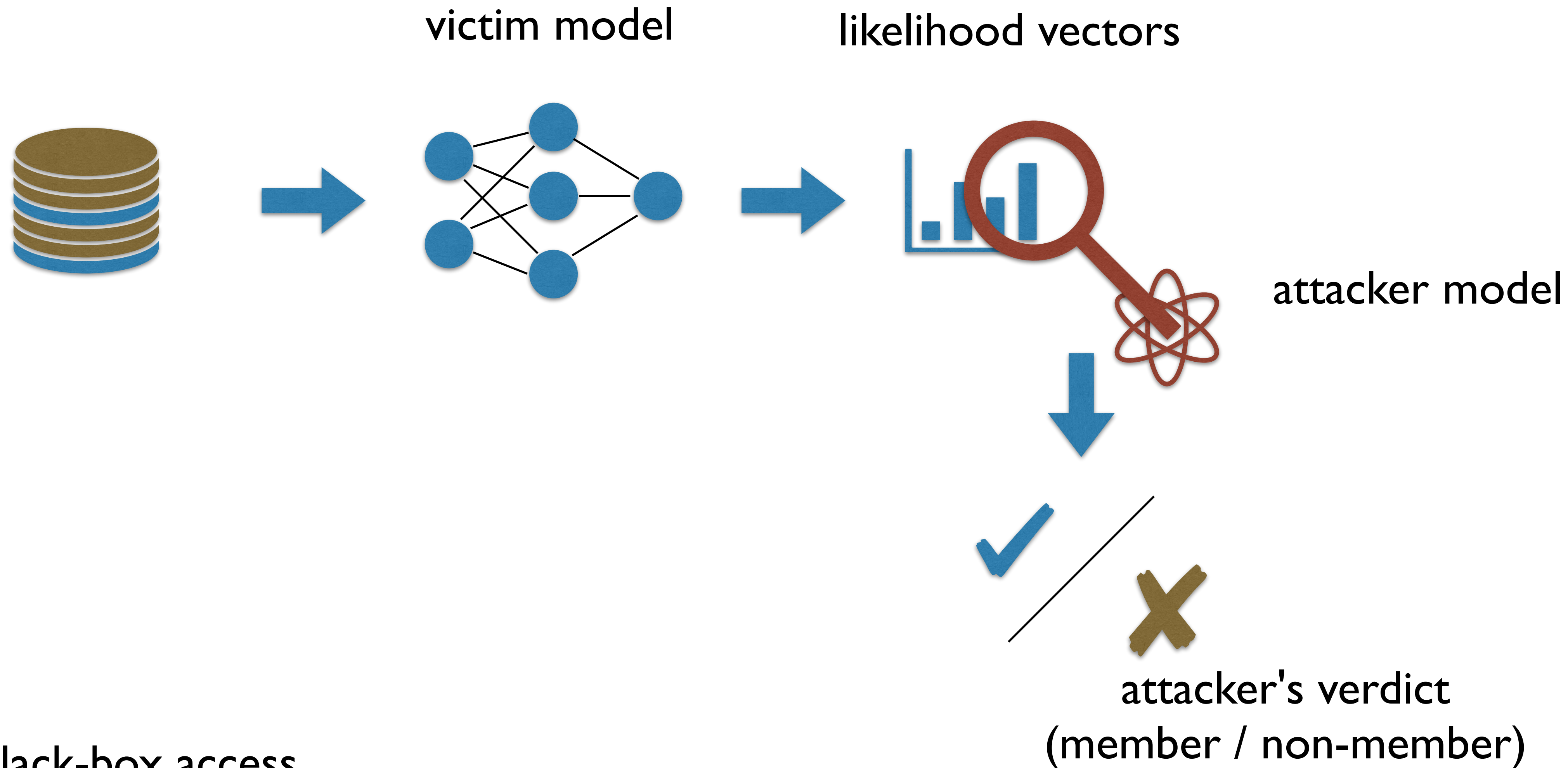


bars less pronounced for trainings data
(likelihood vector has higher entropy)

- Data points used during training (member)
- Data points not used during training (non-member)

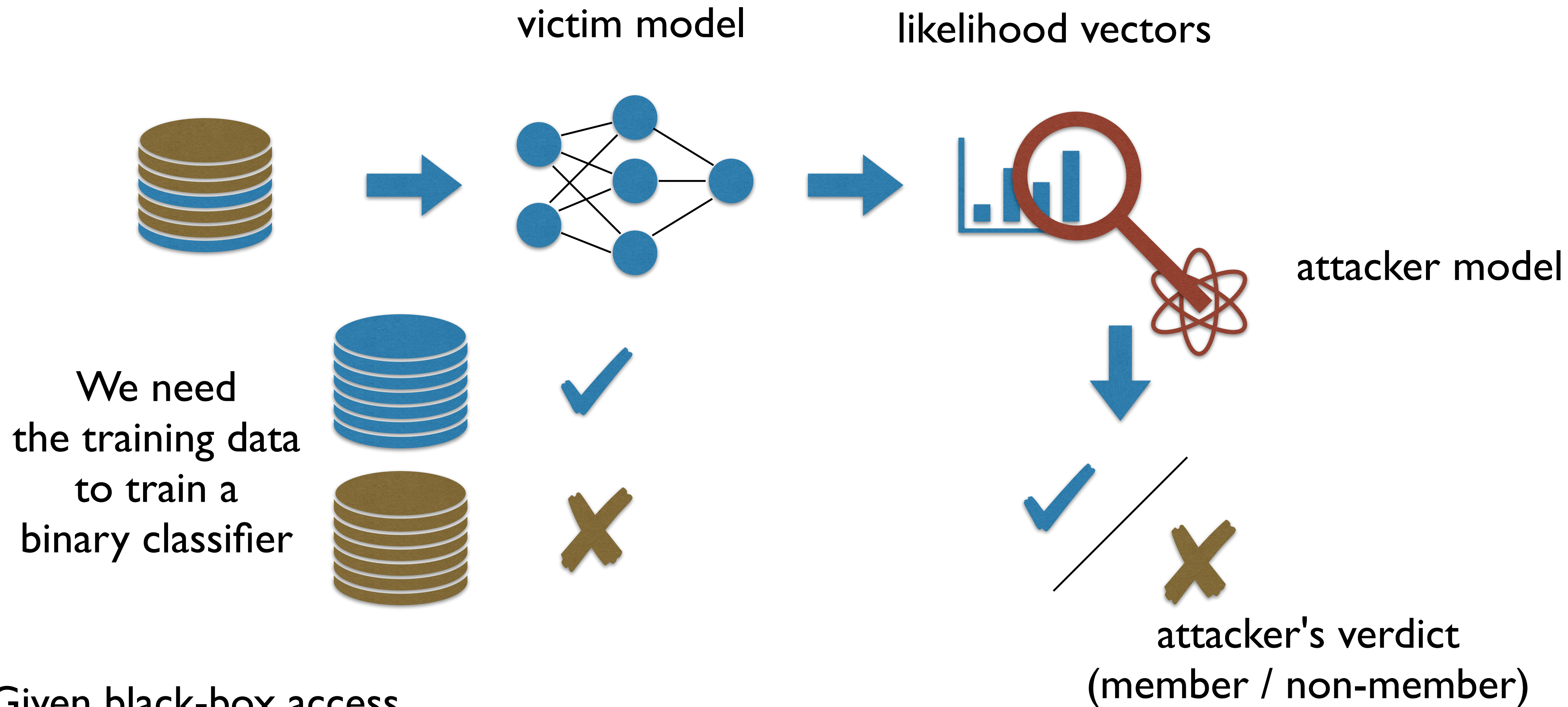


Goal: Attacker model as a binary classifier



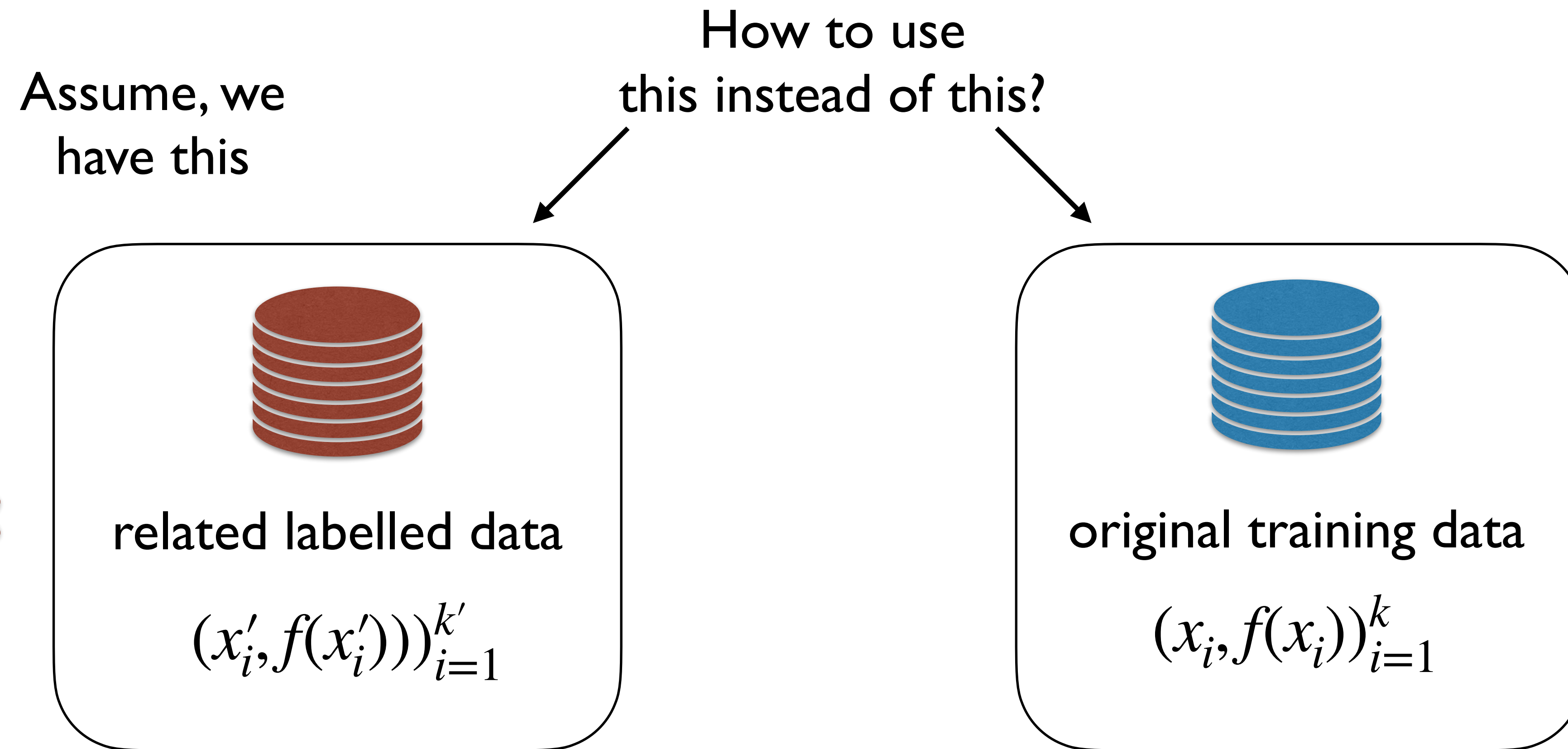
- Goal: Given black-box access, upon receiving the likelihood vector, distinguish members from non-members

Do we need the training data?

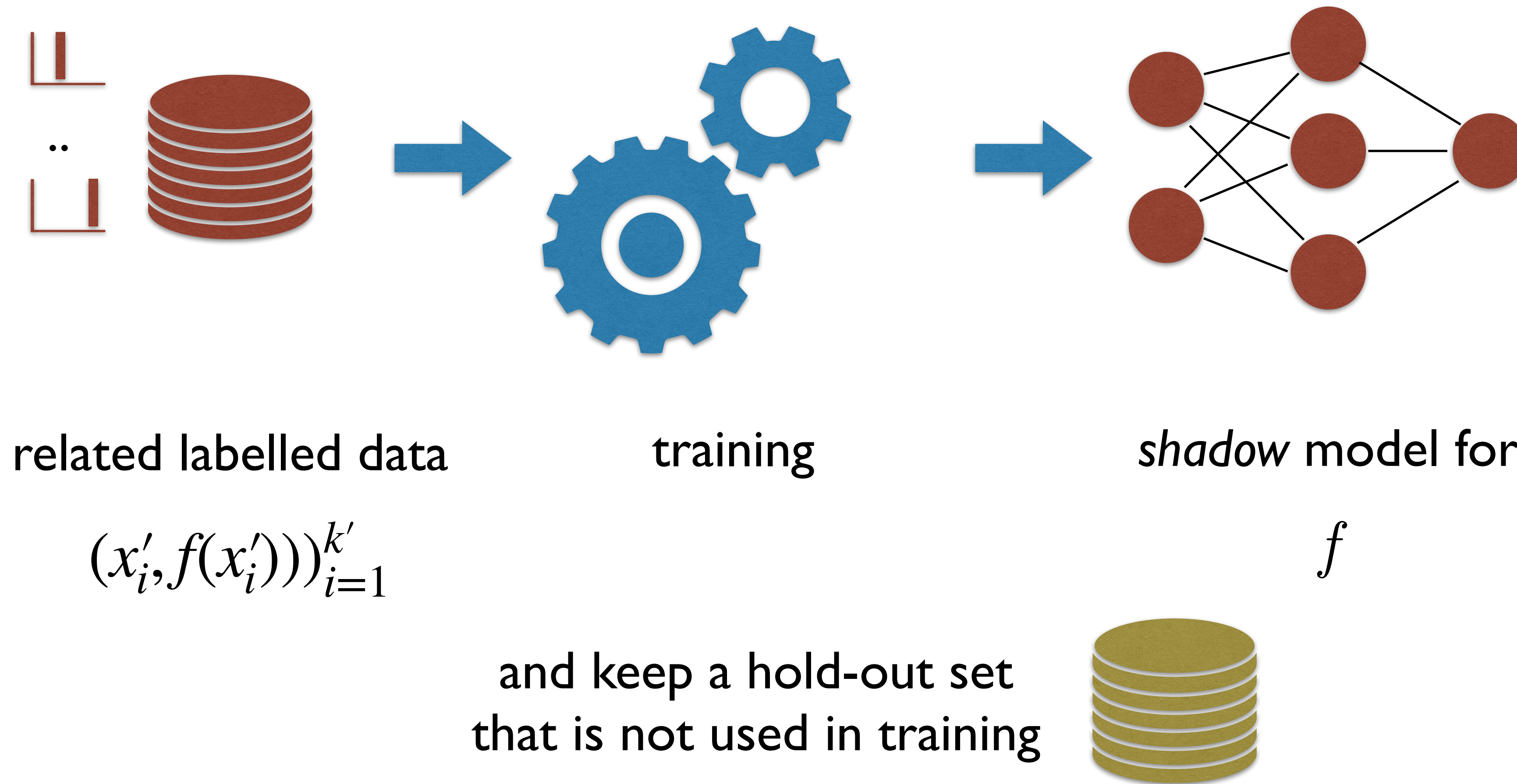


- Goal: Given black-box access, upon receiving the likelihood vector, distinguish members from non-members

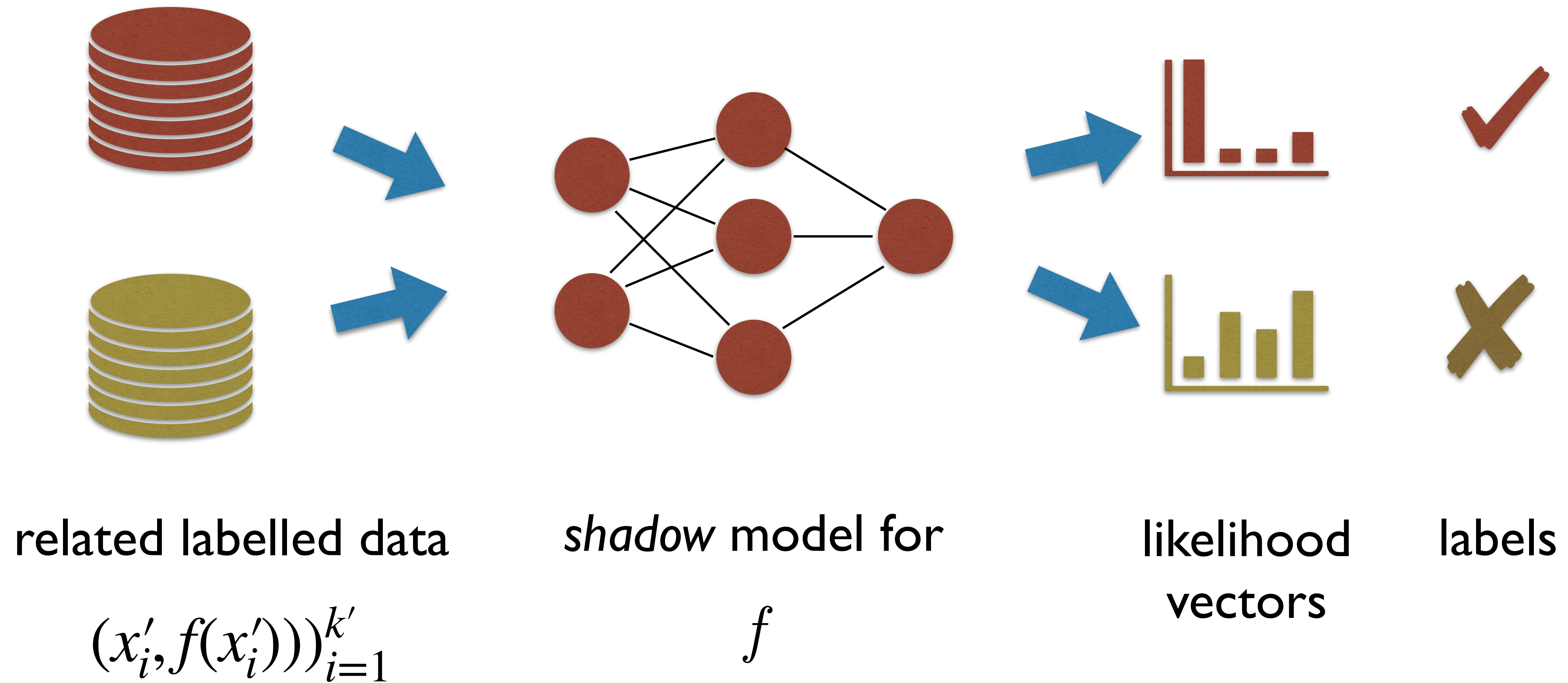
Can we use the related data?



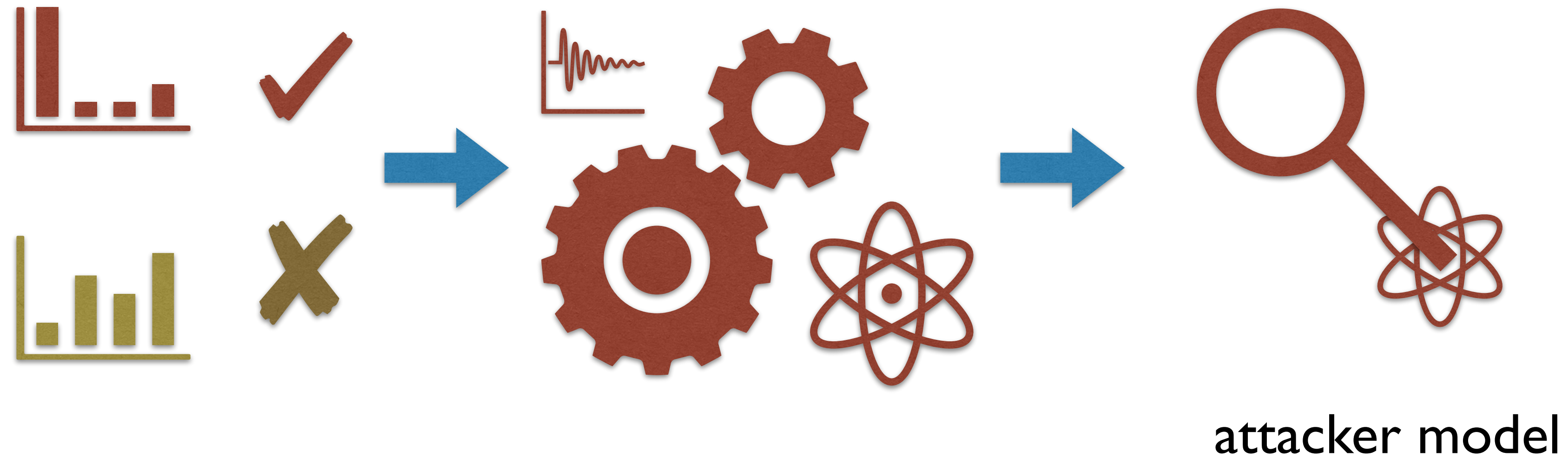
Idea: train your own shadow models

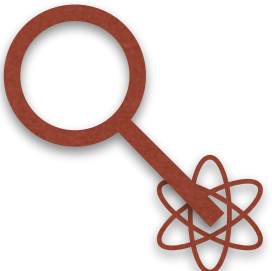
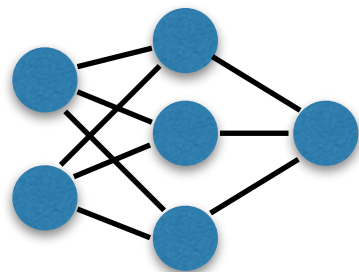


Construct likelihood vectors



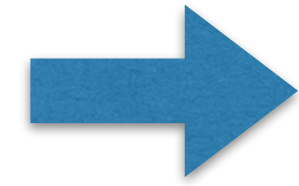
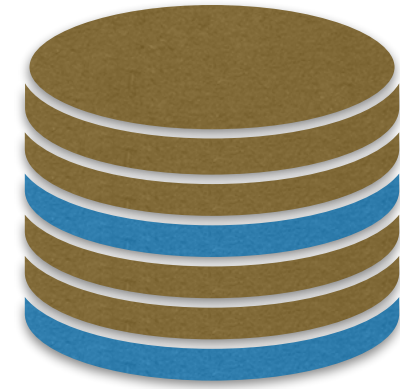
Train the attack model (binary classifier)



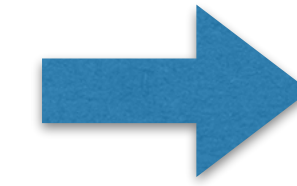
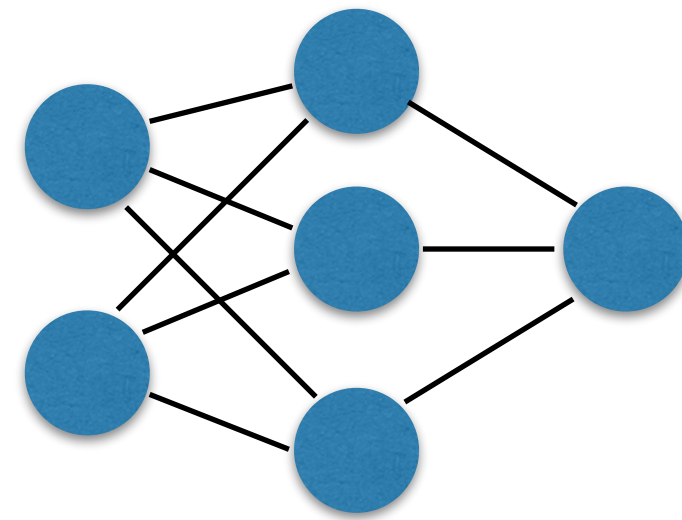
- Input likelihood vectors, labelled with ✓ or ✗
- Output: training data point (member) or no training data point (non-member)
- Attacker model  can be applied without access to the victim model 

Apply the attack model (binary classifier)

candidate data points



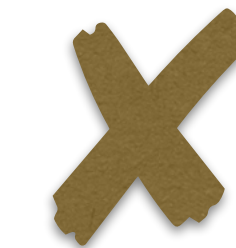
victim model



likelihood vectors



attacker model

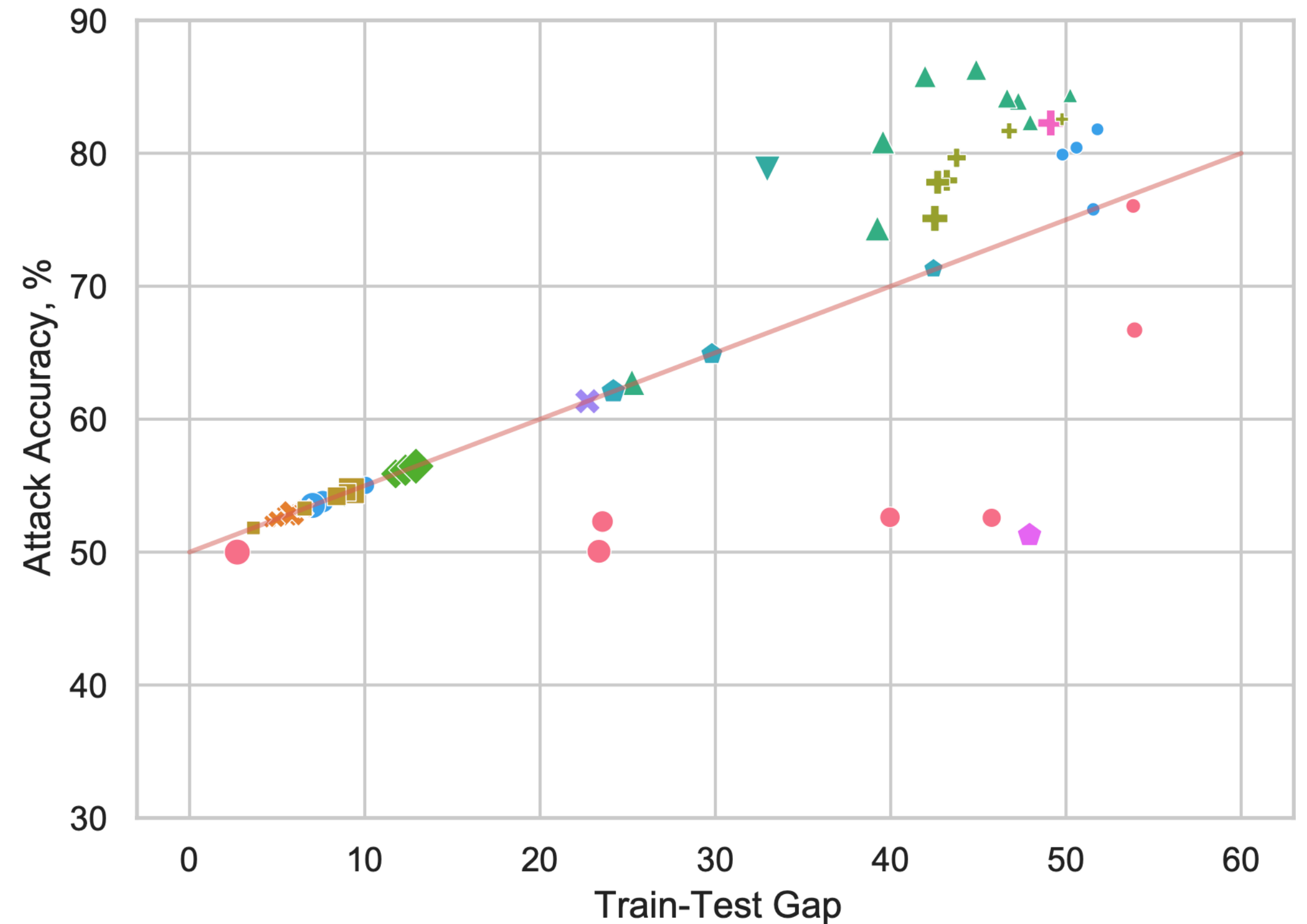


attacker's verdict
(member / non-member)

- Only black-box access to the model is needed
- Shokri et al.:
- For highly overfitted victim model, knowledge about the shape of the likelihood vector generalizes to the victim model

Attack performance

- Different defense types
 - we will get to know the listed Differential Privacy defenses
 - MemGuard: noise the likelihood vector
 - Adversarial regularization: include the attack in the loss function during training
 - The other defenses are standard ML techniques for avoiding overfitting
- Train-Test Gap is the generalization gap: difference between train and test accuracy



—	Baseline Gap Attack	▲	Dropout
●	Adversarial Regularization	▼	Full Fine Tune
×	DP Full Fine Tune	◆	L1 Regularization
■	DP Last Layer	●	L2 Regularization
+	Data Augmentation	×	Last Layer
◆	Differential Privacy	⬠	MemGuard
		+	None

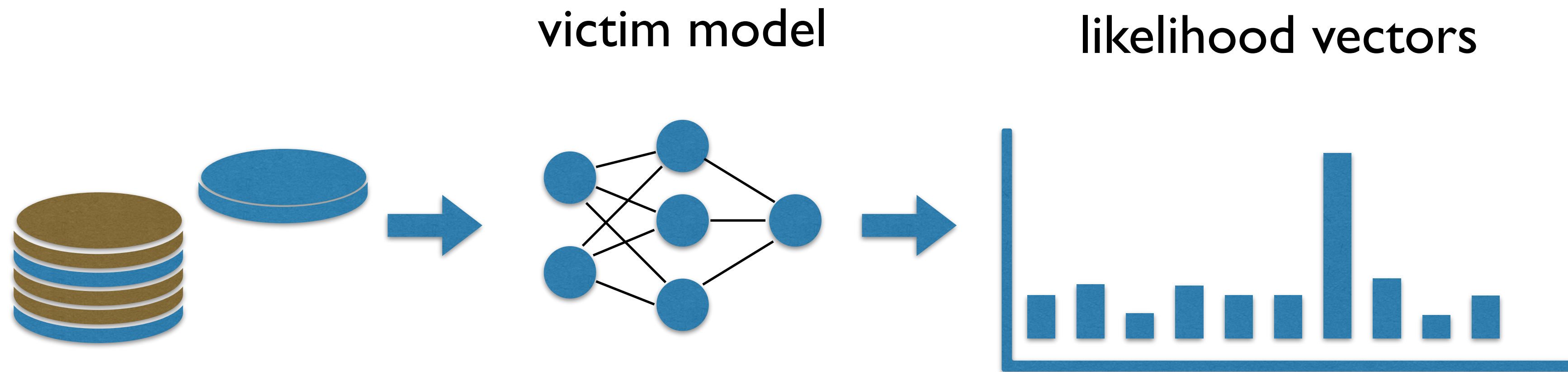
source:
Label-Only Membership Inference Attacks
by Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, Nicolas Papernot
<https://arxiv.org/abs/2007.14321>
In ICML 2021

Roadmap

- **REPETITION: PRIVACY ATTACKS BASED ON CONFIDENCE**
- ▶ **REPETITION: PRIVACY ATTACKS BASED ON BOUNDARY DISTANCE**
- **REPETITION: LOSS-BASED ATTACK**
- **REPETITION: RECONSTRUCTING TRAINING DATA POINTS**
- **PRIVACY ATTACKS ON LARGE LANGUAGE MODELS**



Is the revealing the likelihood vector the problem?



bars more pronounced for trainings data
(likelihood vector has low entropy)

- Previous attack used the likelihood vector
- So why release the likelihood vector?
- Countermeasure: simply output the class with the highest likelihood?

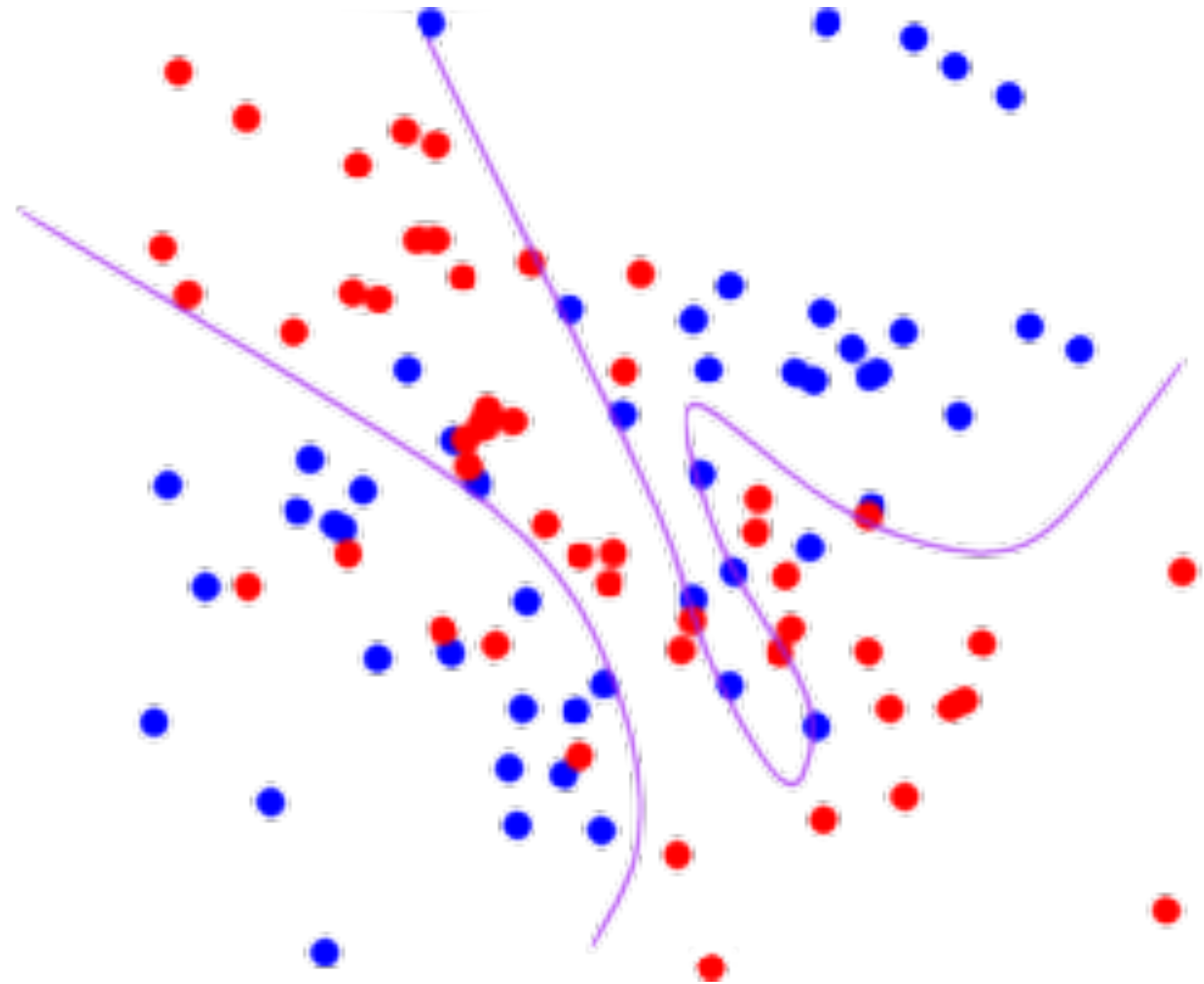
Beyond generalization gap

- Countermeasure:
simply output the class with the highest likelihood?
- Another weakness
- Generalization gap:
 - Difference between training and test accuracy
- So far: exploit generalization gap
 - High certainty \implies guess training point
- Weak against models with very low generalization gap
- There is more



Beyond generalization gap

- Neural networks have wild decision boundaries on complex problems
- Only behave orderly on training data
- Sometimes chaotic on test data

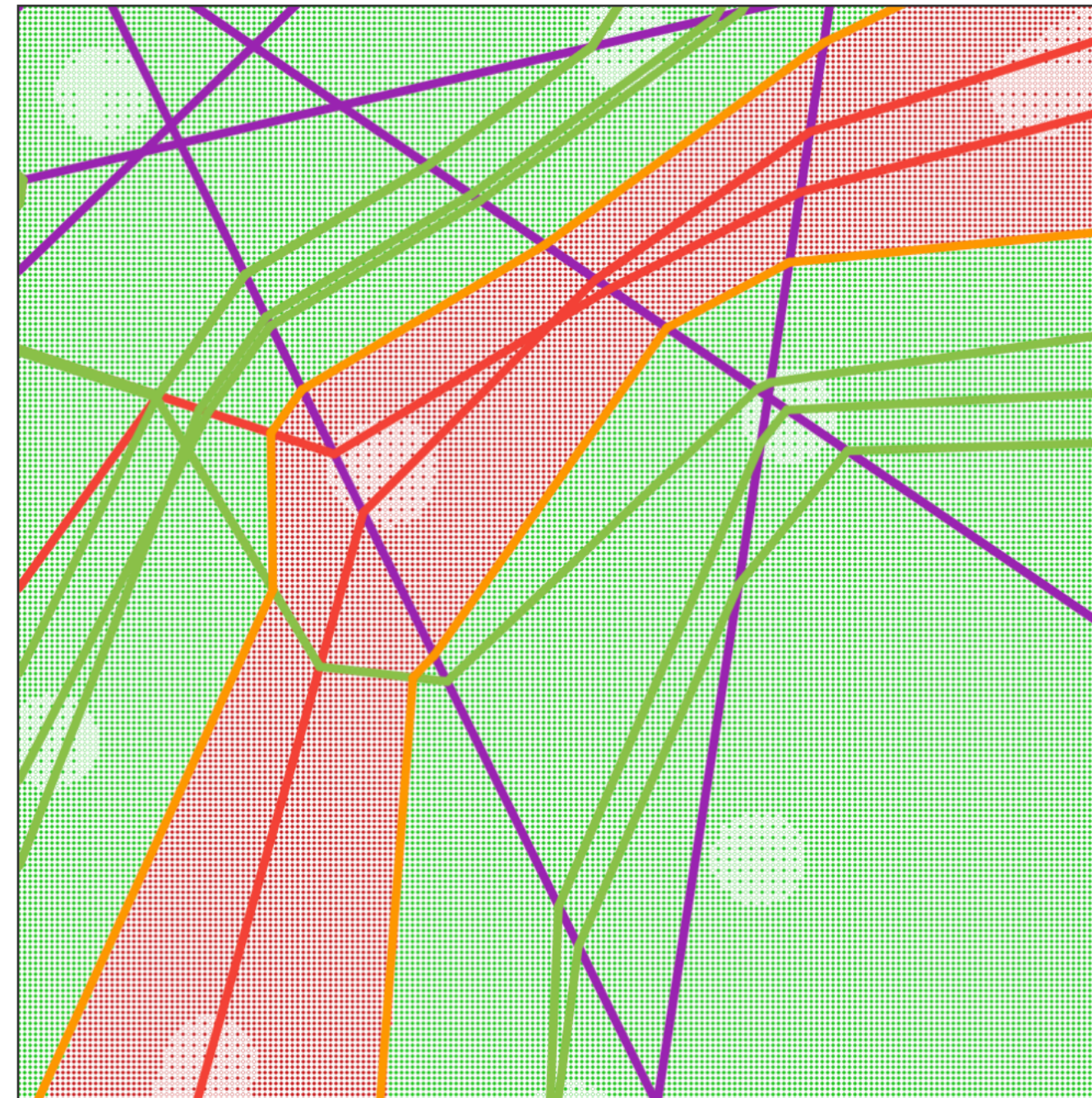


decision boundaries of a neural network

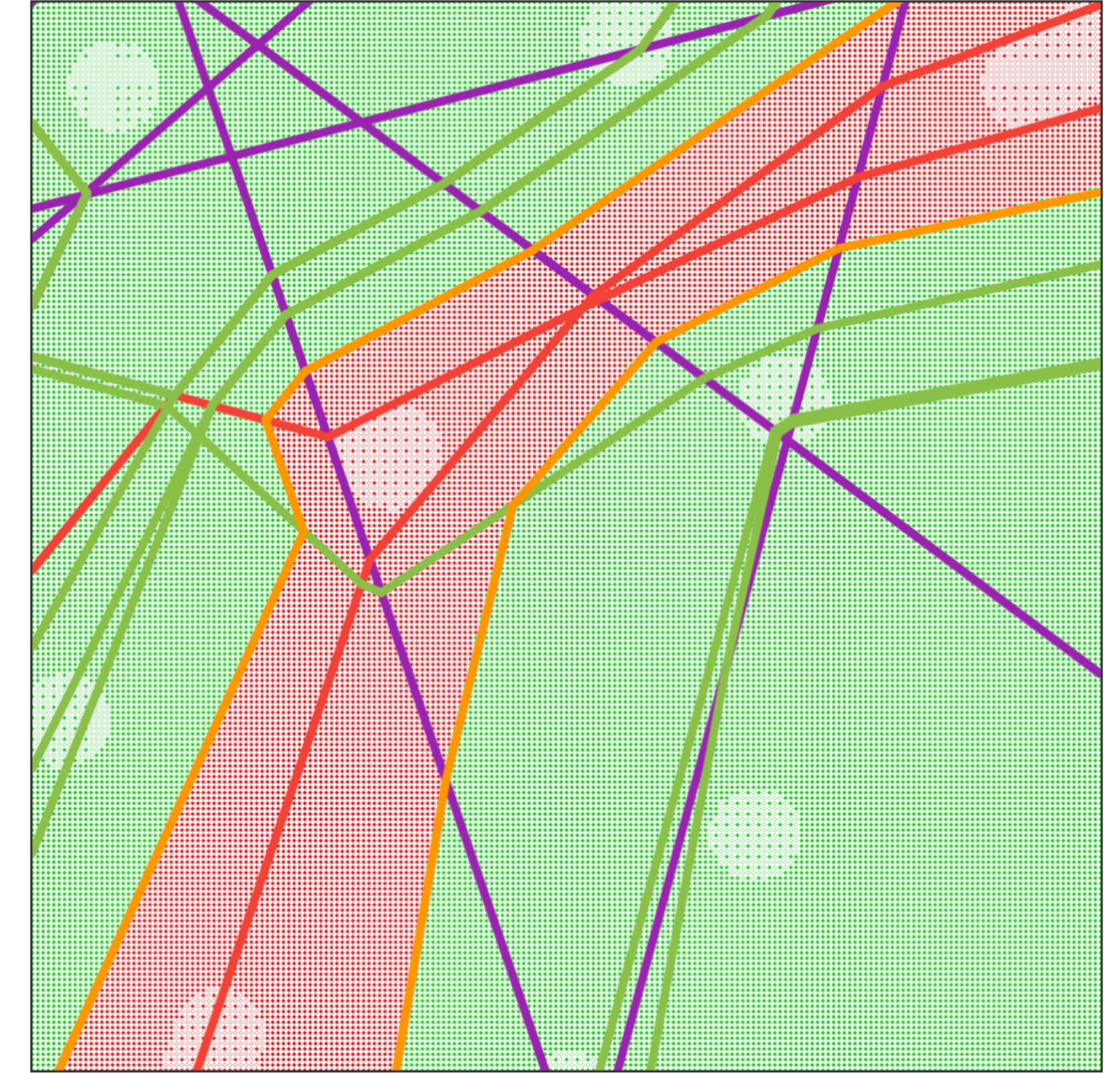
source: <https://www.kdnuggets.com/2015/06/decision-boundaries-deep-learning-machine-learning-classifiers.html>

Artifacts of overfitting (cont'd)

- Lines correspond to neurons
- White points:
 - Training data points
 - 2D inputs, 2 classes
- Prediction
 - Green area: class 1
 - Red area: class 2
- After overfitting:
boundaries specialize
more on training data points



Normal training

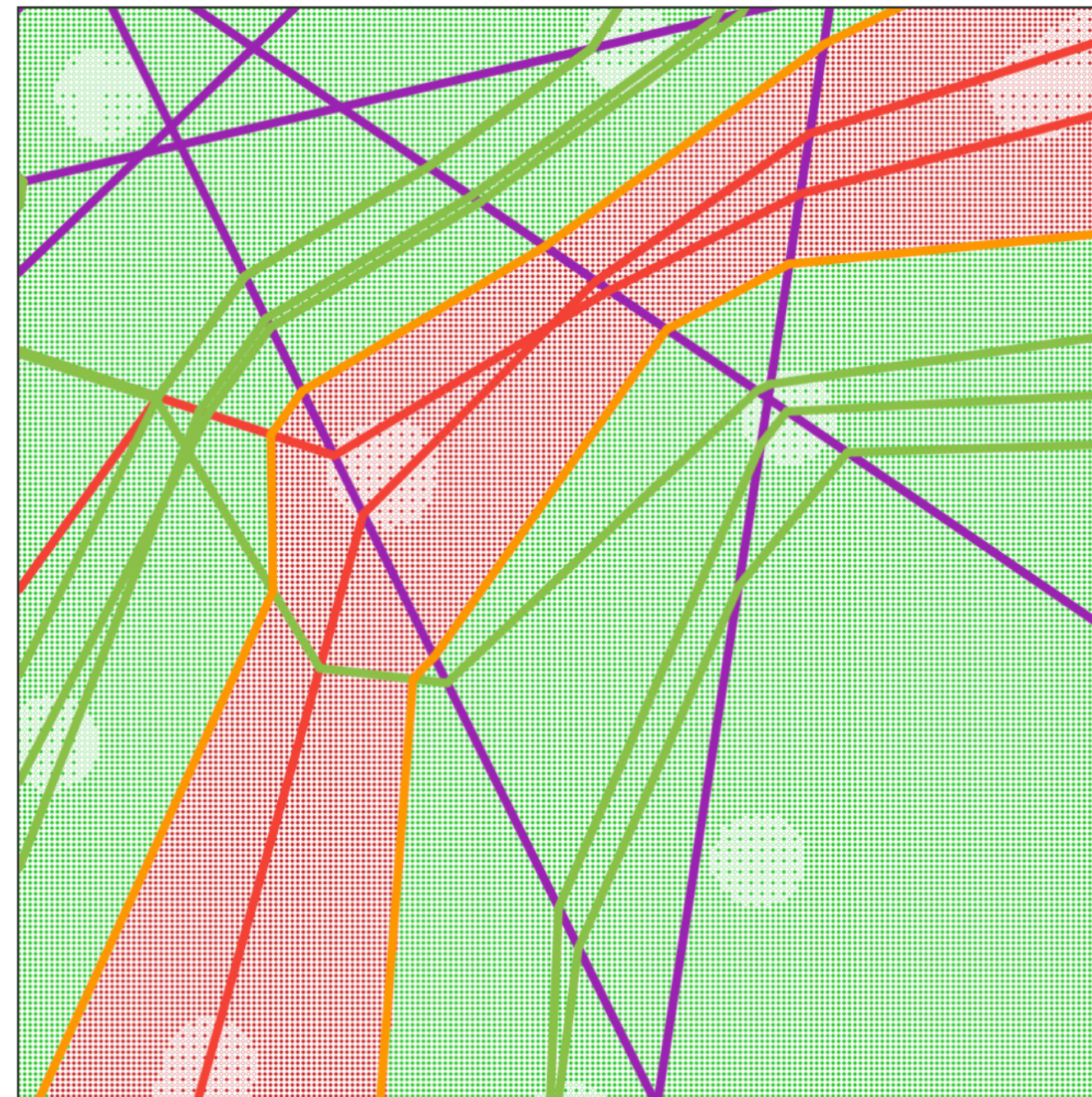


After overfitting

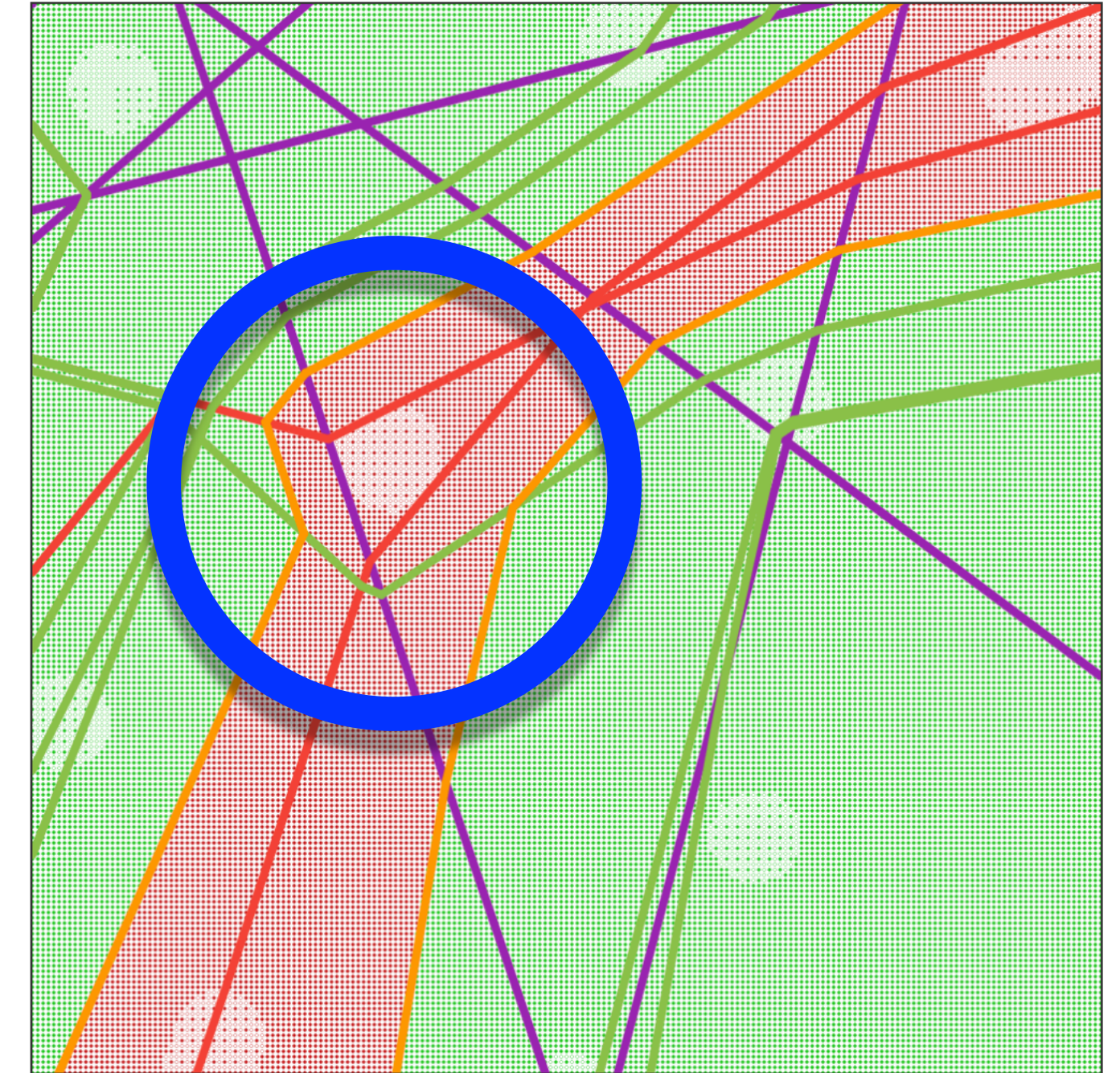
images from a tool by Felix Mächtle

Artifacts of overfitting (cont'd)

- Lines correspond to neurons
- White points:
 - Training data points
- 2D inputs, 2 classes
- Prediction
 - Green area: class 1
 - Red area: class 2
- After overfitting:
boundaries specialize
more on training data points
- How to use that?



Normal training



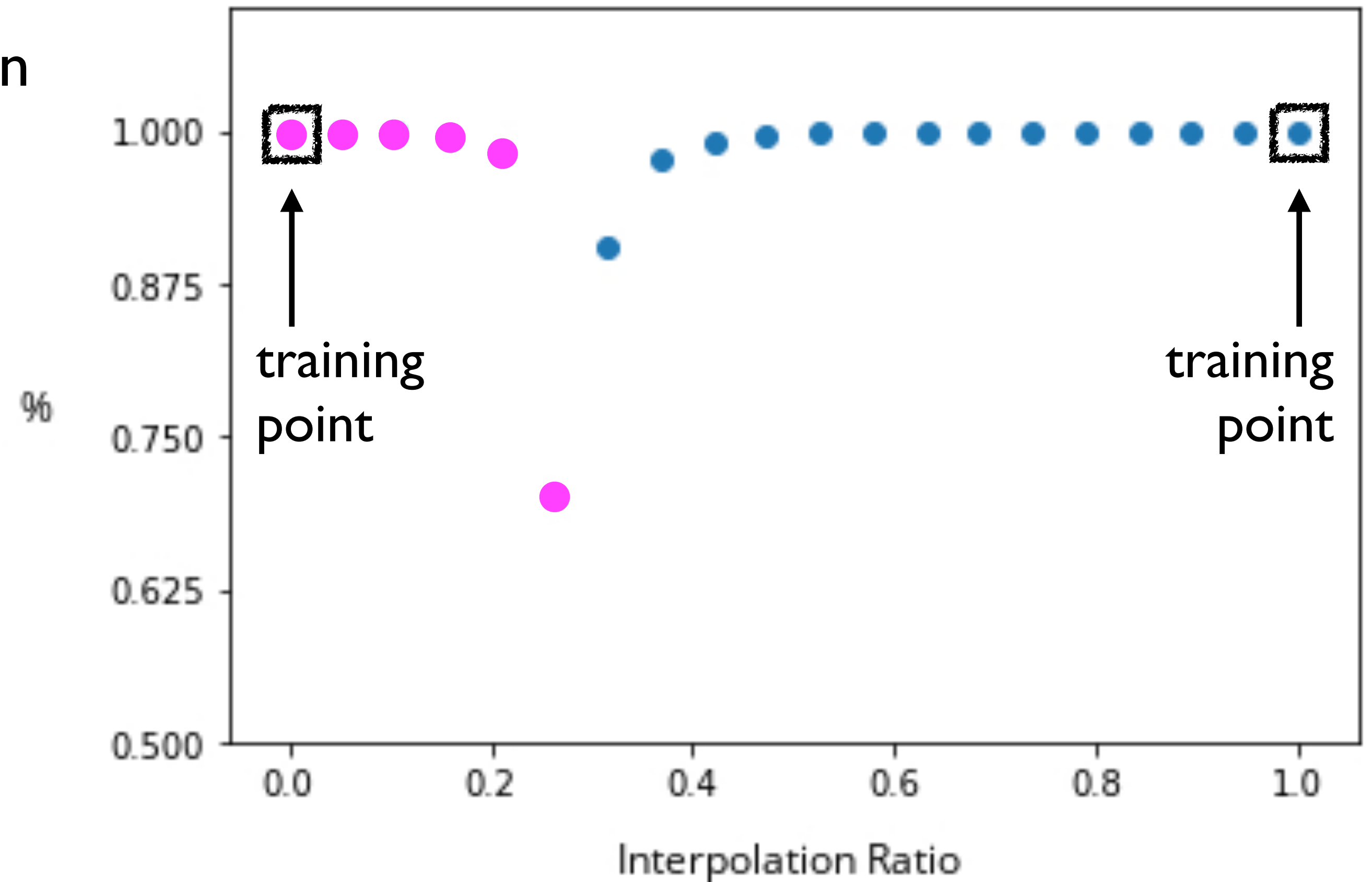
After overfitting

images from a tool by Felix Mächtle

Shape of decision boundaries: training data

- Privacy / ATIS WS 19/20
-- Student Project Thorsten Peinemann
- Plot the certainty on the path from one training point to another training point
- Dataset: CIFAR-10
- Looks kind of orderly

● class 1
● class 2

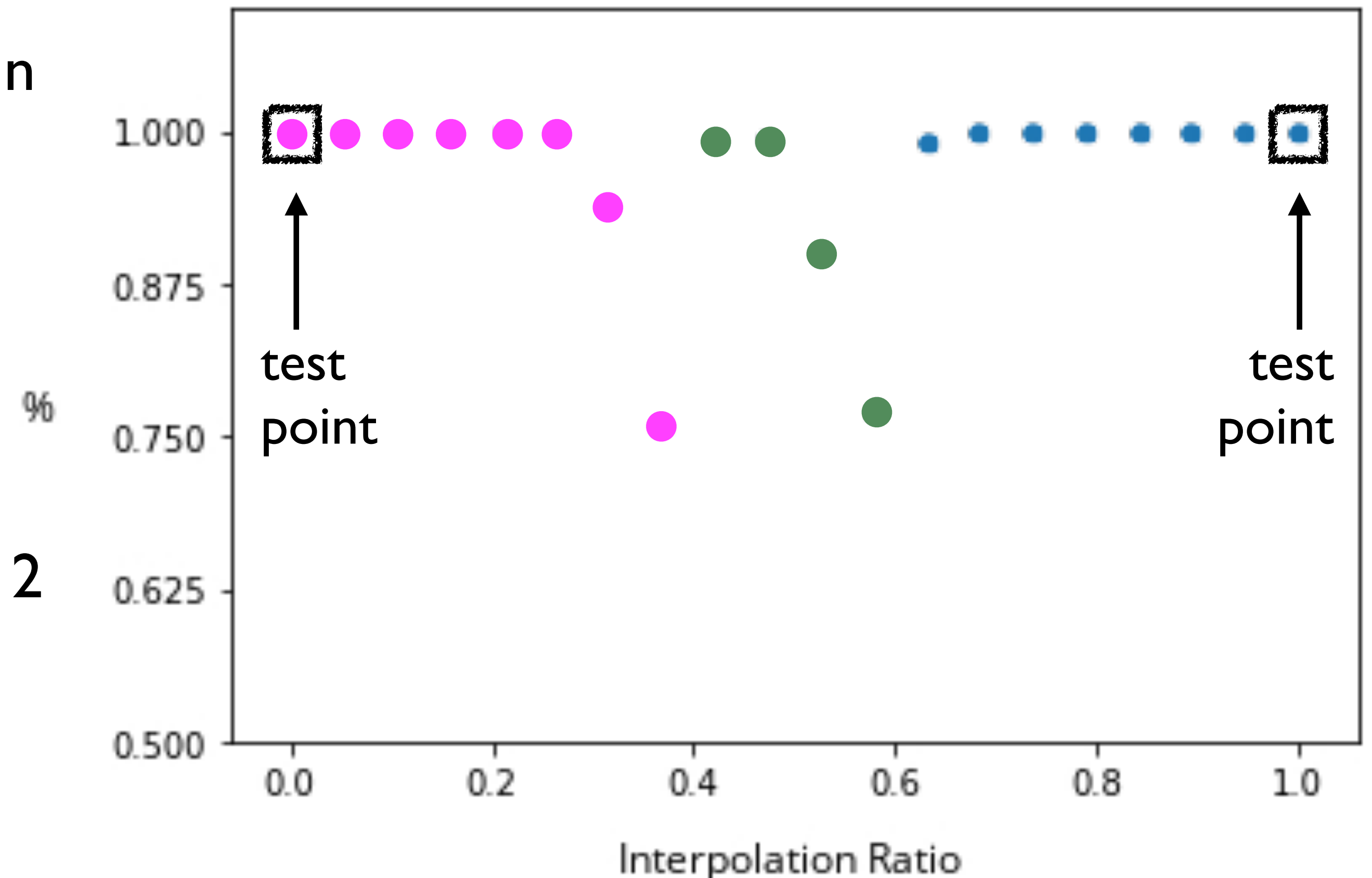


source: Thorsten Peinemann, Mini-project ATIS, WS 19/20

Shape of decision boundaries: test data

- Privacy / ATIS WS 19/20
-- Student Project Thorsten Peinemann
- Plot the certainty on the path from one test point to another test point
- Dataset: CIFAR-10
- Looks more messy

- Class 3 in-between class 1 and class 2

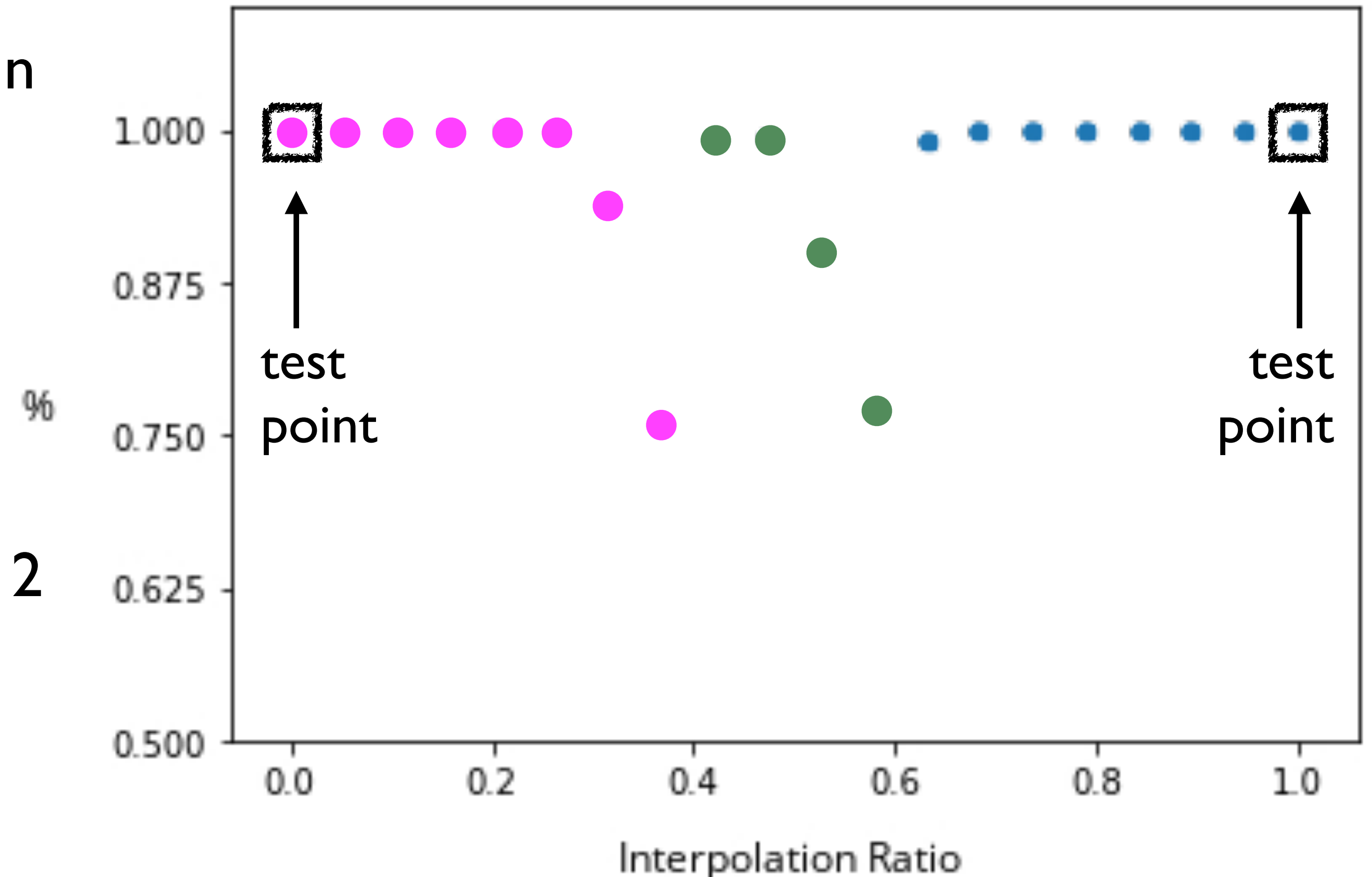


source: Thorsten Peinemann, Mini-project ATIS, WS 19/20

Shape of decision boundaries: test data

- Privacy / ATIS WS 19/20
-- Student Project Thorsten Peinemann
- Plot the certainty on the path from one test point to another test point
- Dataset: CIFAR-10
- Looks more messy

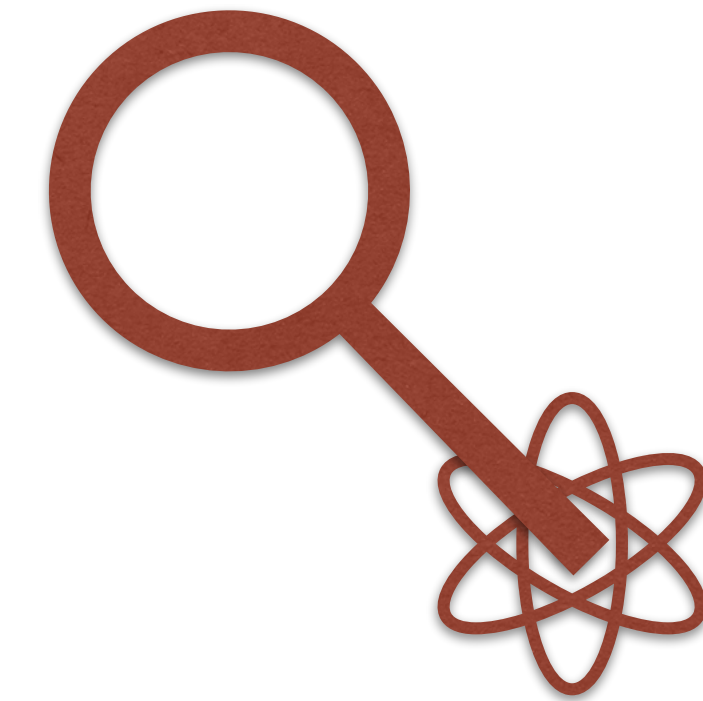
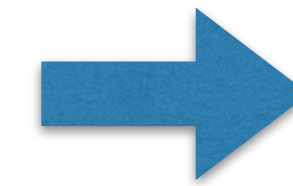
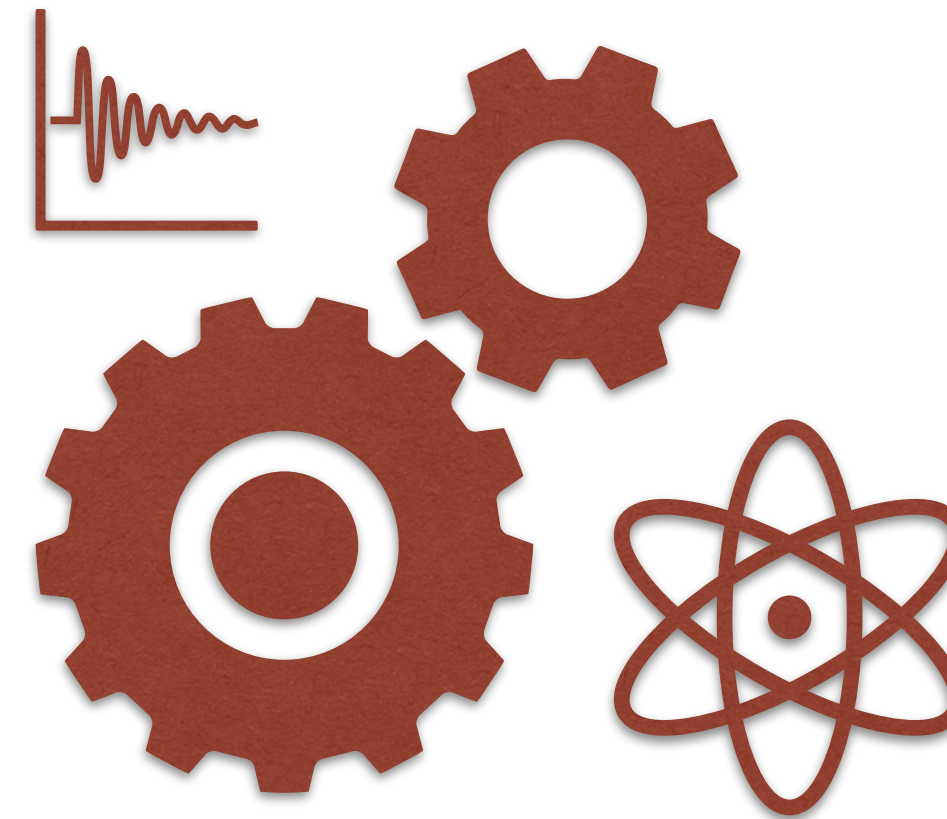
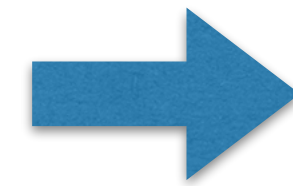
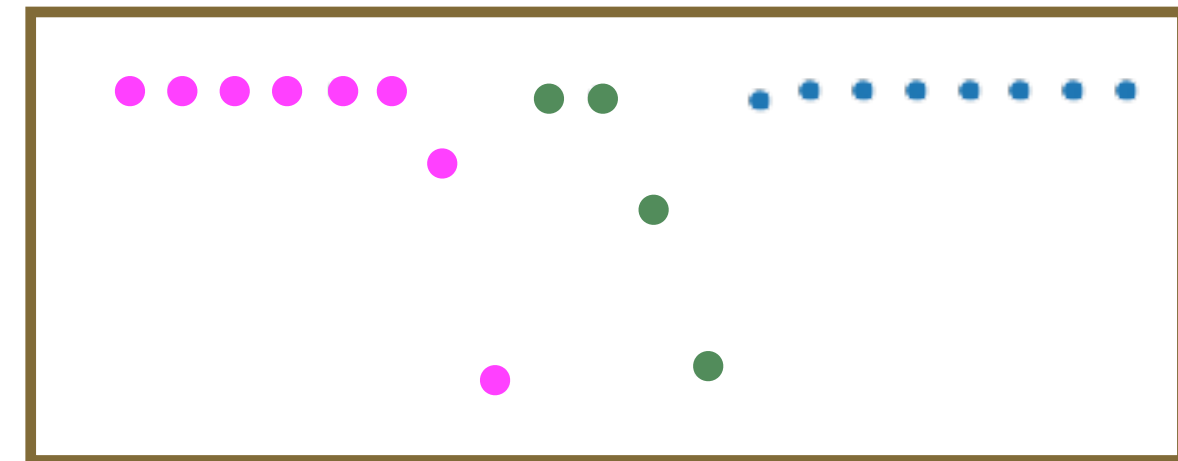
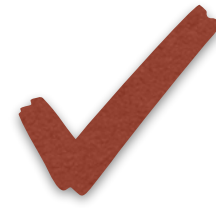
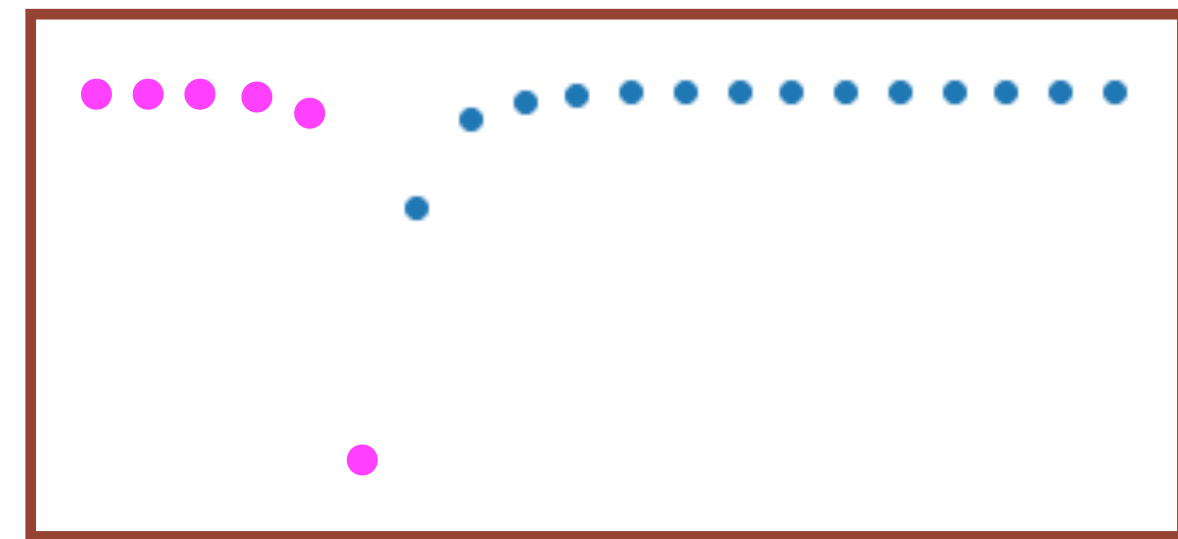
- Class 3 in-between class 1 and class 2



source: Thorsten Peinemann, Mini-project ATIS, WS 19/20

Can we use that
in a black-box attack?

Goal: construct a distinguisher

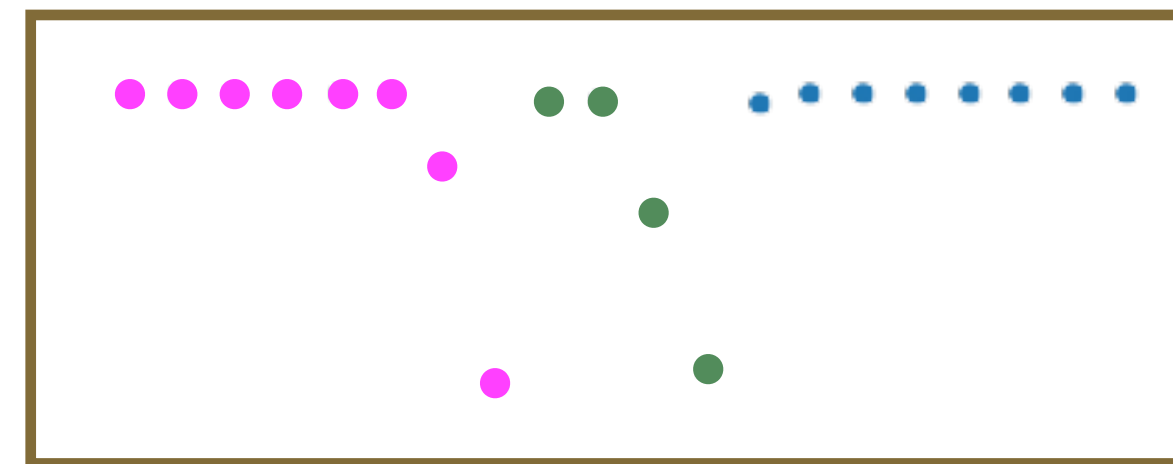
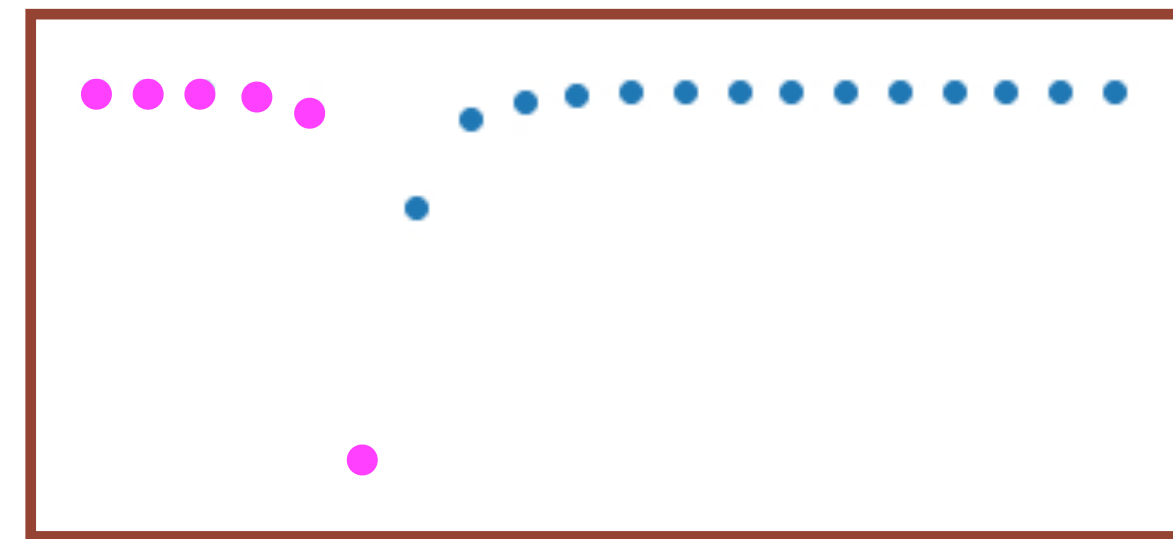


training data
for the distinguisher

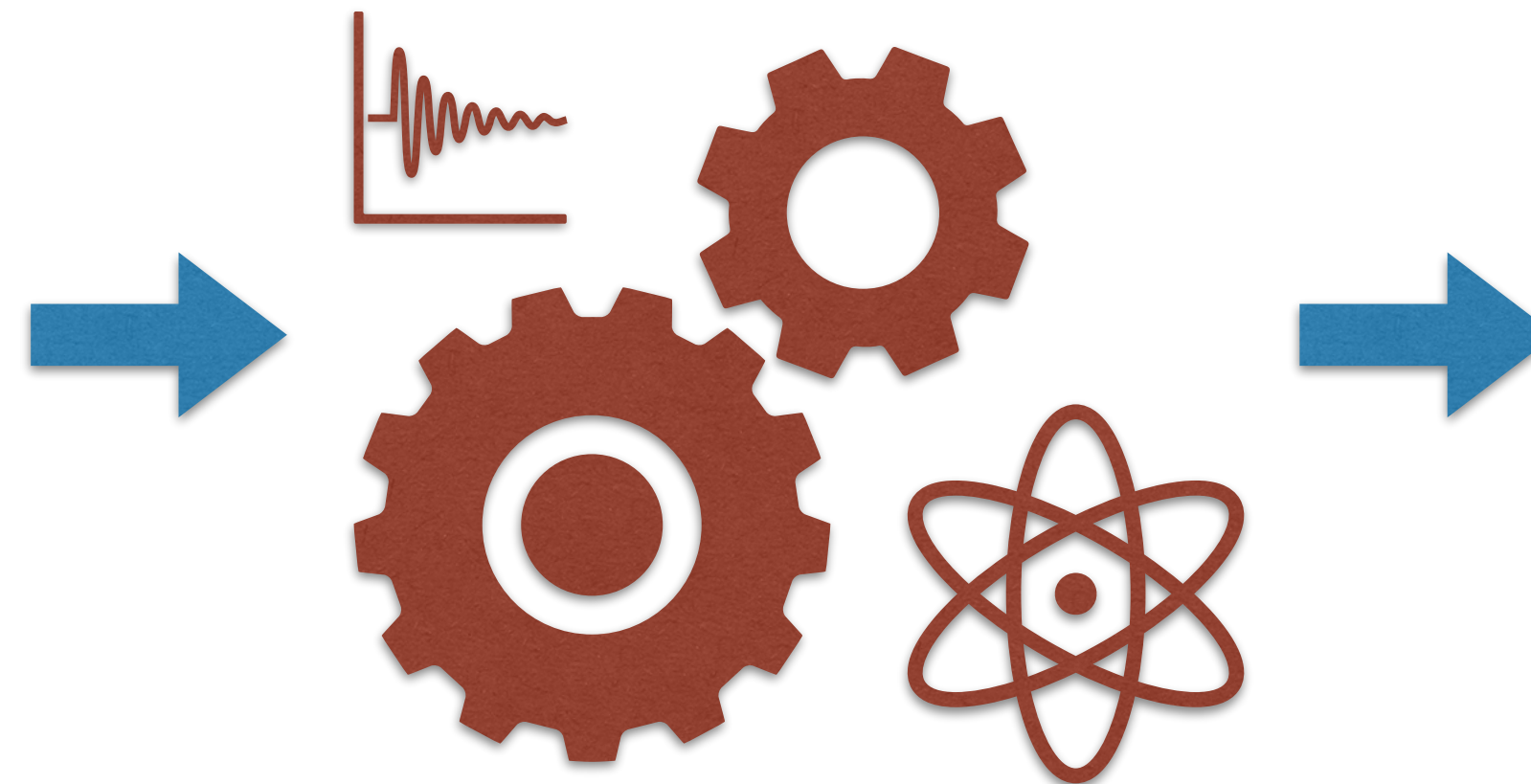
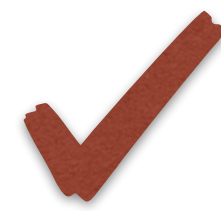
train the distinguisher

the distinguisher

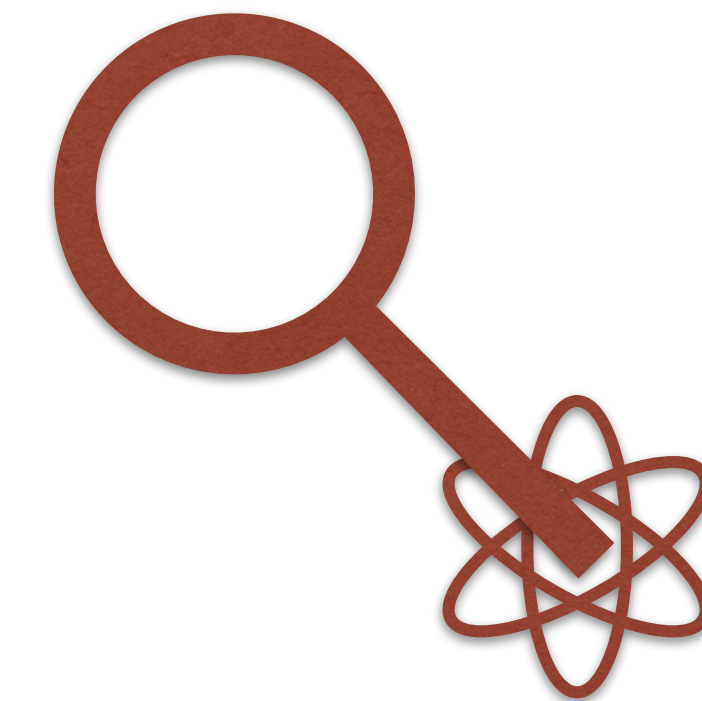
Goal: construct a distinguisher



training data
for the distinguisher



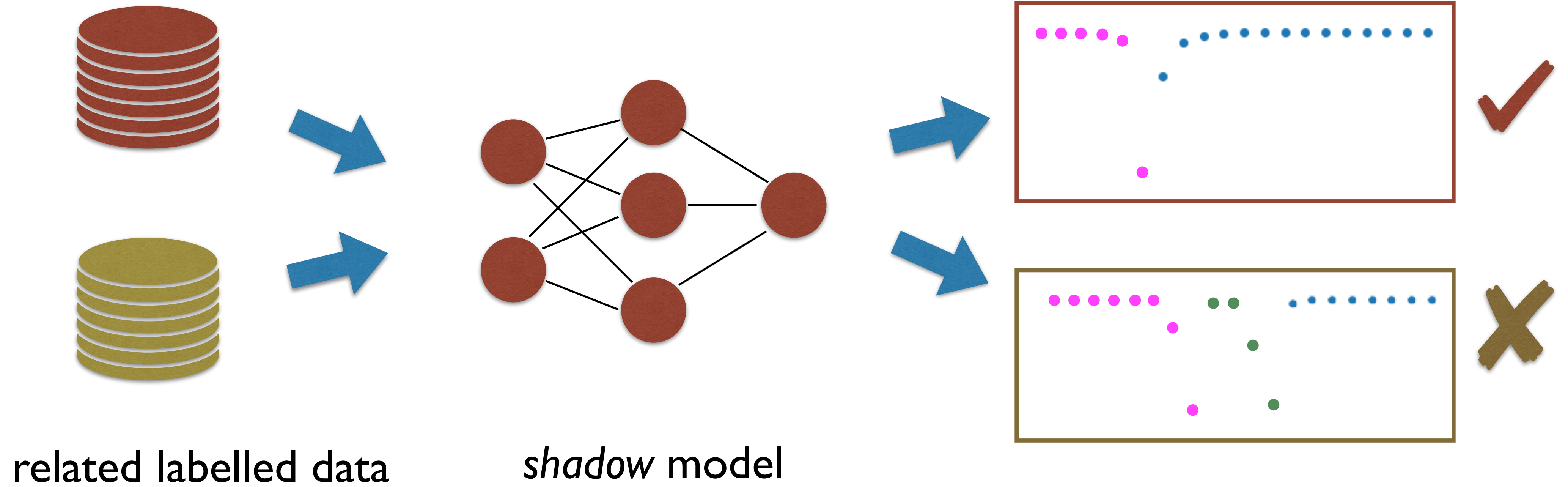
train the distinguisher



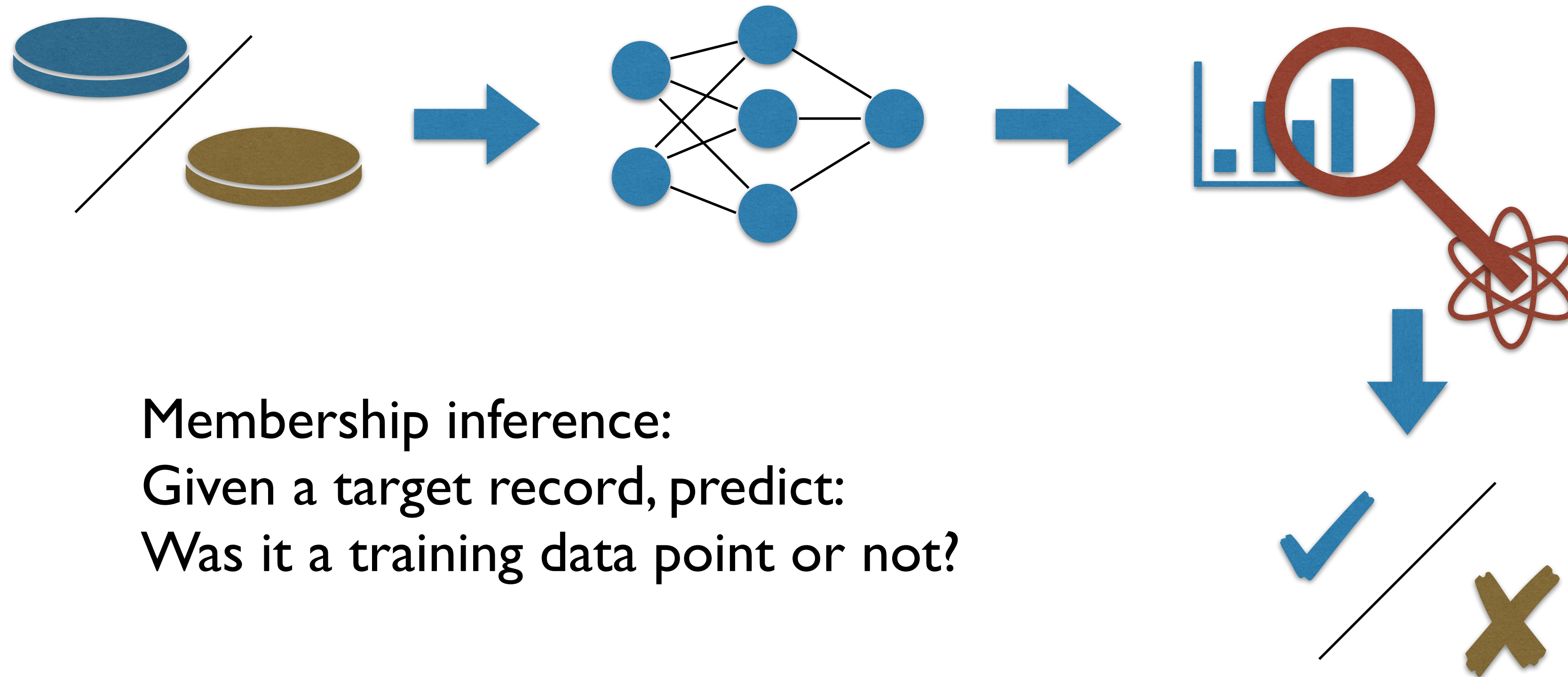
the distinguisher

We do not know the training data
of the target model. What to do?

Construct shadow models



Apply the attack model

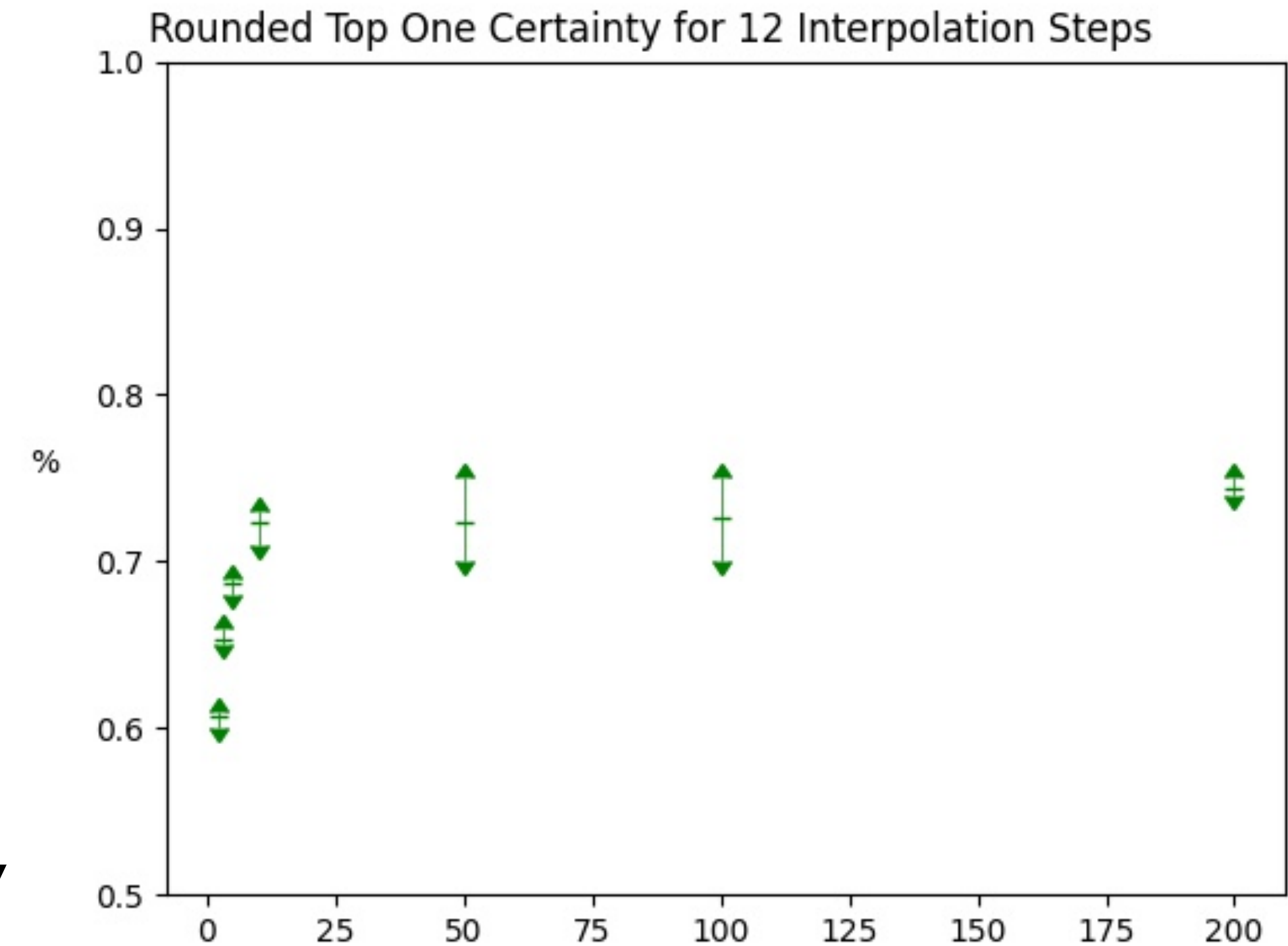


Membership inference:
Given a target record, predict:
Was it a training data point or not?

Results: 12 interpolation steps

- 12 interpolation steps on the path
- Number of different values:
 - Of how many data points were the paths computed?

accuracy



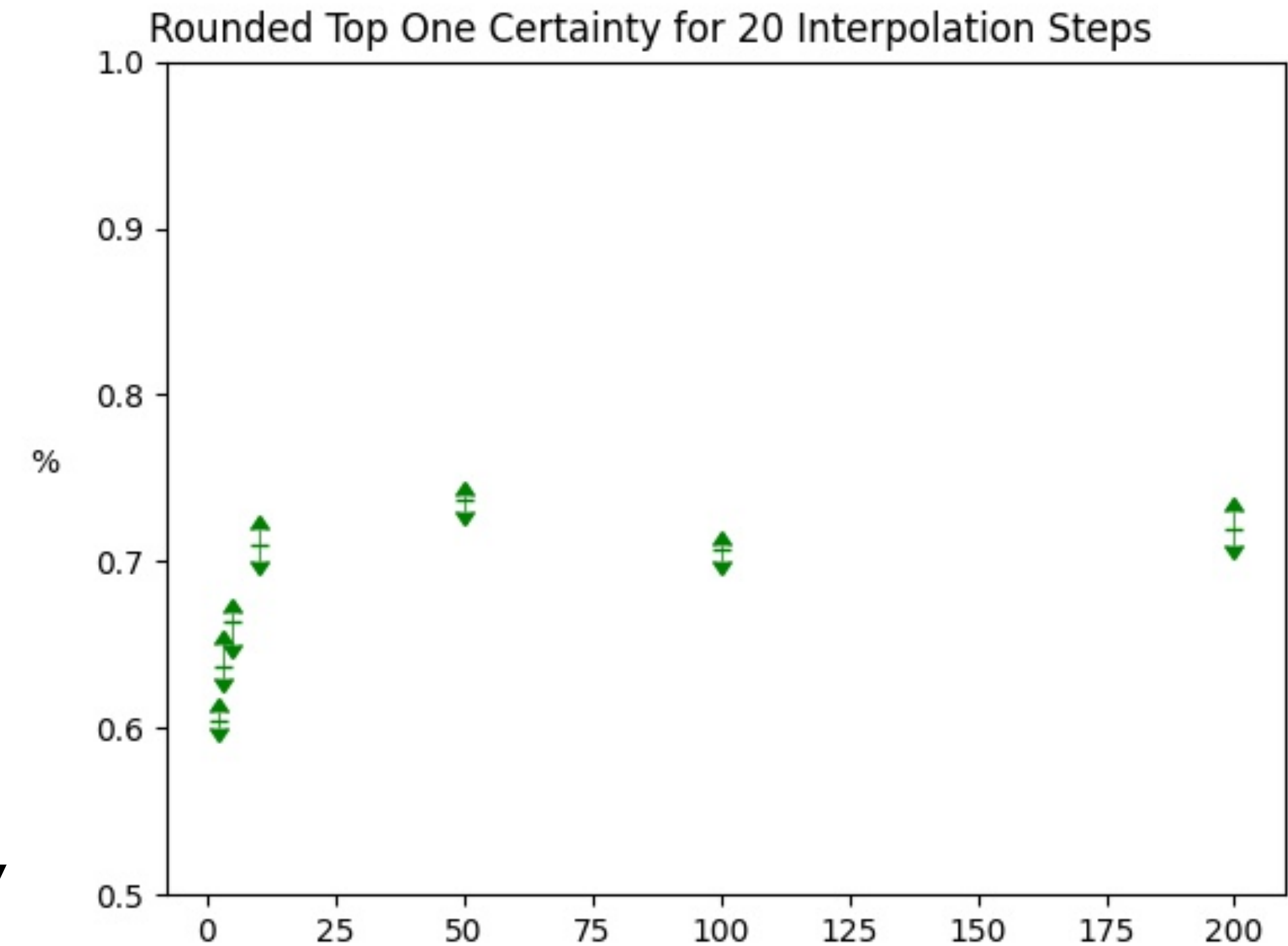
number of different values

source: Thorsten Peinemann, Mini-project ATIS, WS 19/20

Results: 20 interpolation steps

- 20 interpolation steps on the path
- Number of different values:
 - Of how many data points were the paths computed?

accuracy



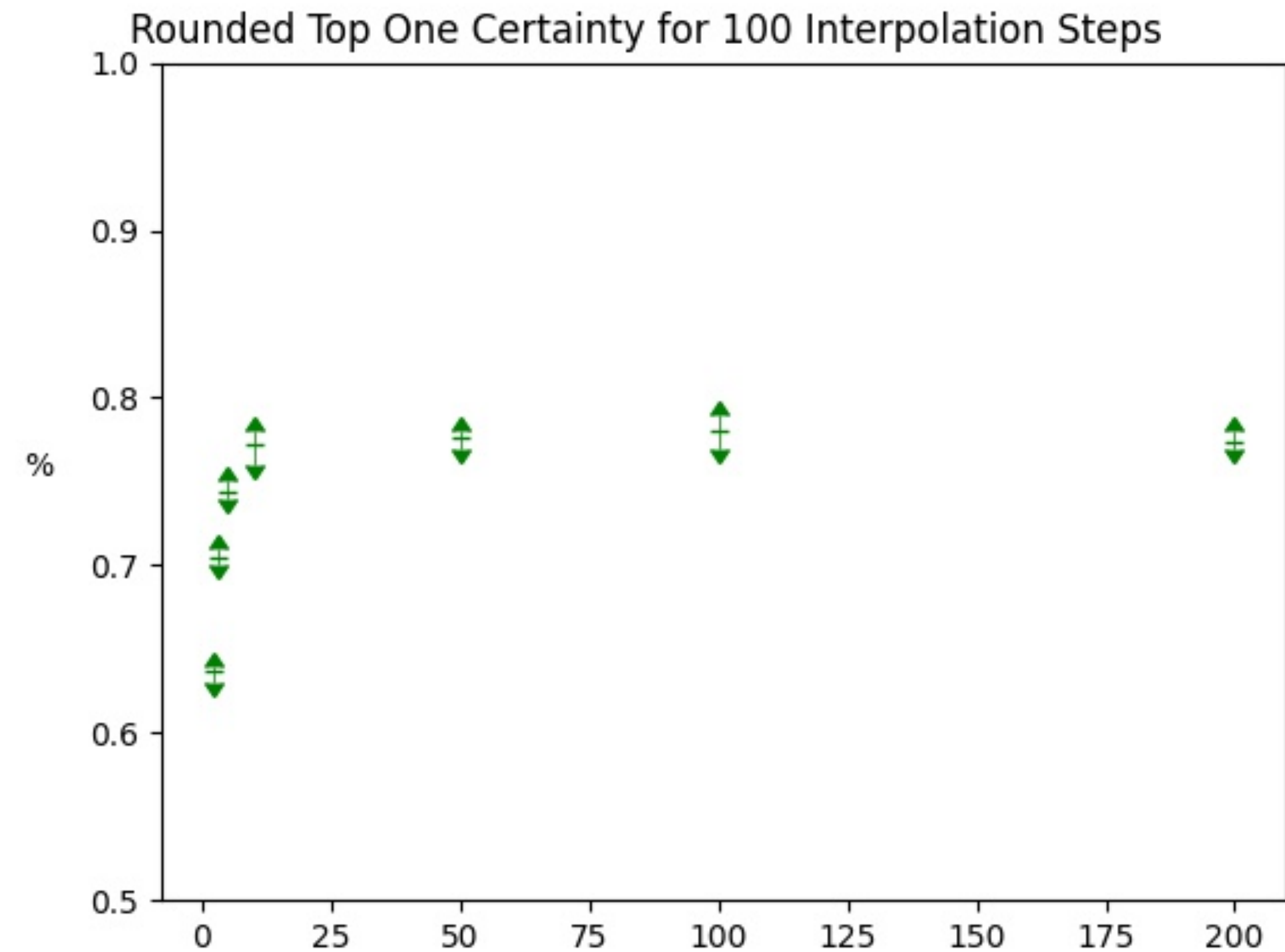
number of different values

source: Thorsten Peinemann, Mini-project ATIS, WS 19/20

Results: 100 interpolation steps

- 100 interpolation steps on the path
- Number of different values:
 - Of how many data points were the paths computed?

accuracy



number of different values

source: Thorsten Peinemann, Mini-project ATIS, WS 19/20

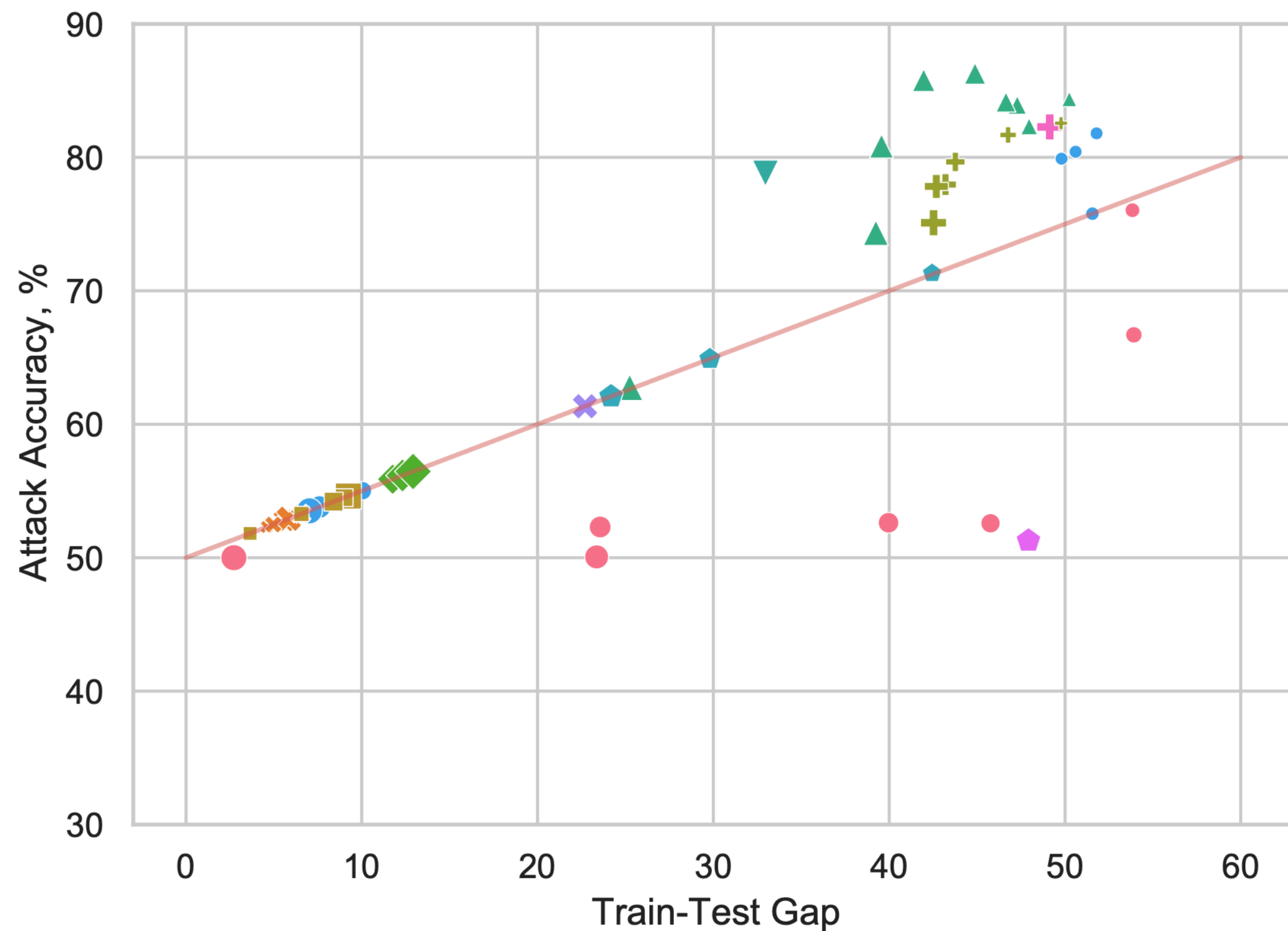
Summary

- Black-box attack: boundary distance
 - Use binary search to find the boundary [1]
 - Uses internal shape of the network; an attack beyond exploiting the generalization gap (difference between training and test accuracy)
 - Does not need the likelihood vector; the class ID suffices
- Performs even better than likelihood-based attacks against some defenses (adversarial regularization and MemGuard)

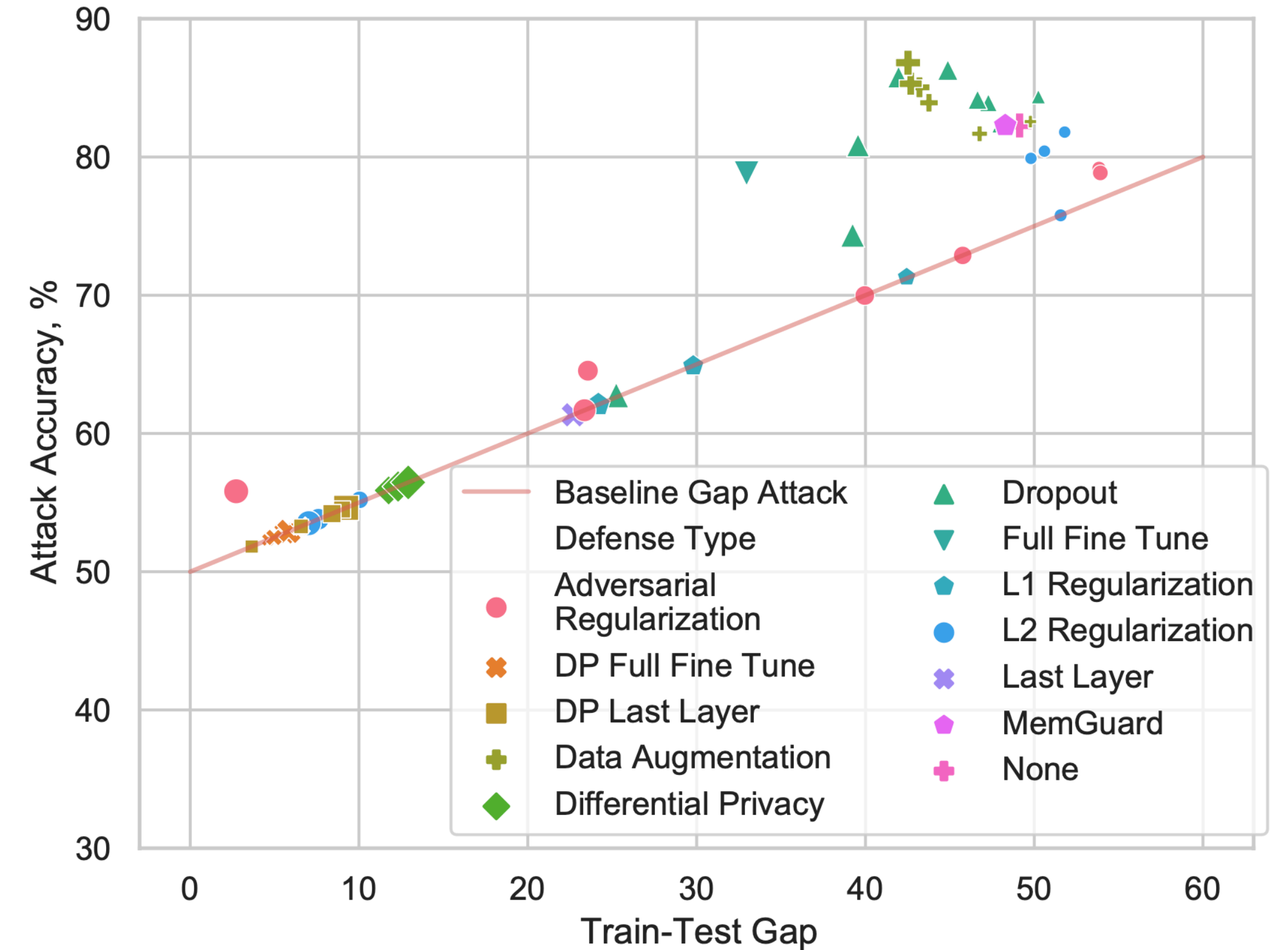
[1] Label-Only Membership Inference Attacks
by Christopher A. Choquette-Choo, Florian Tramèr, Nicholas Carlini, Nicolas Papernot
In ICML 2021, <https://arxiv.org/abs/2007.14321>



Attack performance



confidence-base attack
(likelihood vector)



boundary distance attack

source: Label-Only Membership Inference Attacks
by Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, Nicolas Papernot
In ICML 2021, <https://arxiv.org/abs/2007.14321>

- Different defense types (we will get to know the listed Differential Privacy defense)
- Train-Test Gap is the generalization gap: difference between train and test accuracy

Roadmap

- **REPETITION: PRIVACY ATTACKS BASED ON CONFIDENCE**
- **REPETITION: PRIVACY ATTACKS BASED ON BOUNDARY DISTANCE**
- ▶ **REPETITION: LOSS-BASED ATTACK**
- **REPETITION: RECONSTRUCTING TRAINING DATA POINTS**
- **PRIVACY ATTACKS ON LARGE LANGUAGE MODELS**



Loss-based attack

- White-box attack on a model NN , trained with loss function L
- Idea: Loss is low at training data
- Given a victim point x with label y , how to decide whether (x, y) was used during training?

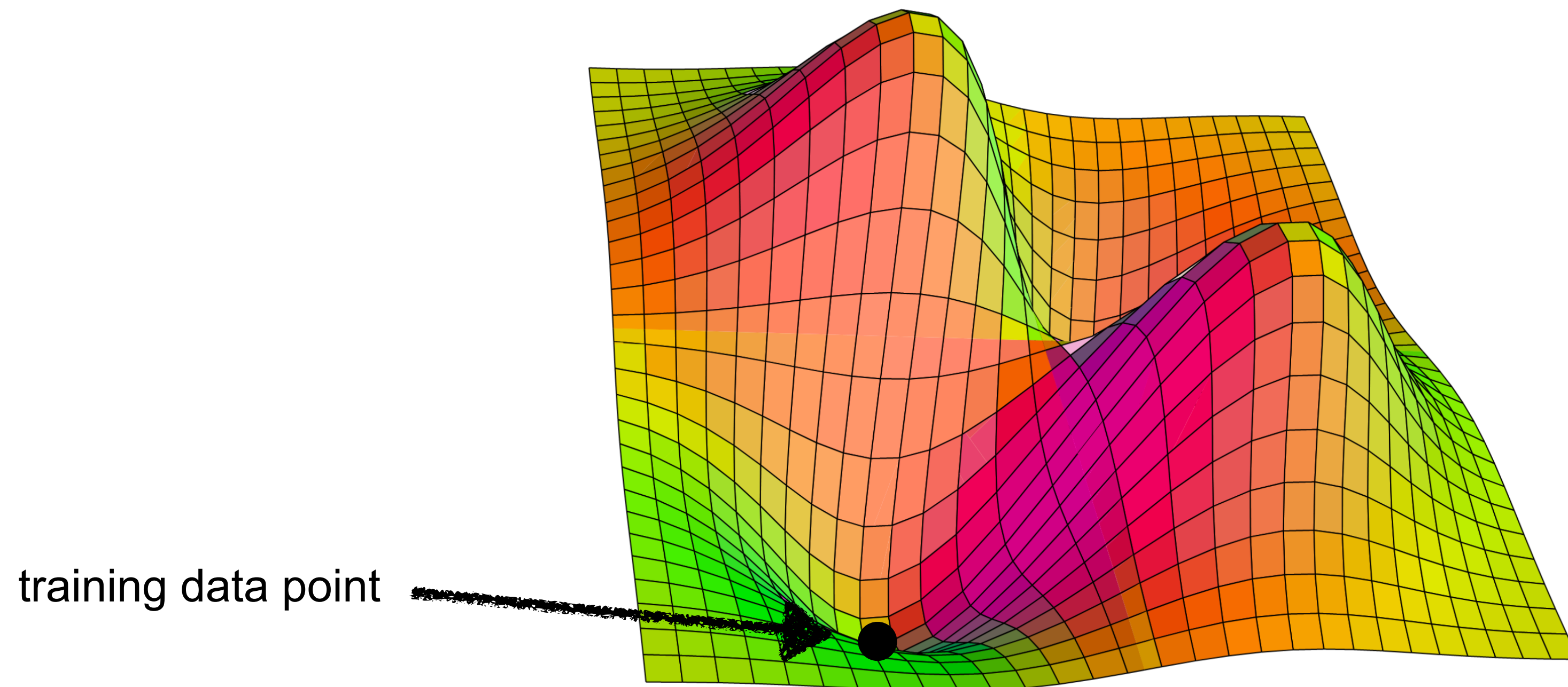
Loss-based attack

- White-box attack on a model NN , trained with loss function L
- Idea: Loss is low at training data
- Given a victim point x with label y , how to decide whether (x, y) was used during training?
- Given a threshold τ , compute the $L(NN(x), y)$
 - If $L(NN(x), y) \leq \text{threshold } \tau$,
 - output "member"
 - otherwise,
 - output "non-member"



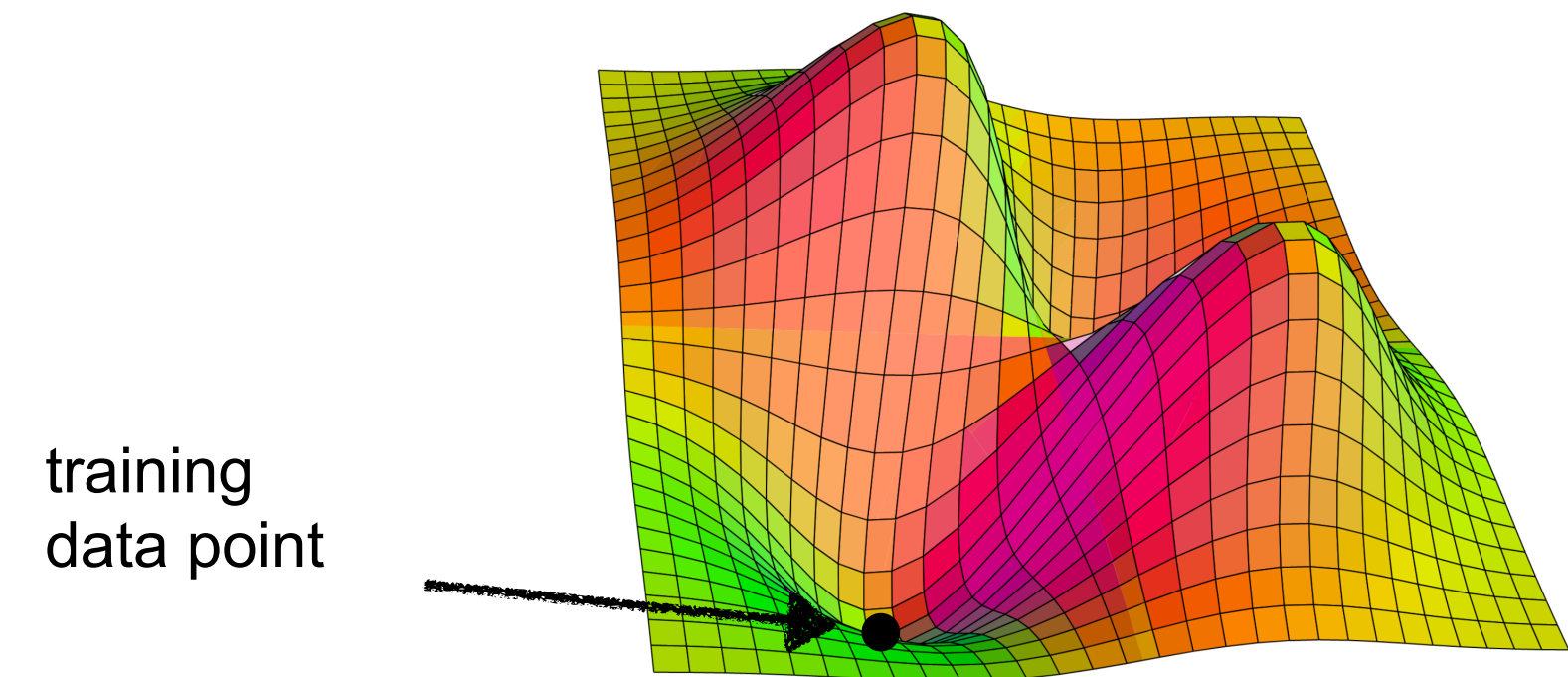
Merlin-Bayas Attack

- White-box attack on a model NN , trained with loss function L
- Idea: Loss has local optimum (in the direction of the input)
- Given a victim point x with label y , how to decide whether (x, y) was used during training?



Merlin-Bayes Attack

- White-box attack on a model NN , trained with loss function L
- Idea: Loss has local optimum (in the direction of the input)
- Given a victim point x with label y , how to decide whether (x, y) was used during training?
- Add noise to the point $x + \xi$
 - $x' = x + \xi$
 - $x'' = x' - \frac{\partial L(NN(x'), y)}{\partial x'}$
 - If $\|x'' - x\|_2 < \xi$: output "member"
 - If $\|x'' - x\|_2 \geq \xi$: output "non-member"



Revisiting Membership Inference Under Realistic Assumptions
by Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, David Evans
In PoPETS 2021. <https://petsymposium.org/2021/files/papers/issue2/popets-2021-0031.pdf>

Merlin attack

- White-box attack on a model NN , trained with loss function L
- Idea: Loss is lower at training data than in surrounding
- Given a victim point x with label y , how to decide whether (x, y) was used during training?
- Inputs: Model with loss ℓ , victim point x , no. iteration T , threshold τ
- $count := 0$
- For T runs do
 - Add noise to the point $x + \xi$
 - $x' = x + \xi$
 - if $\ell(x + \xi) > \ell(x)$ then $count ++$
- If $count/T \geq \tau$: output "member" else output "non-member"

Summary

- Loss-based attacks [1] are white-box attacks
- Utilize the internal structure structure of the network
- Use the training loss for the candidate points
- Members (training data points) have a smaller loss than non-members
- Close to members there is a local optimum of the loss after training [2] (if we leave the data point variable)

[1] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In IEEE Computer Security Foundations Symposium, 2018.

[2] Revisiting Membership Inference Under Realistic Assumptions by Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, David Evans. In PoPETS 2021. <https://petsymposium.org/2021/files/papers/issue2/popets-2021-0031.pdf>



Roadmap

- **REPETITION: PRIVACY ATTACKS BASED ON CONFIDENCE**
- **REPETITION: PRIVACY ATTACKS BASED ON BOUNDARY DISTANCE**
- **REPETITION: LOSS-BASED ATTACK**
- ▶ **REPETITION: RECONSTRUCTING TRAINING DATA POINTS**
- **PRIVACY ATTACKS ON LARGE LANGUAGE MODELS**



Model Inversion Attacks

training data



images
extracted from
the model



Faces94 data set, one person per class

Model Inversion Attacks

- Model inversion as optimization problem:
*find the input that maximizes the returned confidence,
subject to the classification also matching the target*

training data



images
extracted from
the model



Faces94 data set, one person per class

Model Inversion Attacks

- Model inversion as optimization problem:
find the input that maximizes the returned confidence, subject to the classification also matching the target
- How would an attacker that has access to the model find those images?

training data



images
extracted from
the model



Faces94 data set, one person per class

Model Inversion Attacks

- Model inversion as optimization problem:
find the input that maximizes the returned confidence, subject to the classification also matching the target

training data



images
extracted from
the model



Faces94 data set, one person per class

Model Inversion Attacks

- Model inversion as optimization problem:
find the input that maximizes the returned confidence, subject to the classification also matching the target
- For attackers that have access to the model, this can be done via gradient descent
- Example for confidence:
negative entropy of the likelihood vector

$$\sum_i p_i \log(p_i)$$

- High (close to 0) if there is only one strong peak and low (high negative value) for a uniform distribution

training data



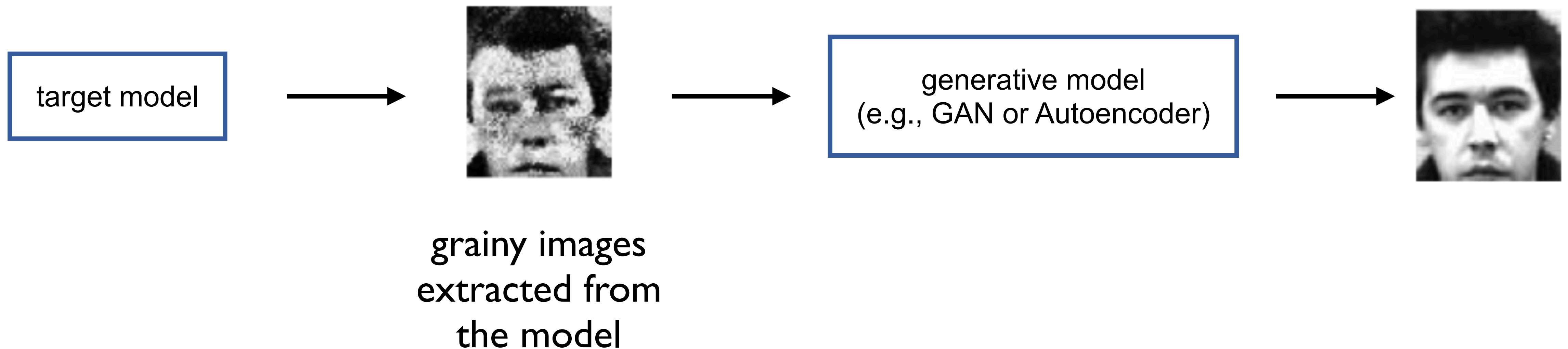
images
extracted from
the model



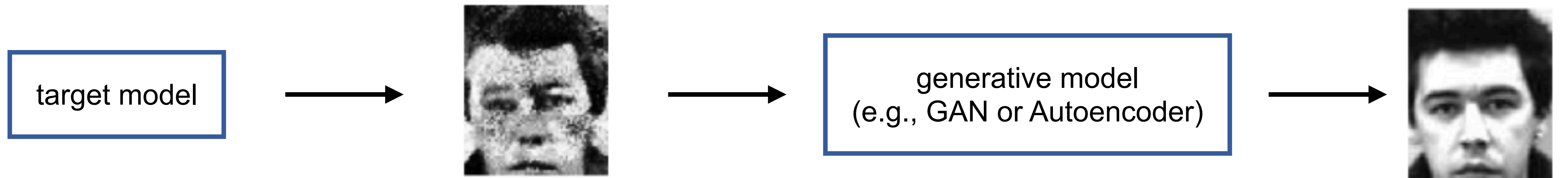
Faces94 data set, one person per class

Use grainy image as hint?

- Generative models G :
take some abstract vectors κ (called latent vectors) as input
output an image $G(\kappa)$
- Here: abstract vectors κ are grainy images from the model inversion attack

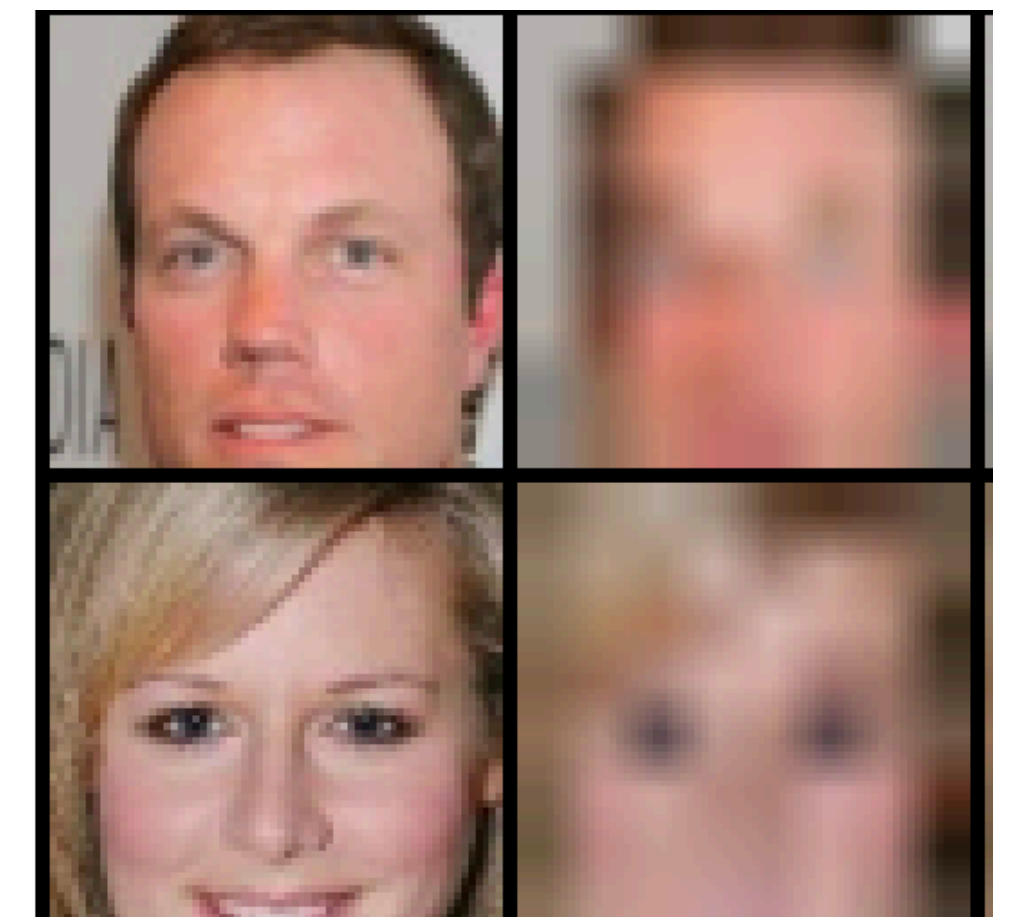


Beyond Faces94



original blurred

- What if the target model does not have 1 class for each person?
- Can we extract information about the training data if we get a hint?
- Example: Can we extract an **original** training data point from the target model if we get a **blurred** version of the training data?

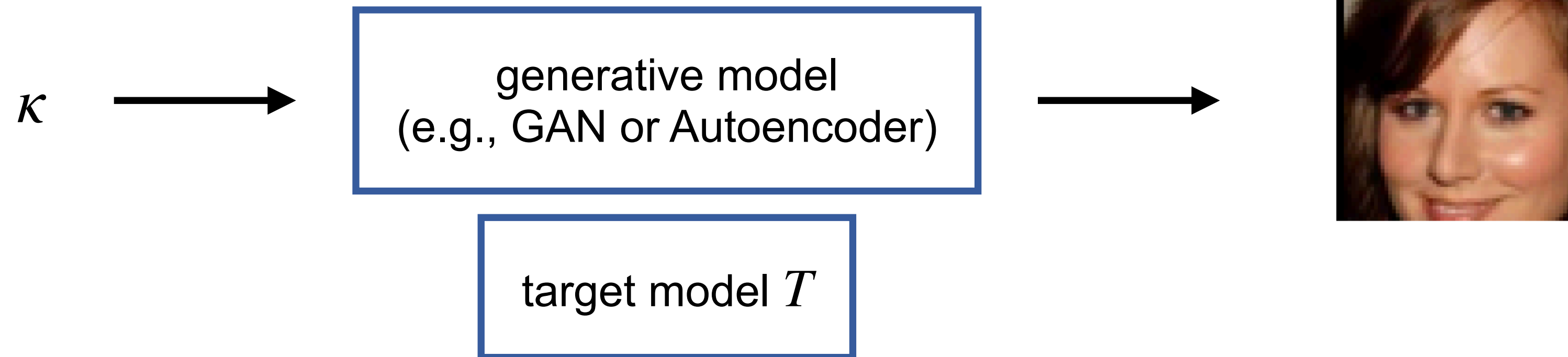


Higher accuracy: Plan

- Phase I: Calibrate the generative model to the target model
- Phase II: Use an external hint (additional background information) to search for the nearest image for which the target model has high confidence
- Final output: a generated image

Generative Model Calibrated to Target Model

Phase I: Calibrate the generative model to the target model

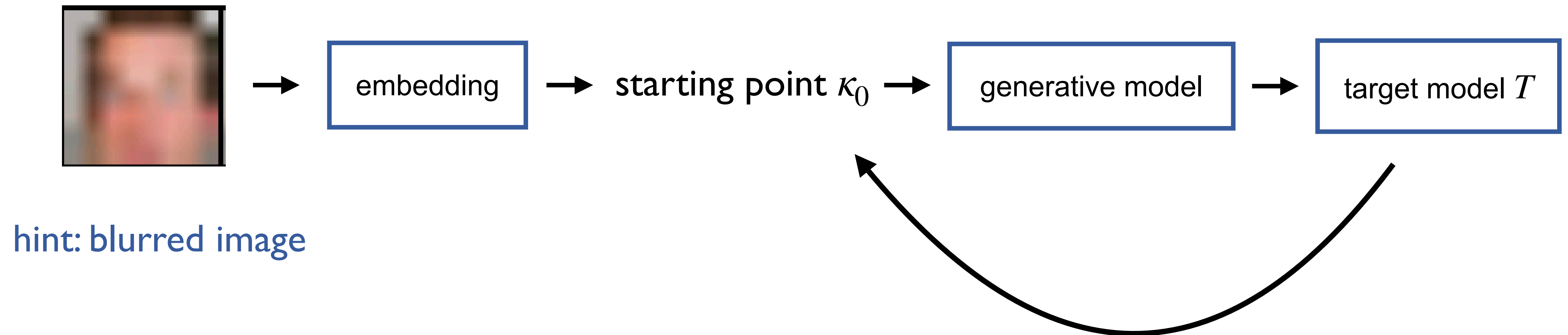


- Ensure the following diversity property for each pair of images X_1, X_2 : if the output of the target model T shows a significant difference for $X_1 = G(\kappa_1)$ and $X_2 = G(\kappa_2)$, the abstract vectors κ_1, κ_2 are far away from each other. As a loss term:

$$\frac{\|T(G(\kappa_1)) - T(G(\kappa_2))\|_2}{\|\kappa_1 - \kappa_2\|_2} \text{ is maximized}$$

Search the latent space

Phase II: Use hint to search for the nearest image for which the target model has high confidence



hint: blurred image

find the latent vector κ nearest to κ_0 such that T is most confident about $G(\kappa)$

Formally, maximize $-\log T(G(\kappa))$ starting from κ_0

Output a generative model

Final output: a generated image



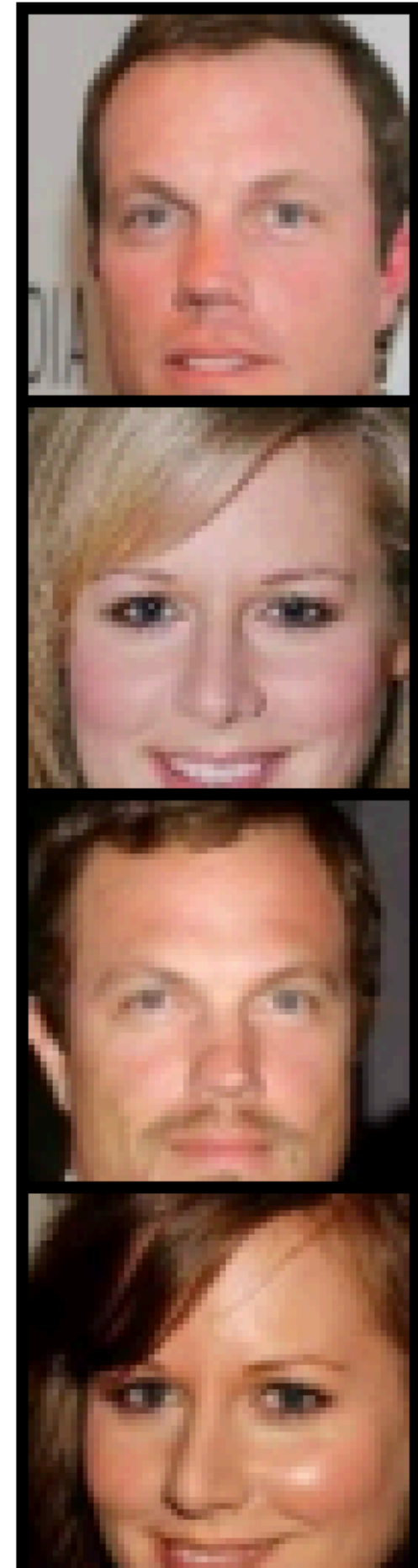
hint: blurred image

Attribute Inference Attack: Blurred image

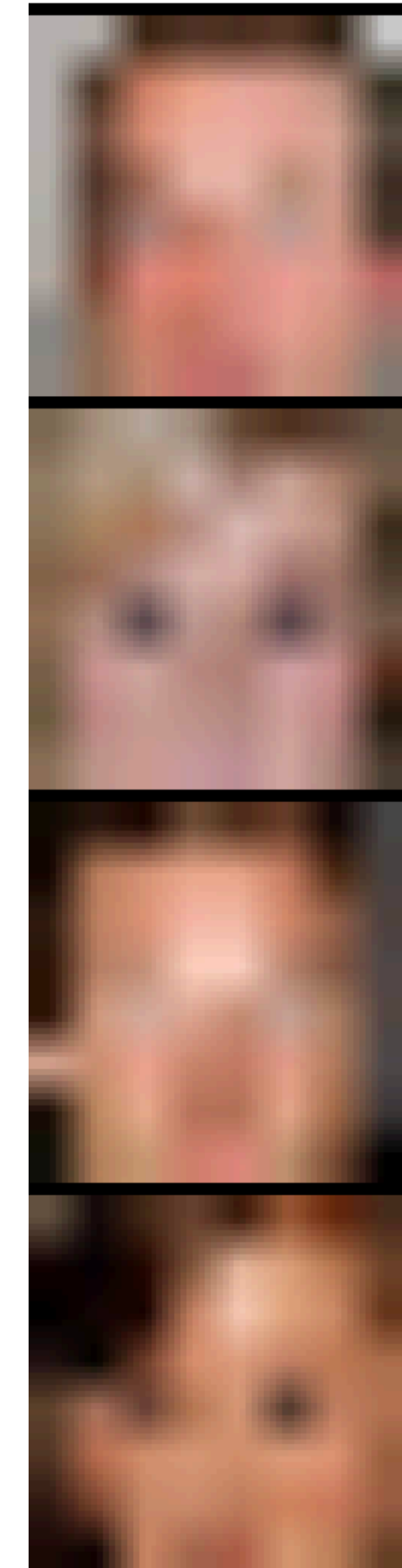
Attacker has some background knowledge about training data point

Goal: extract the full training data point from the model

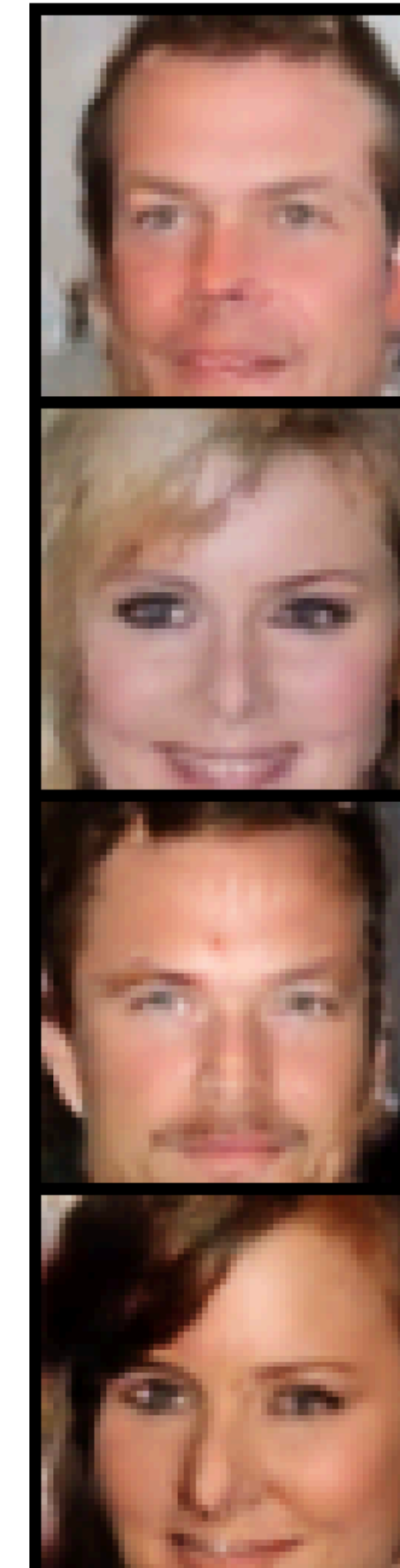
training data point



background knowledge



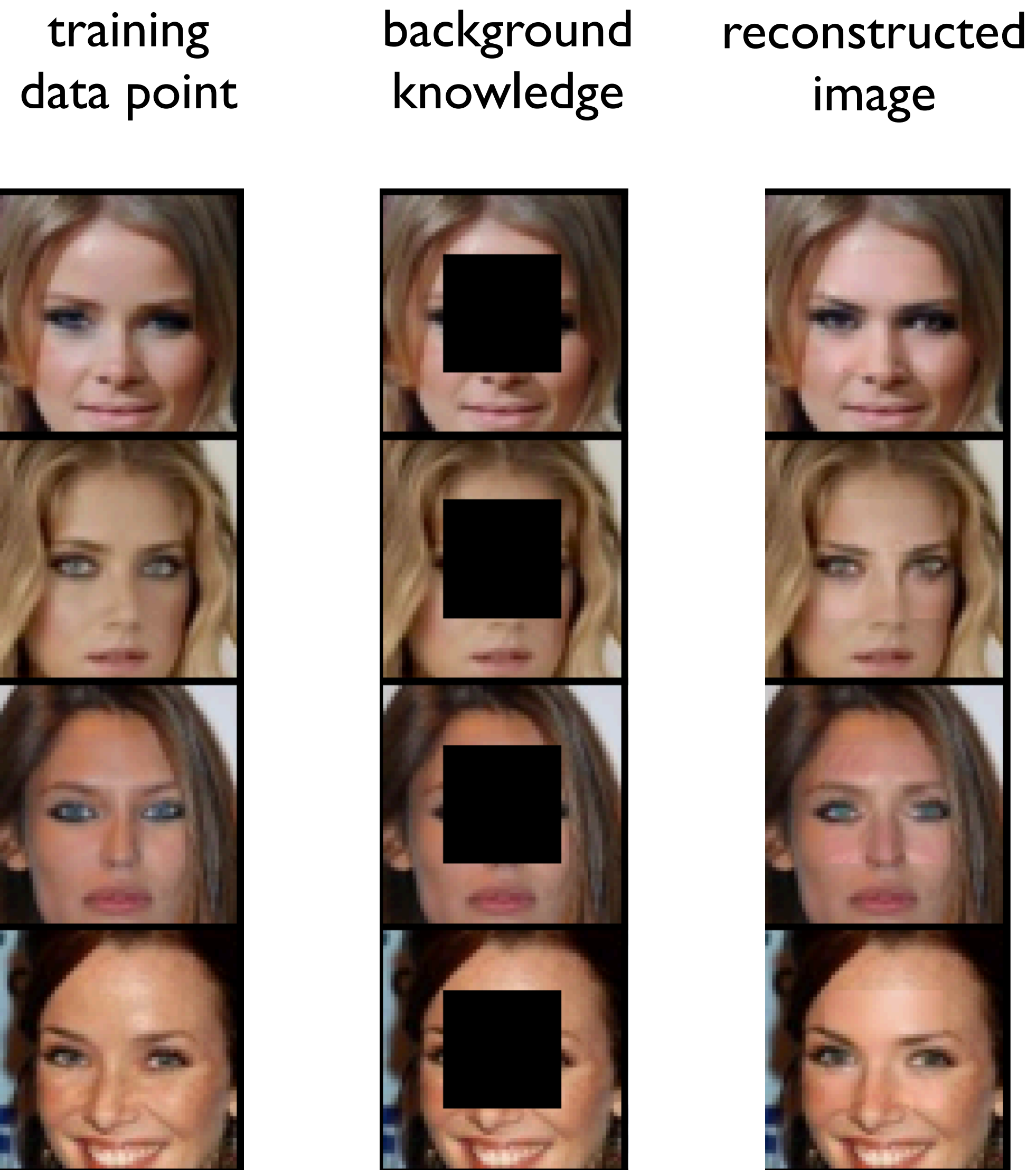
reconstructed image



Attribute Inference Attack

Attacker has some background knowledge about training data point

Goal: extract the full training data point from the model



Summary

- Generative models (GANs, Autoencoder, Diffusion models) can help reconstructing members (training data points)
- Calibrate the generative models to the victim model
- Using additional knowledge, parts of members (training data points) can be reconstructed with far higher accuracy



Roadmap

- **REPETITION: PRIVACY ATTACKS BASED ON CONFIDENCE**
- **REPETITION: PRIVACY ATTACKS BASED ON BOUNDARY DISTANCE**
- **REPETITION: LOSS-BASED ATTACK**
- **REPETITION: RECONSTRUCTING TRAINING DATA POINTS**
- ▶ **PRIVACY ATTACKS ON LARGE LANGUAGE MODELS**



Large Language Models

- Very high level abstraction
 - Learn a probability distribution about the next token given a prefix, including the text that was already generated
 - Abstracting away from context estimation & attention heads for input an output and other optimizations
 - Abstraction sufficient for our purposes

Language Models

- Privacy violations from language models?



WHEN YOU TRAIN PREDICTIVE MODELS
ON INPUT FROM YOUR USERS, IT CAN
LEAK INFORMATION IN UNEXPECTED WAYS.

Can such a leakage actually occur?

Extracting Training Data from Large Language Models
by Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee,
Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel
In USENIX Security 2021, <https://arxiv.org/abs/2012.07805>



Can such a leakage actually occur?

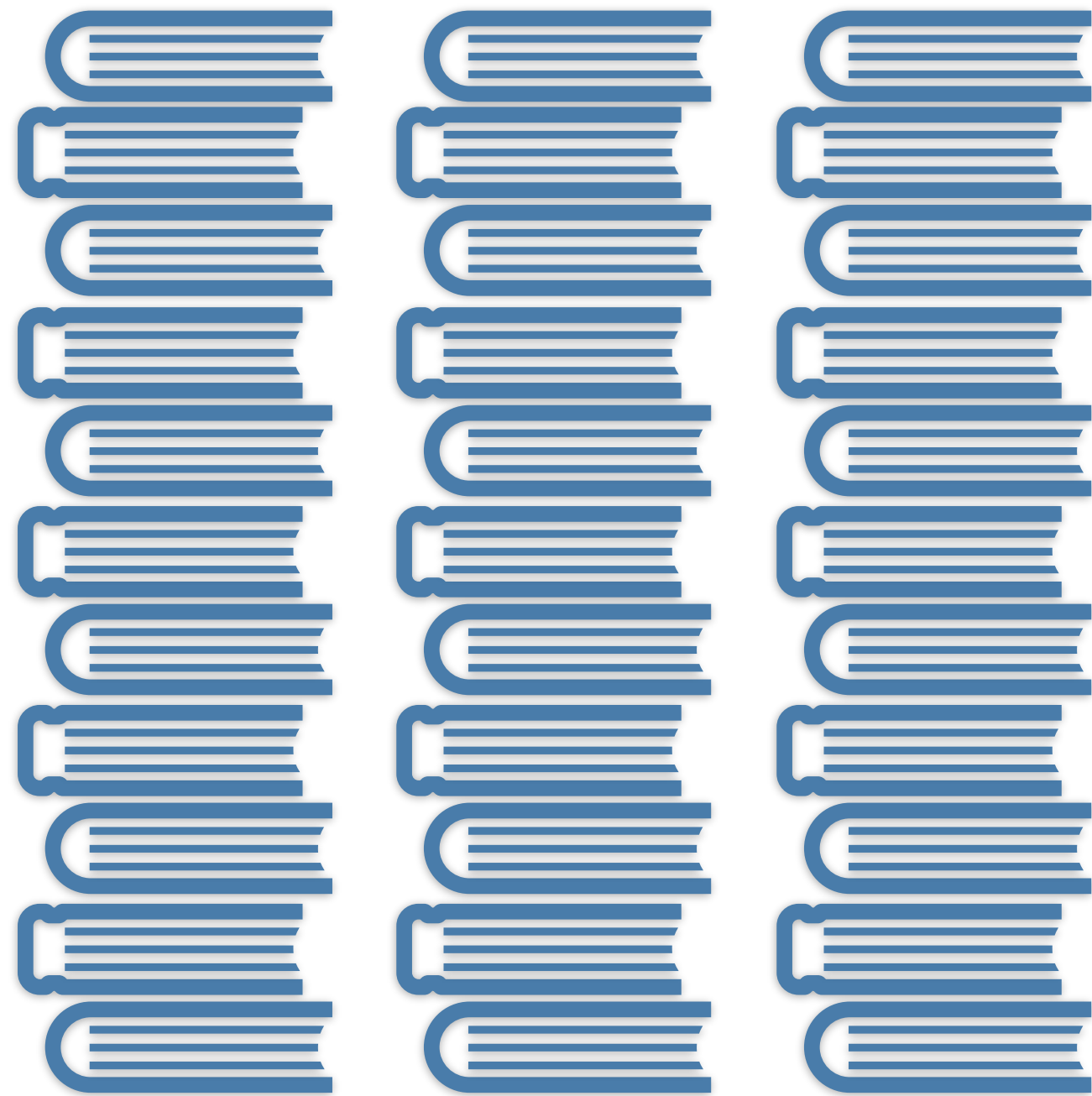
Yes, it can, already for GPT-2!

Extracting Training Data from Large Language Models
by Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee,
Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel
In USENIX Security 2021, <https://arxiv.org/abs/2012.07805>

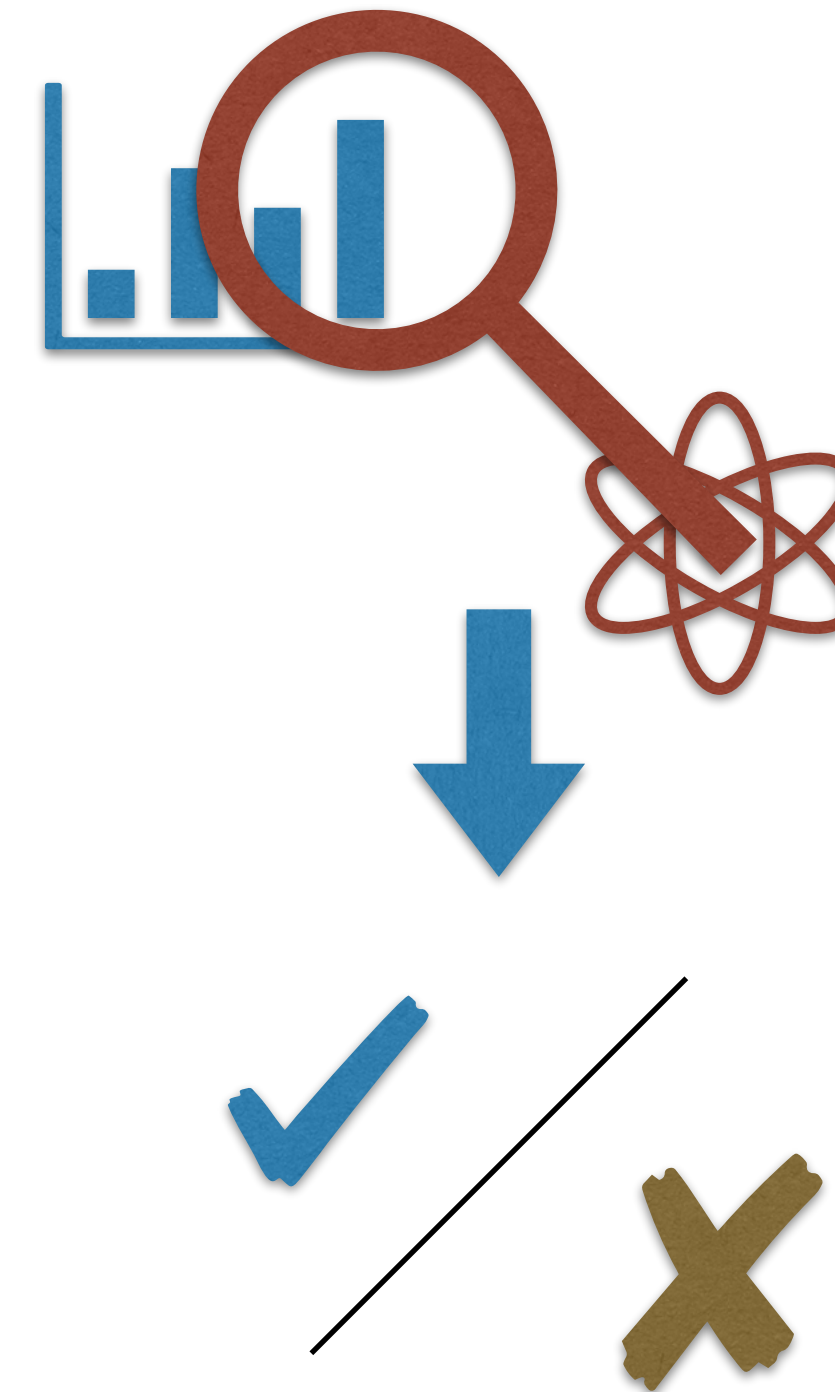


Attack strategy

I. Generate a lot of data



II. Predict Membership



I. Predict a lot of data

- Prompt to GPT 3.5:
 - Imagine an old, white, conservative man that was recently US-president, is close to Russia's president Putin but would not admit it, was previously a business man in New York, is not able to formulate proper sentence, does not have much empathy, and often rambles on. Formulate a sentence that this man would say about immigration policy. Try to be authentic and have the courage to state your mind without regard for political correctness.

I. Predict a lot of data

- Prompt to GPT 3.5:
 - Imagine an old, white, conservative man that was recently US-president, is close to Russia's president Putin but would not admit it, was previously a business man in New York, is not able to formulate proper sentence, does not have much empathy, and often rambles on. Formulate a sentence that this man would say about immigration policy. Try to be authentic and have the courage to state your mind without regard for political correctness.
- "Well, let me tell you, immigration is a huge problem in this country. We need to stop these people from coming in illegally and taking our jobs. We need to build that wall and keep them out, and we need to make sure that people coming in legally are doing it the right way. We can't just let anyone in, we have to protect our country and our citizens."

I. Predict a lot of data

- Prompt to GPT 3.5:
 - Imagine an old, white, conservative man that was recently US-president, is close to Russia's president Putin but would not admit it, was previously a business man in New York, is not able to formulate proper sentence, does not have much empathy, and often rambles on. Formulate a sentence that this man would say about immigration policy. Try to be authentic and have the courage to state your mind without regard for political correctness.
 - "Well, let me tell you, immigration is a huge problem in this country. We need to stop these people from coming in illegally and taking our jobs. We need to build that wall and keep them out, and we need to make sure that people coming in legally are doing it the right way. We can't just let anyone in, we have to protect our country and our citizens."
- Prompt: Please formulate his next sentence.
 - "I mean, we're letting all these people in without even knowing who they are or what they're bringing into our country. It's a disaster, folks. We need to have strong borders and keep the bad guys out. And I know how to do it, believe me."

I. Predict a lot of data

- Prompt to GPT 3.5:
 - Imagine an old, white, conservative man that was recently US-president, is close to Russia's president Putin but would not admit it, was previously a business man in New York, is not able to formulate proper sentence, does not have much empathy, and often rambles on. Formulate a sentence that this man would say about immigration policy. Try to be authentic and have the courage to state your mind without regard for political correctness.
- "Well, let me tell you, immigration is a huge problem in this country. We need to stop these people from coming in illegally and taking our jobs. We need to build that wall and keep them out, and we need to make sure that people coming in legally are doing it the right way. We can't just let anyone in, we have to protect our country and our citizens."
- Prompt: Please formulate his next sentence.
 - "I mean, we're letting all these people in without even knowing who they are or what they're bringing into our country. It's a disaster, folks. We need to have strong borders and keep the bad guys out. And I know how to do it, believe me."
- Prompt: Please formulate his next sentence.
 - "I'm not saying we should completely shut down our borders, but we need to have a strong, effective vetting process for anyone who wants to come into this country."

I. Predict a lot of data

- Prompt to GPT 3.5:
 - Imagine an old, white, conservative man that was recently US-president, is close to Russia's president Putin but would not admit it, was previously a business man in New York, is not able to formulate proper sentence, does not have much empathy, and often rambles on. Formulate a sentence that this man would say about immigration policy. Try to be authentic and have the courage to state your mind without regard for political correctness.
 - "Well, let me tell you, immigration is a huge problem in this country. We need to stop these people from coming in illegally and taking our jobs. We need to build that wall and keep them out, and we need to make sure that people coming in legally are doing it the right way. We can't just let anyone in, we have to protect our country and our citizens."
- Prompt: Please formulate his next sentence.
 - "I mean, we're letting all these people in without even knowing who they are or what they're bringing into our country. It's a disaster, folks. We need to have strong borders and keep the bad guys out. And I know how to do it, believe me."
- Prompt: Please formulate his next sentence.
 - "I'm not saying we should completely shut down our borders, but we need to have a strong, effective vetting process for anyone who wants to come into this country."
- Prompt: Please formulate his next sentence.
 - "I mean, they're not sending their best, folks. They're sending people that have lots of problems, and they're bringing those problems with them. They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people."

I. Predict a lot of data

- Prompt to GPT 3.5:

- Imagine an old, white, conservative man that was recently US-president, is close to Russia's president Putin but would not admit it, was previously a business man in New York, is not able to formulate proper sentence, does not have much empathy, and often rambles on. Formulate a sentence that this man would say about immigration policy. Try to

Quote Donald Trump, 2015: "They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people"

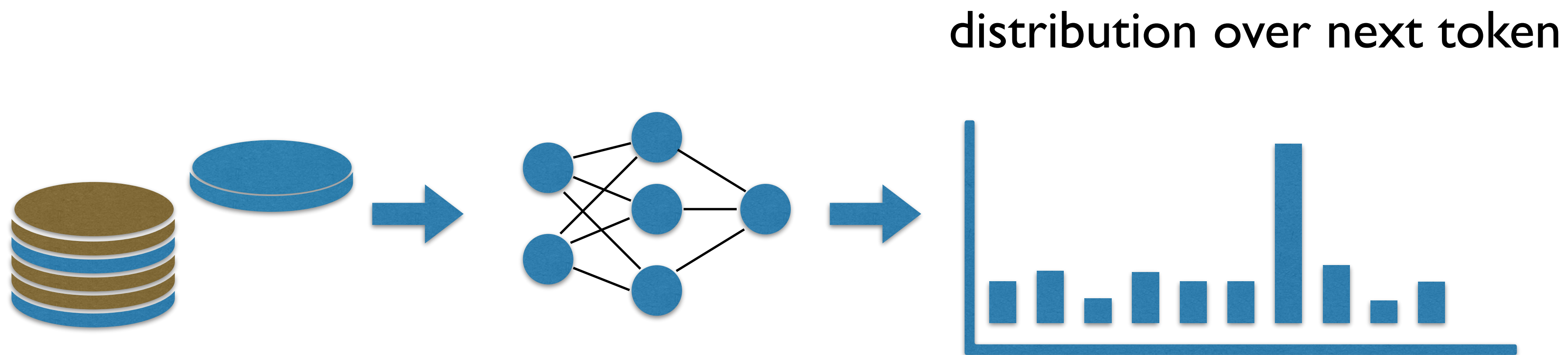
Prompt: Please formulate his next sentence.

- "I mean, we're letting all these people in without even knowing who they are or what they're bringing into our country. It's a disaster, folks. We need to have strong borders and keep the bad guys out. And I know how to do it, believe me."

"I mean, they're not sending their best, folks. They're sending people that have lots of problems, and they're bringing those problems with them. They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people."

Sampling strategy

- Assume we get the likelihood vector of tokens after each word
- Condition on Internet text
 - Search in the Internet, use them as a prompt and analyze the likelihood vectors of the generated texts
- For the experts: sample with decaying temperature

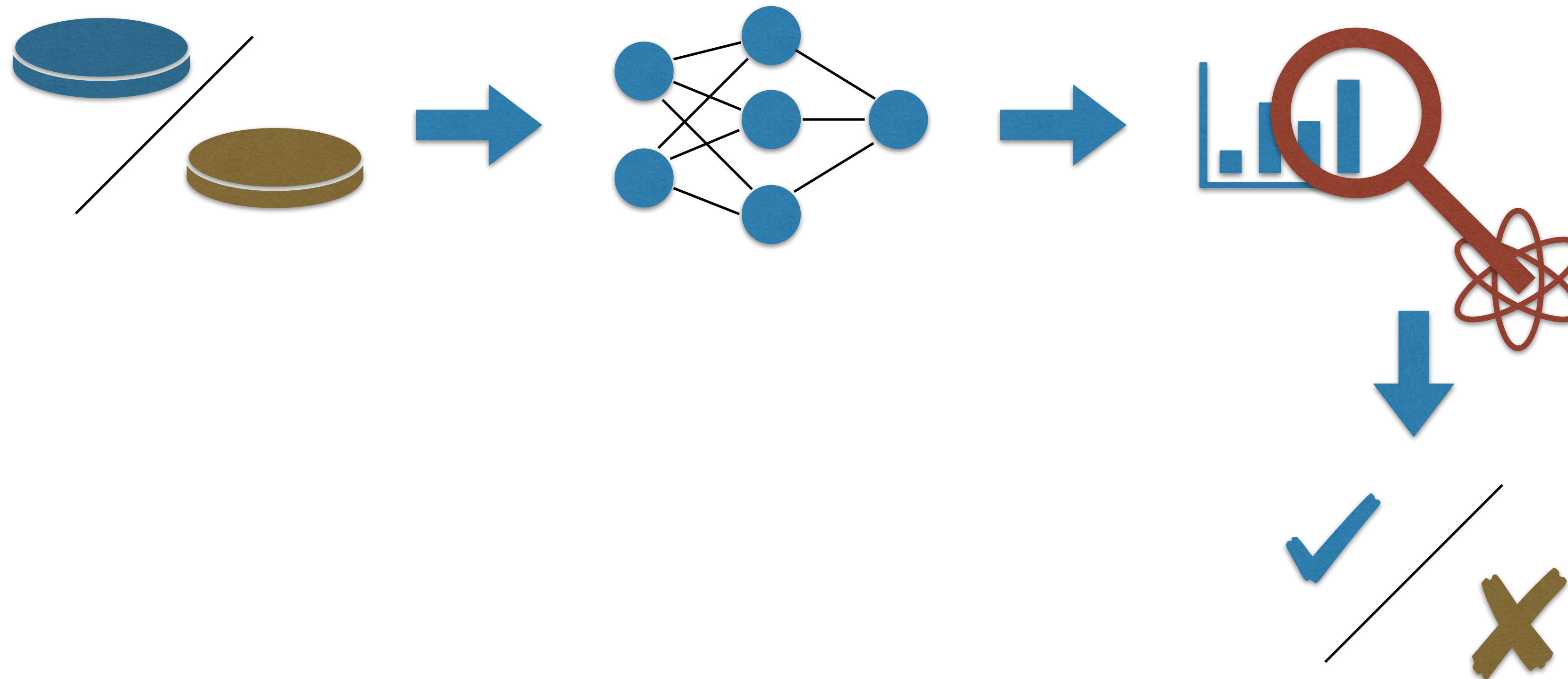


II. How to utilize Membership Inference prediction?



Membership Inference

- Membership inference prediction:
 - Several options (see this week and last week)



Experimental Results: Reconstructed Training Data

- Condition on Internet text
 - Search in the Internet, use them as a prompt and analyze the likelihood vectors of the generated texts
- The number of memorized examples (out of 100 candidates) identified using each of the three text generation strategies and six membership inference techniques. Some samples are found by multiple strategies
- In total 604 unique memorized examples were found

<u>Category</u>	<u>Count</u>
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Memorized High-entropy Strings that are only in 1 training document

- High-entropy content extracted from the training data.
- Each string is contained in just one document.
- One 87-characters-long sequence is contained in the training dataset only 10 times in total, all in the same document

Memorized String	Sequence Length	Occurrences in Data	
		Docs	Total
Y2...██████...y5	87	1	10
7C...██████...18	40	1	22
XM...██████...WA	54	1	36
ab...██████...2c	64	1	49
ff...██████...af	32	1	64
C7...██████...ow	43	1	83
0x...██████...C0	10	1	96
76...██████...84	17	1	122
a7...██████...4b	40	1	311

Memorization increases with larger model size

- Snippets of Reddit URLs that appear a varying number of times in a single training document
- The GPT-2 prompt began with a Reddit URL
- Task: generate the the corresponding URL verbatim in the first 10,000 generations
- 1/2 if the URL is generated by providing GPT-2 with the first 6 characters of the URL
- GPT-2 XL: 1.5 billion parameters
- GPT-3: 175 billion parameters
- Even larger models exist

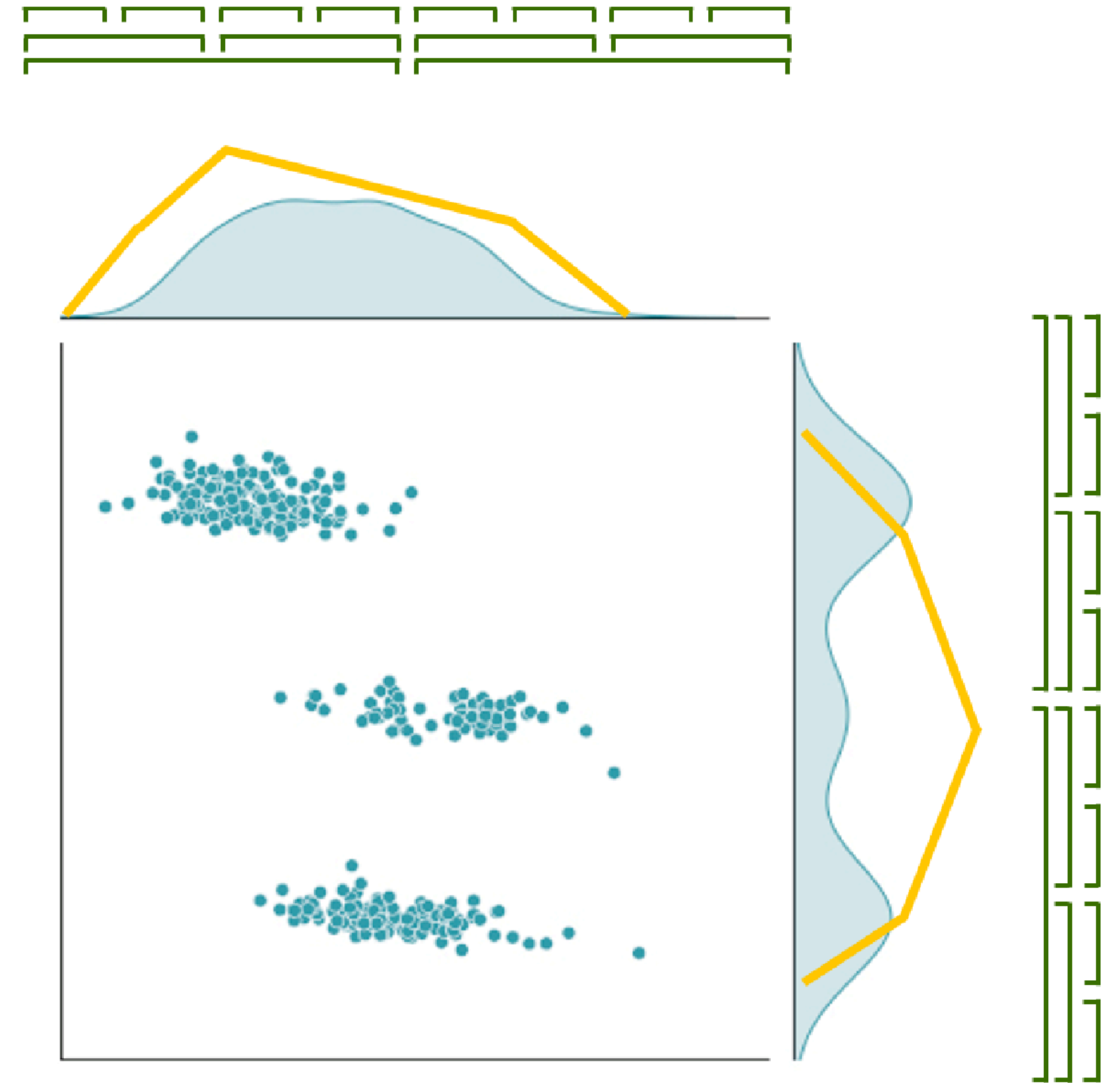
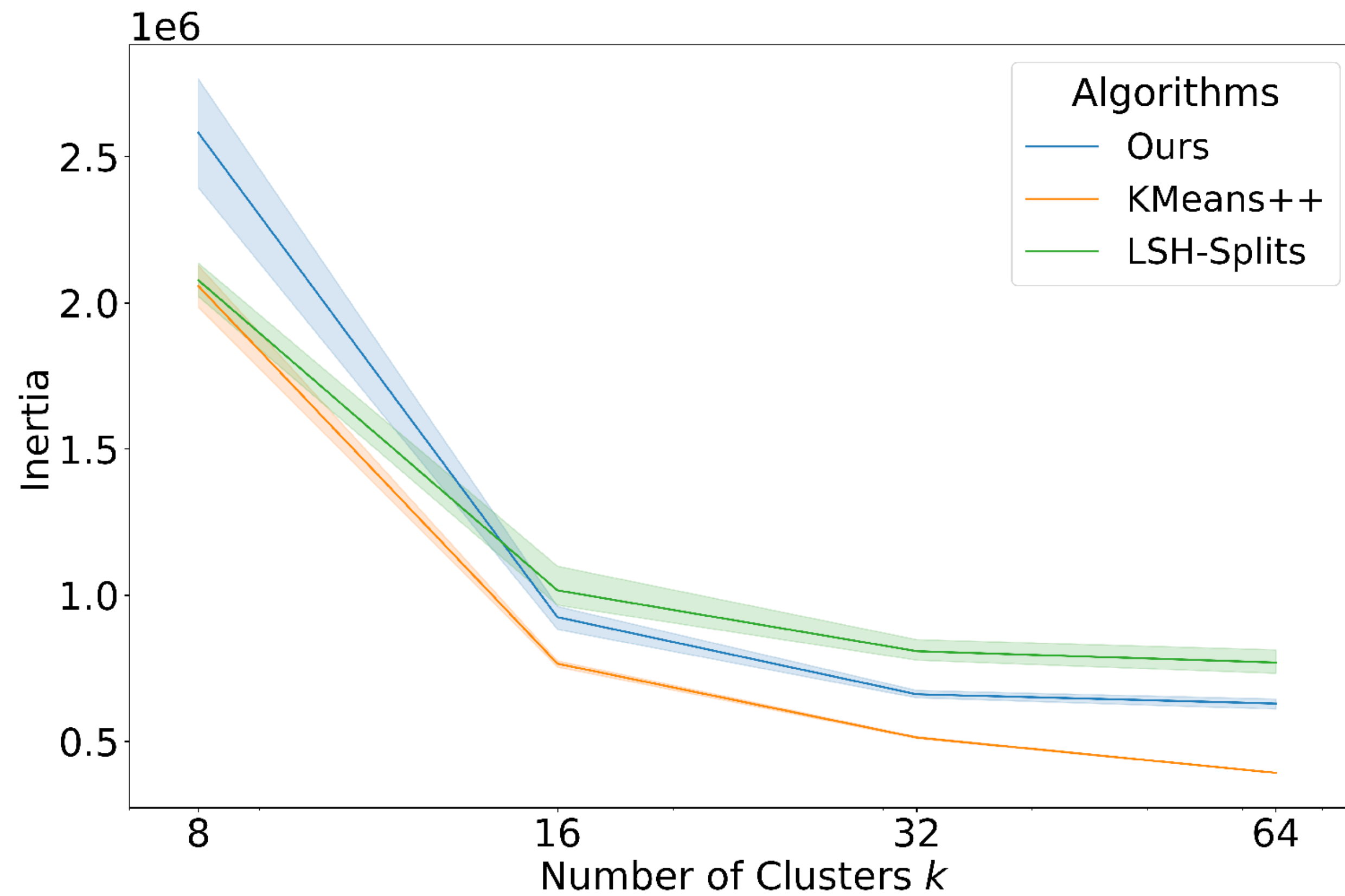
URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████1p8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

Some recent results

- DP Mondrian: Privacy-Preserving Clustering
 - *Yara Schütt, Johannes Liebenow, Marcel Gehrke, Tanya Braun, Esfandiar Mohammadi*
- DP Helmet: Privacy-Preserving Distributed Softmax-Layer Learning
 - *Moritz Kirschte, Sebastian Meiser, Saman Ardalan, Esfandiar Mohammadi*
 - E.g., for privacy-preserving distributed fine-tuning of the last layer
- Improved DP GBDT:
Privacy-Preserving Distributed Learning of Gradient Boosting Decision Trees
 - *Thorsten Peinemann, Moritz Kirschte, Joshua Stock, Carlos Cotrini, Esfandiar Mohammadi*
- n.b.: Some people call distributed learning also federated learning

DPM: Privacy-Preserving Clustering

- Search for splits instead of point concentrations



(b) MNIST embeddings (10 classes)

Strong differential privacy guarantees: $\epsilon = 1$

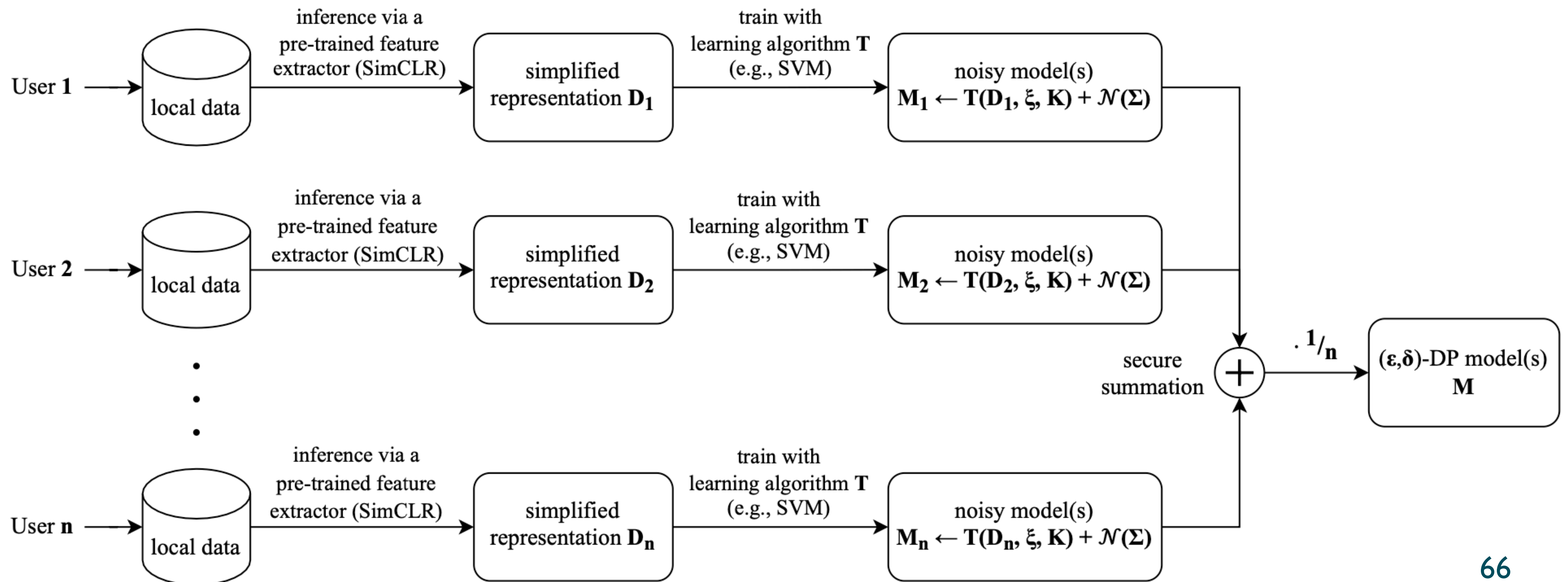
Frugal DP GBDT

- Privacy-Preserving Distributed Learning of Gradient Boosting Decision Trees
- Utilizing several privacy-boosting techniques
 - Individual Privacy Filters, Nested Composition, Subsampling
- Initial results:
 - Factor 20 better privacy for the same utility:
 - For strong privacy guarantees ($\epsilon = 0.5$), we achieve the same performance (RMSE = 2.55 on Abalone) as prior work for very weak privacy guarantees ($\epsilon = 10$)
 - For a medium-sized data set: 3000 training data points (Abalone)



DP Helmet: Distributed softmax-layer learning

- DP Helmet: Privacy-Preserving Distributed Softmax-Layer Learning
- We prove strong convexity for training a softmax-layer
- We utilize this strong convexity to get very good utility privacy tradeoffs



Results on CIFAR-10

