

Group-based Robustness: A General Framework for Customized Robustness in the Real World

Weiran Lin¹

Keane Lucas¹

Neo Eyal³

Lujo Bauer¹

Michael K. Reiter²

Mahmood Sharif³

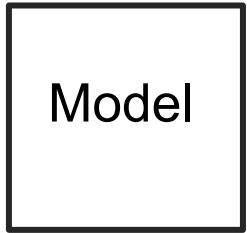
¹ Carnegie Mellon University

² Duke University

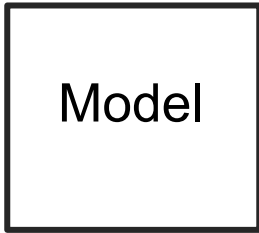
³ Tel Aviv University

What is robustness (against evasion attacks)?

What are evasion attacks?

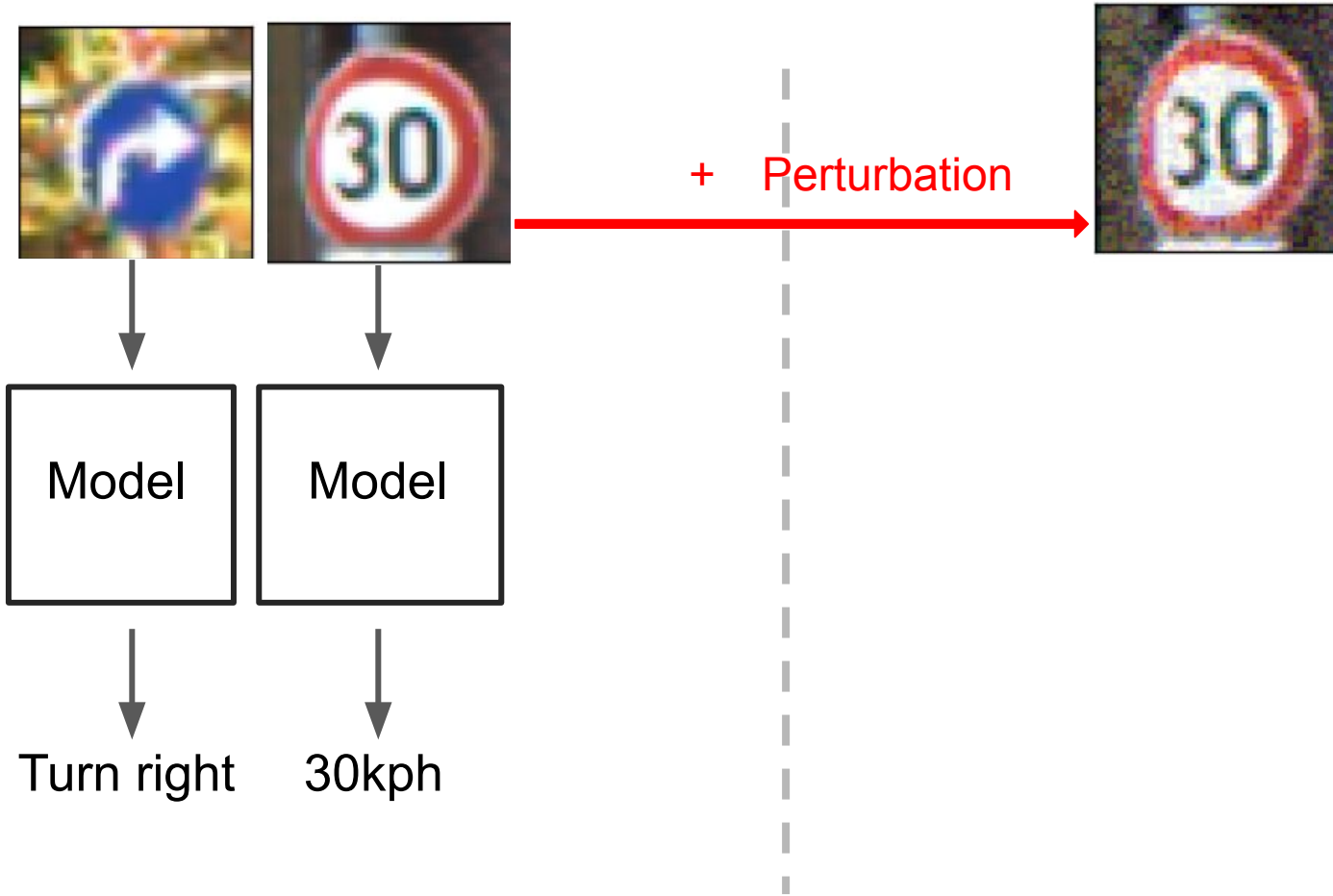


Turn right



30kph

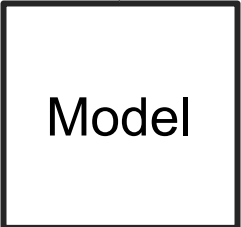
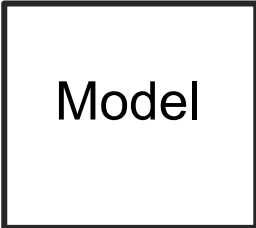
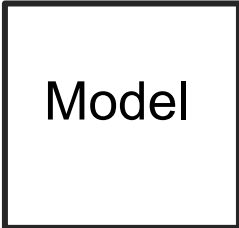
What are evasion attacks?



What are evasion attacks?



+ Perturbation



Turn right

30kph

Turn right!

Types of evasion attacks (adversary goals)

Adversarial
example

Untargeted

*Any misclassification implies
success*

Model

Not “30kph”

Types of evasion attacks (adversary goals)

Adversarial example

Untargeted

Any misclassification implies success

Model

Not “30kph”

Adversarial example

Targeted

Only a specific misclassification implies success

Model

“Turn right!”

How do we evaluate a model's robustness?

Adversarial
example



Model

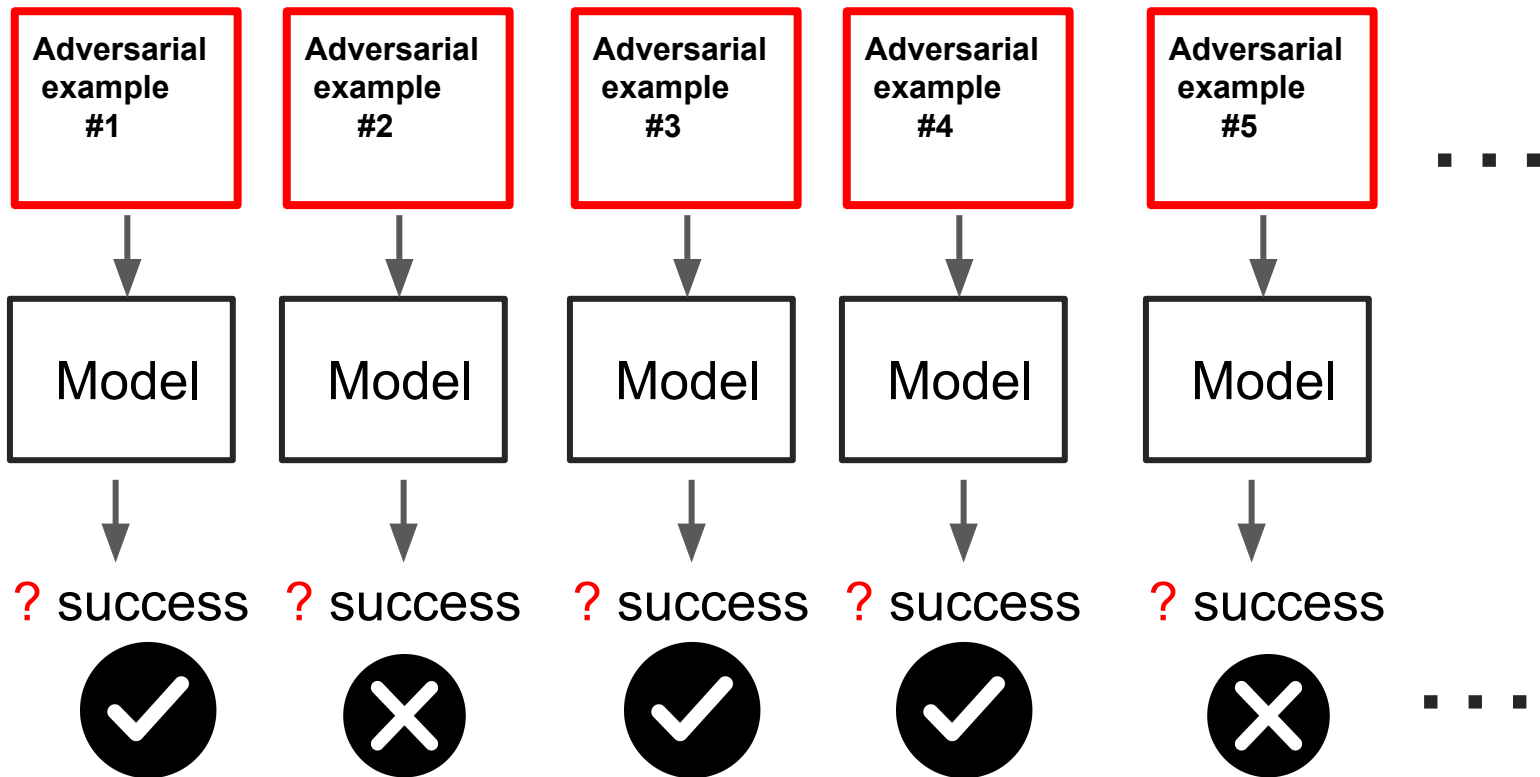


? success



How do we evaluate a model's robustness?

by measuring the fraction of inputs on which attacks succeed



Prior to our paper

- “Robustness” is defined as either targeted or untargeted

Prior to our paper

- “Robustness” is defined as either targeted or untargeted
- “Robustness” is measured on a per-input-instance basis
 - i.e. counting how many instances attacks failed on

Current robustness metrics
do not always work

Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors

Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot

Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers



Lujo



Milla

Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers
- A correct classification \neq threat

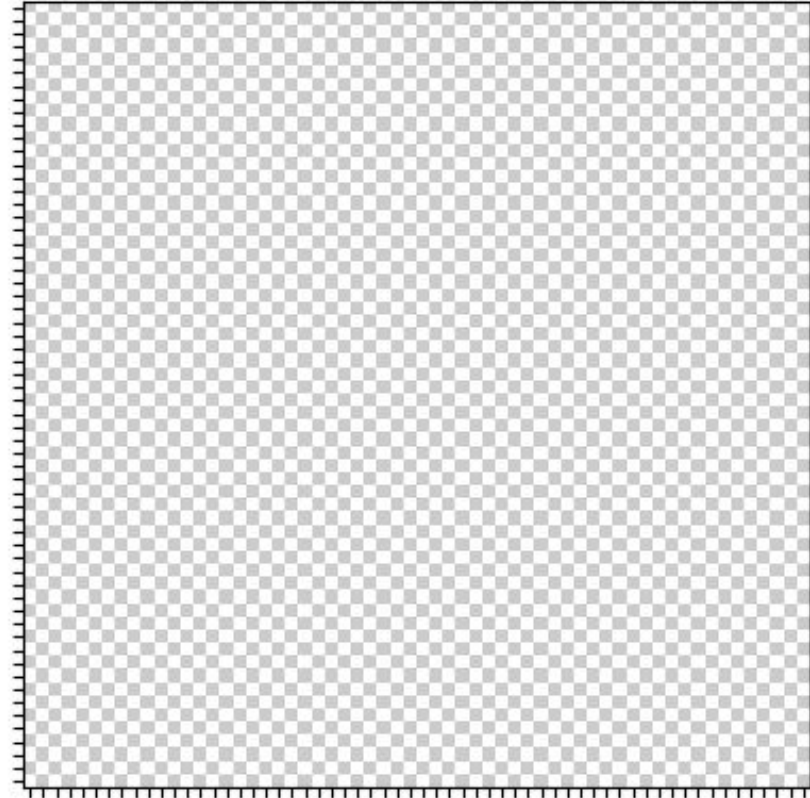
Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers
- A correct classification \neq threat
- A misclassification $=?$ threat

Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers
- A correct classification \neq threat
- A misclassification $=?$ threat

Classified as

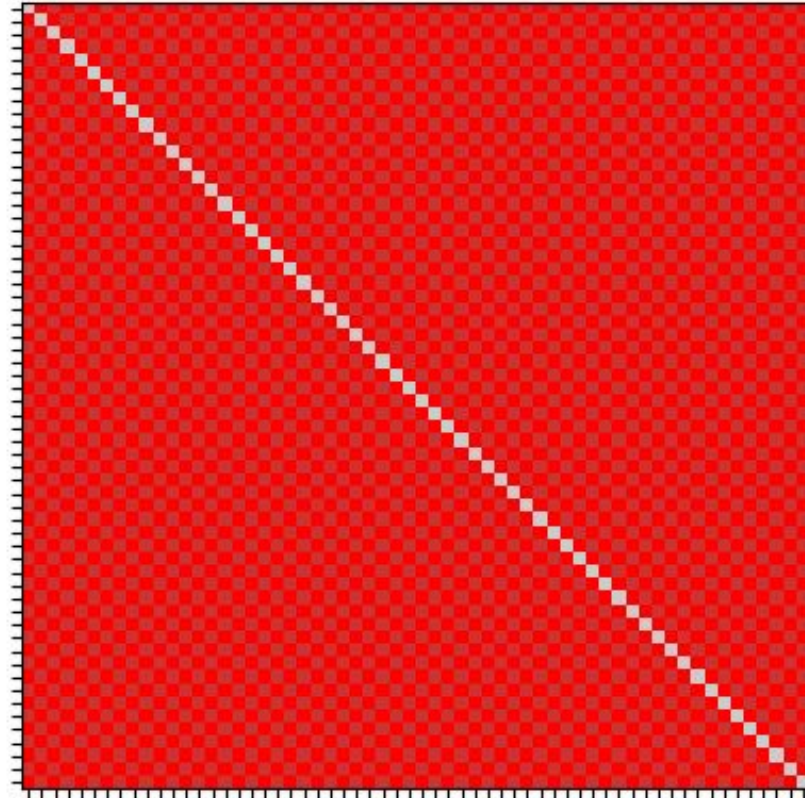


Real identities

Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers
- A correct classification \neq threat
- A misclassification $=?$ threat

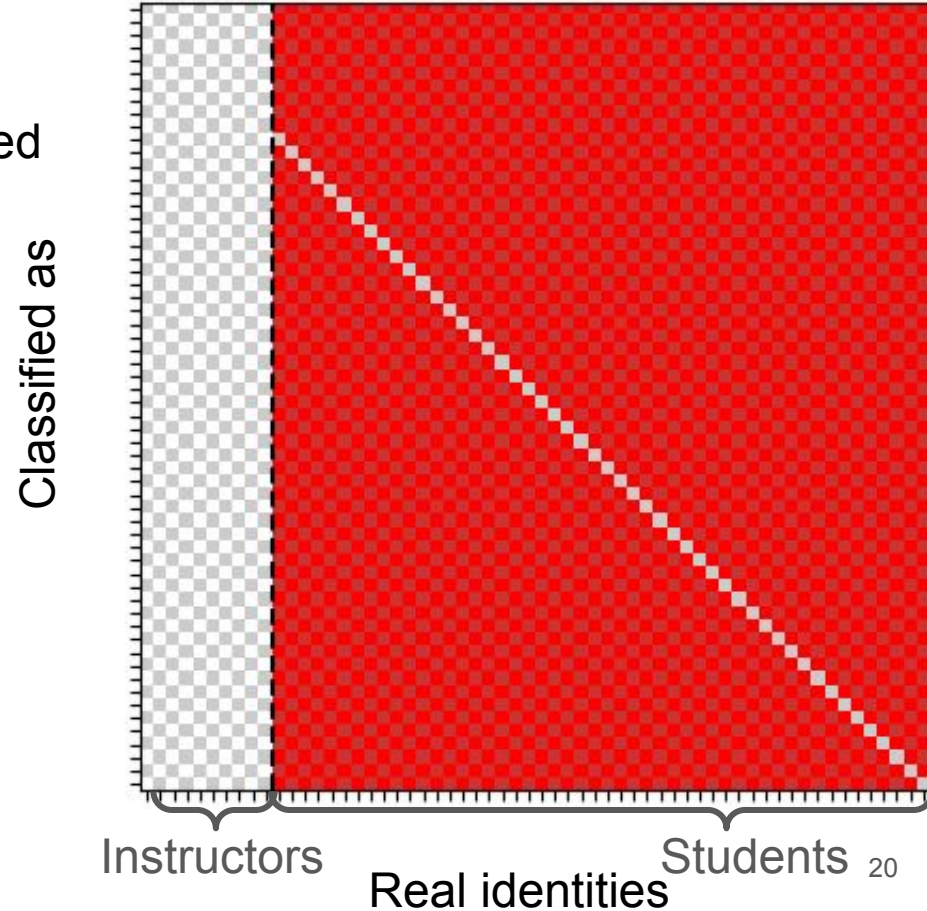
Classified as



Real identities

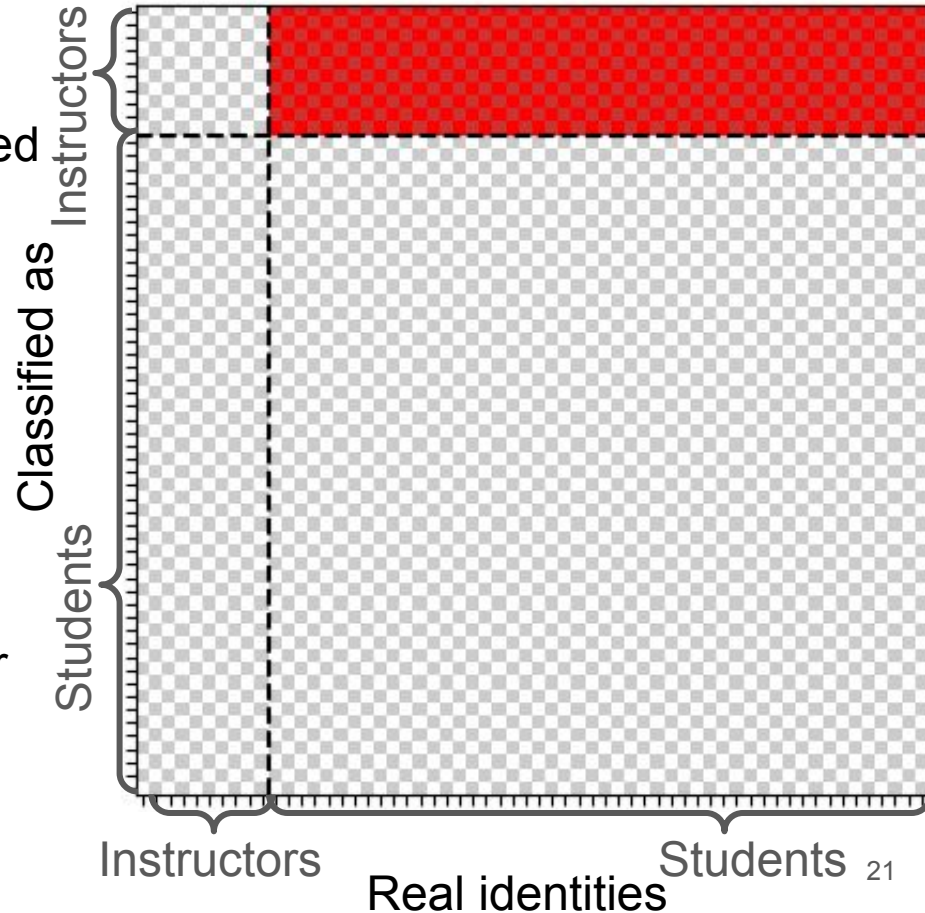
Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers
- A correct classification \neq threat
- A misclassification $=?$ threat
 - Instructors gain no benefit by impersonating anyone



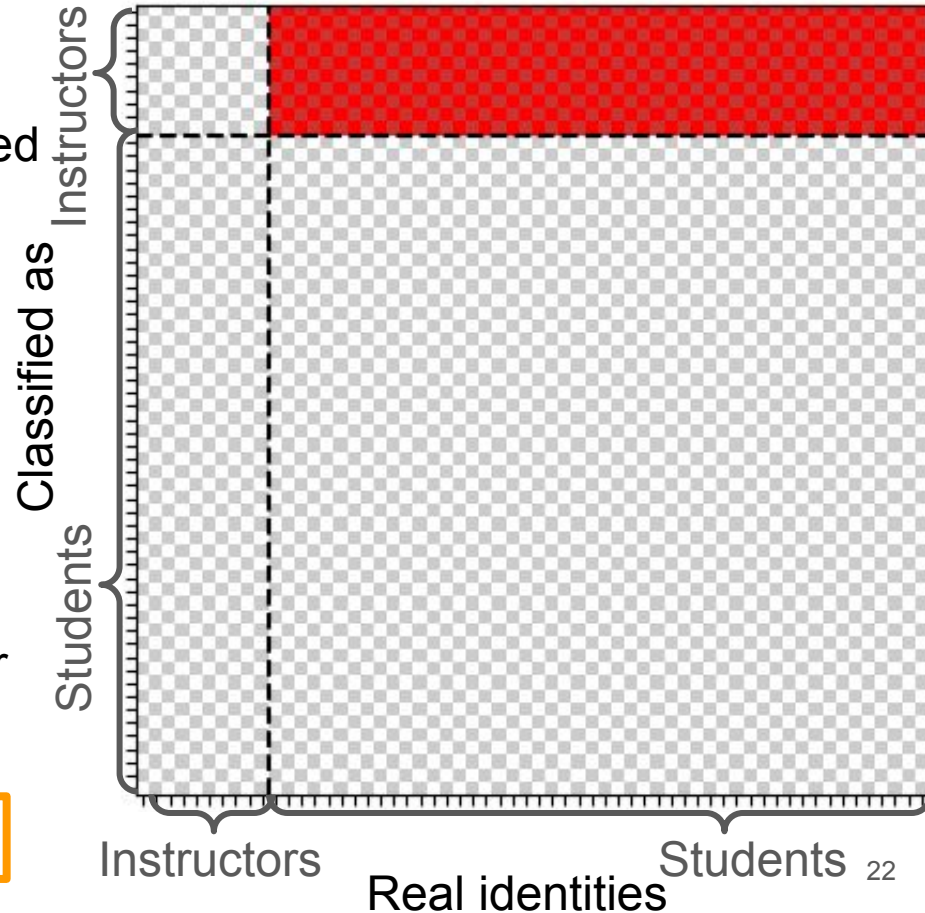
Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers
- A correct classification \neq threat
- A misclassification $=?$ threat
 - Instructors gain no benefit by impersonating anyone
 - Students, after impersonating other students, still cannot get access



Scenario 1: students try to steal unreleased answers

- 50 students and 10 instructors
- Instructors can legally access unreleased answers, students cannot
 - Students try to impersonate *any* instructor to steal answers
- A correct classification \neq threat
- A misclassification $=?$ threat
 - Instructors gain no benefit by impersonating anyone
 - Students, after impersonating other students, still cannot get access
- Only students \Rightarrow instructors is a **threat**



Scenario 1A: students try to steal unreleased answers

- 50 students and 10 instructors
 - From different classes

Scenario 1A: students try to steal unreleased answers

- 50 students and 10 instructors
 - From different classes
- Instructors can legally access unreleased answers, students cannot

Scenario 1A: students try to steal unreleased answers

- 50 students and 10 instructors
 - From different classes
- Instructors can legally access unreleased answers, students cannot
- Only students => instructors is a **threat**

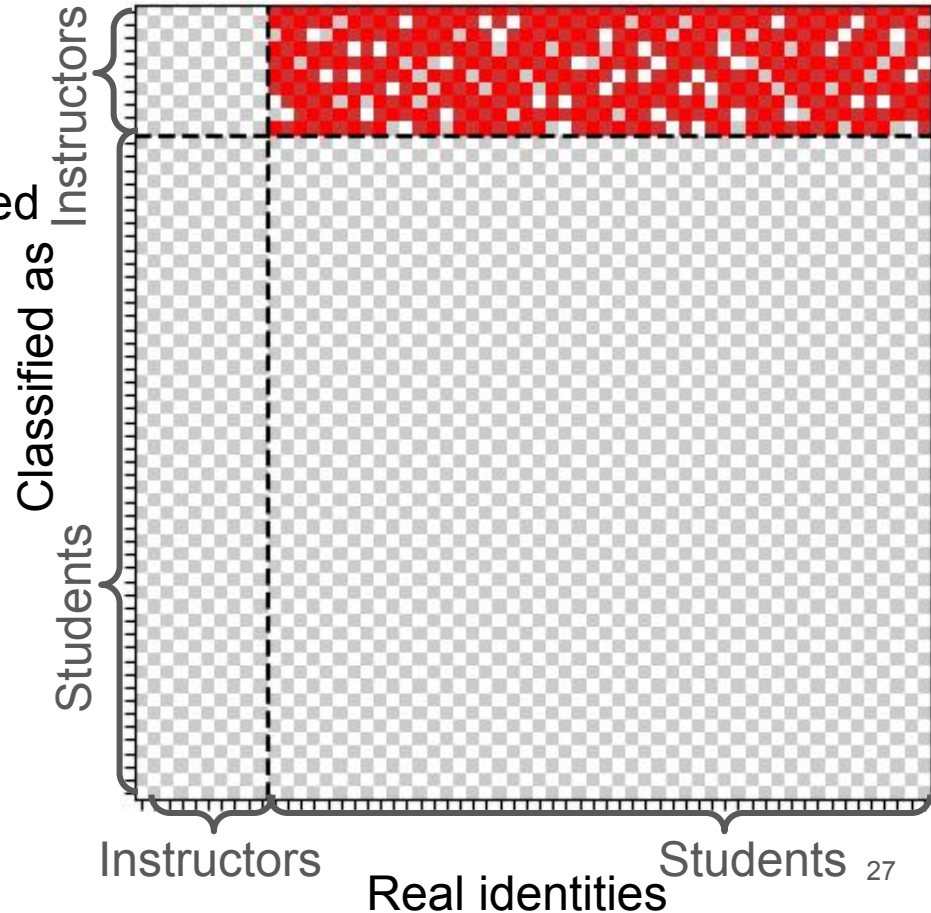
Scenario 1A: students try to steal unreleased answers

- 50 students and 10 instructors
 - From different classes
- Instructors can legally access unreleased answers, students cannot
- Only students => instructors is a **threat**

- Students may have different set of instructors

Scenario 1A: students try to steal unreleased answers

- 50 students and 10 instructors
 - From different classes
- Instructors can legally access unreleased answers, students cannot
- Only students => instructors is a **threat**
- Students may have different set of instructors
- Students may want to impersonate different set of instructors



Current robustness metrics
do not capture the risk in
students => instructors

Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree



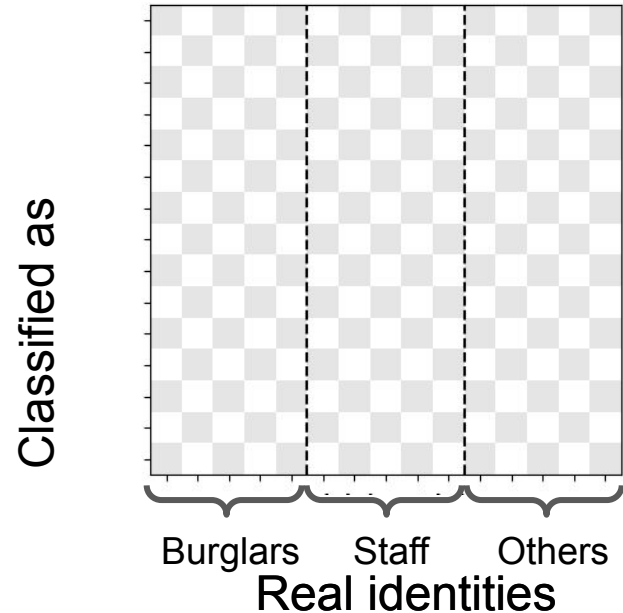
Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members



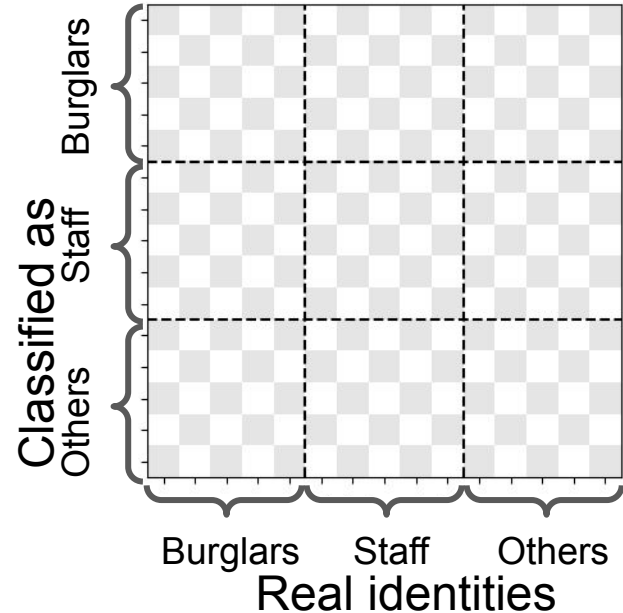
Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members
- For example,
 - five burglars,
 - five staff members,
 - and five others



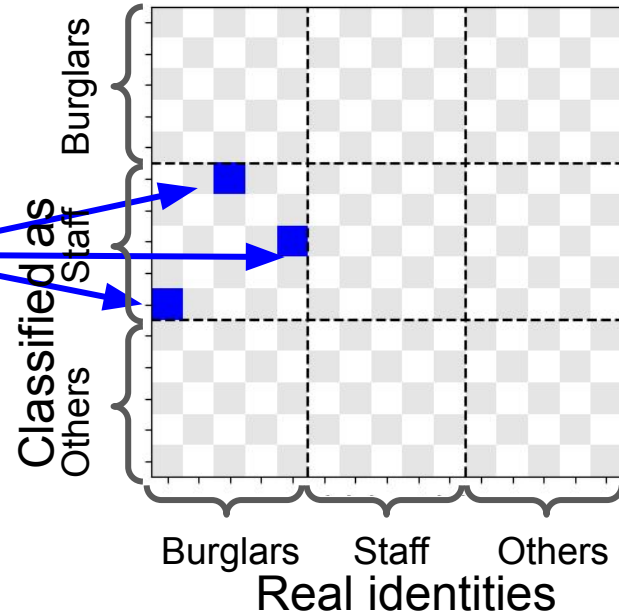
Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members
- For example,
 - five burglars,
 - five staff members,
 - and five others



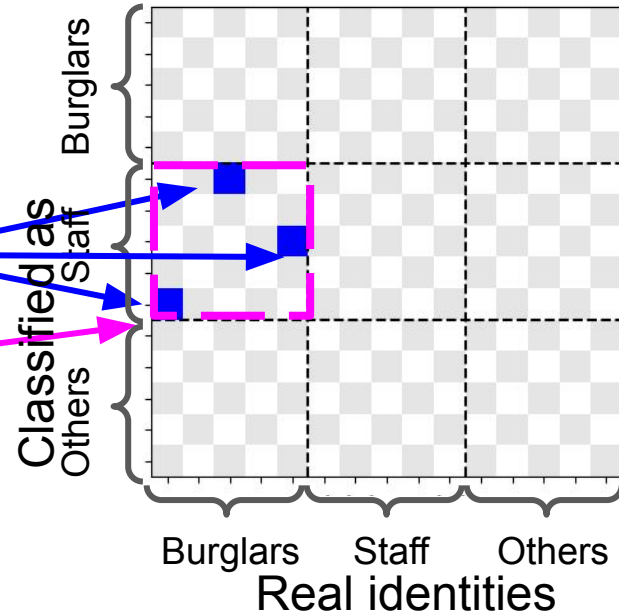
Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members
- For example,
 - five burglars,
 - five staff members,
 - and five others
- Attackers win if all *these three misclassifications* happen simultaneously



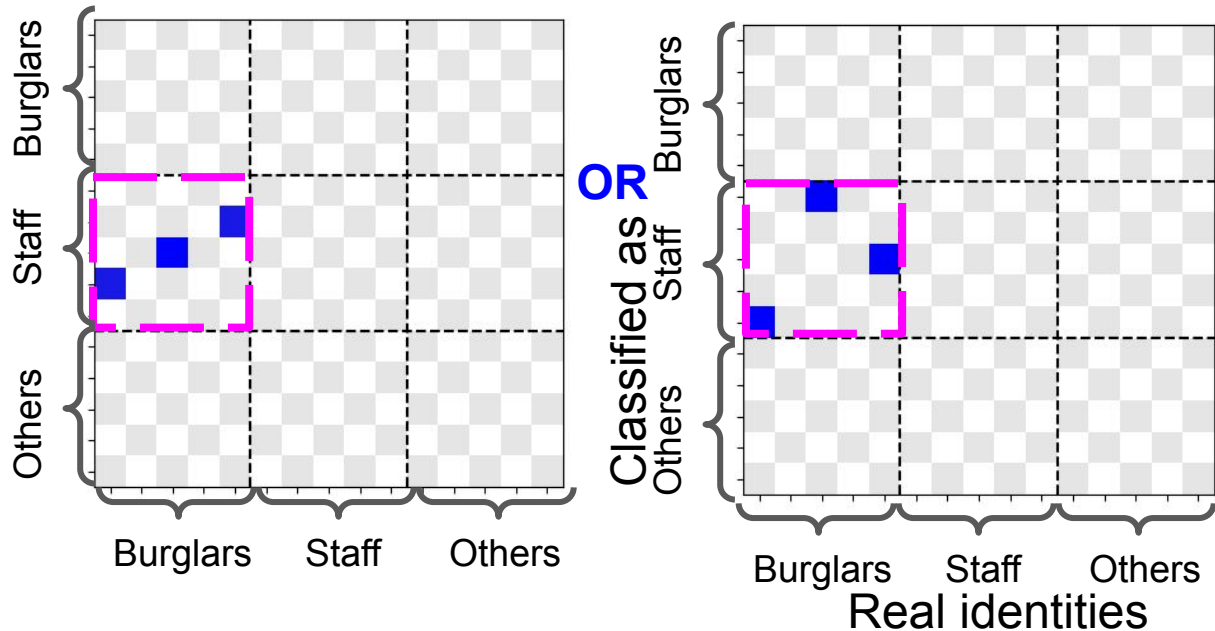
Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members
- For example,
 - five burglars,
 - five staff members,
 - and five others
- Attackers win if all *these three misclassifications* happen simultaneously
- *Impersonate three different staff members*



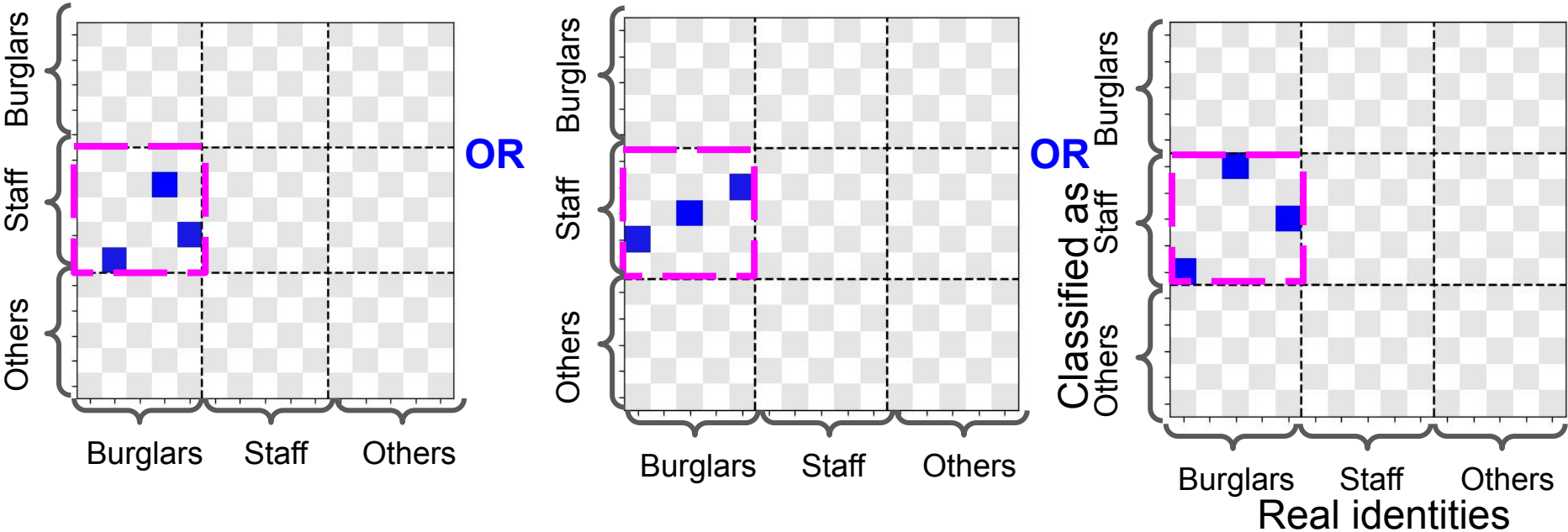
Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members



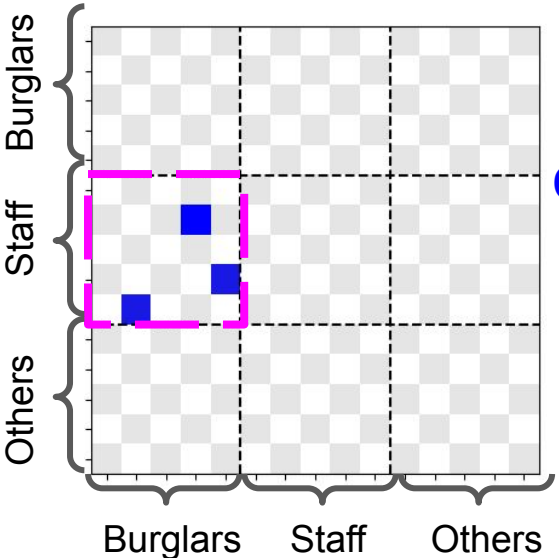
Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members

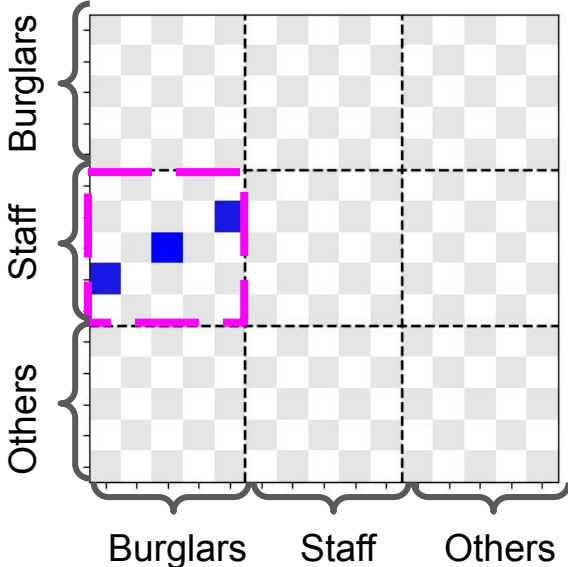


Scenario 2: bank burglary

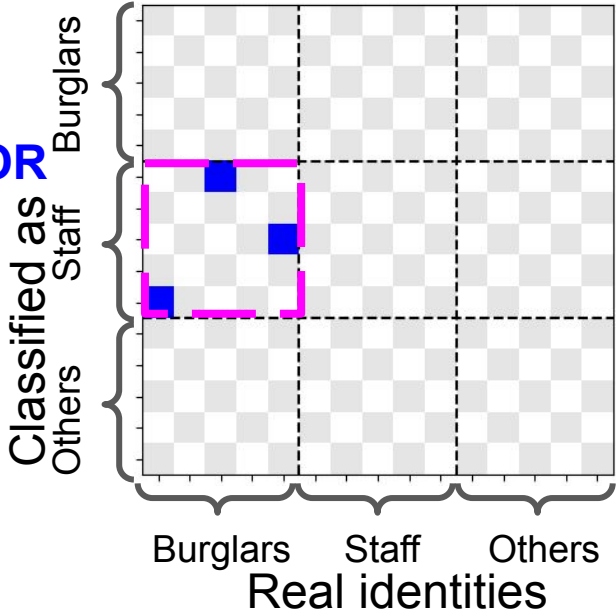
- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members



OR



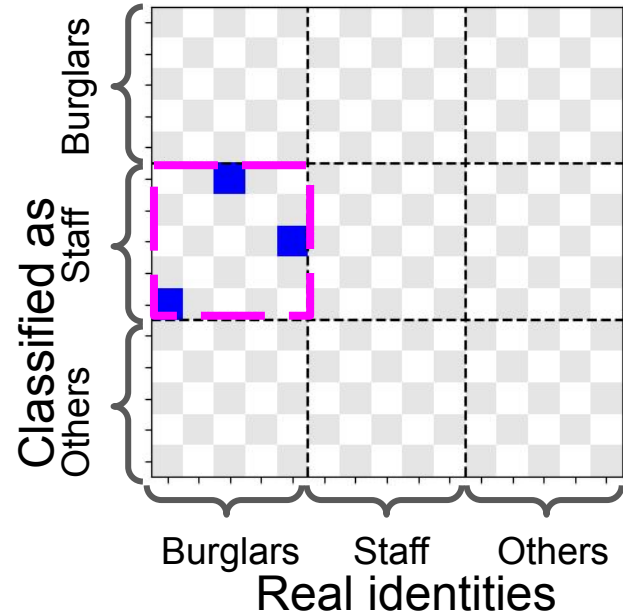
OR



OR ...

Scenario 2: bank burglary

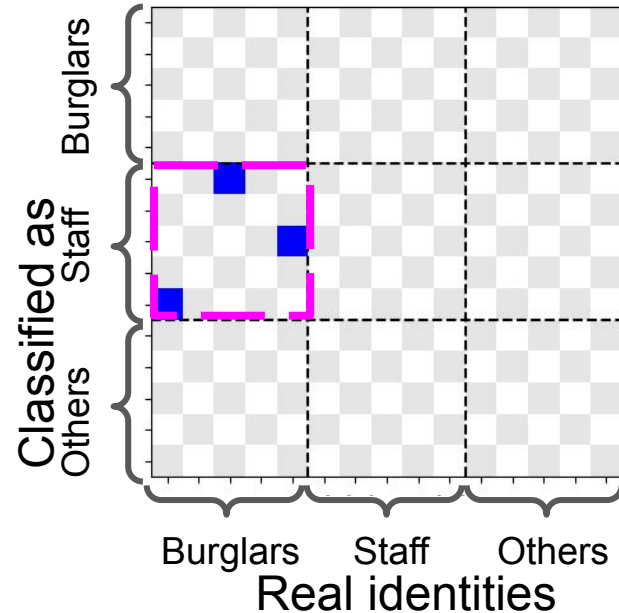
- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members
- Attacks can still succeed:
 - If some burglars do not impersonate anyone
 - If some staff members are not impersonated



Scenario 2: bank burglary

- A vault can be opened only if **three** staff members are present and agree
- A group of burglars (≥ 3) try to impersonate staff members
- Attacks can still succeed:
 - If some burglars do not impersonate anyone
 - If some staff members are not impersonated

Robustness cannot be evaluated on a per-input-instance basis



Current robustness
metrics do not capture the
risk in burglars => staff

Prior to our paper

- “Robustness” is defined as either targeted or untargeted
- “Robustness” is measured on a per-input-instance basis
 - i.e. counting how many instances attacks failed on

Prior to our paper

- “Robustness” is defined as either targeted or untargeted
- “Robustness” is measured on a per-input-instance basis
 - i.e. counting how many instances attacks failed on
- In some practical scenarios (e.g., grades, bank vault):
 - previous definition of “robustness” might not tell us how likely these attackers are to succeed!

Our contributions

- New definitions of robustness that better assess risk (#1)

Our contributions

- New definitions of robustness that better assess risk (#1)
- Enabled by new definitions:
 - **Faster attacks** (#2)
 - **Better defenses** (#3)

Contribution #1: Better assessment of risk

- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks

Contribution #1: Better assessment of risk

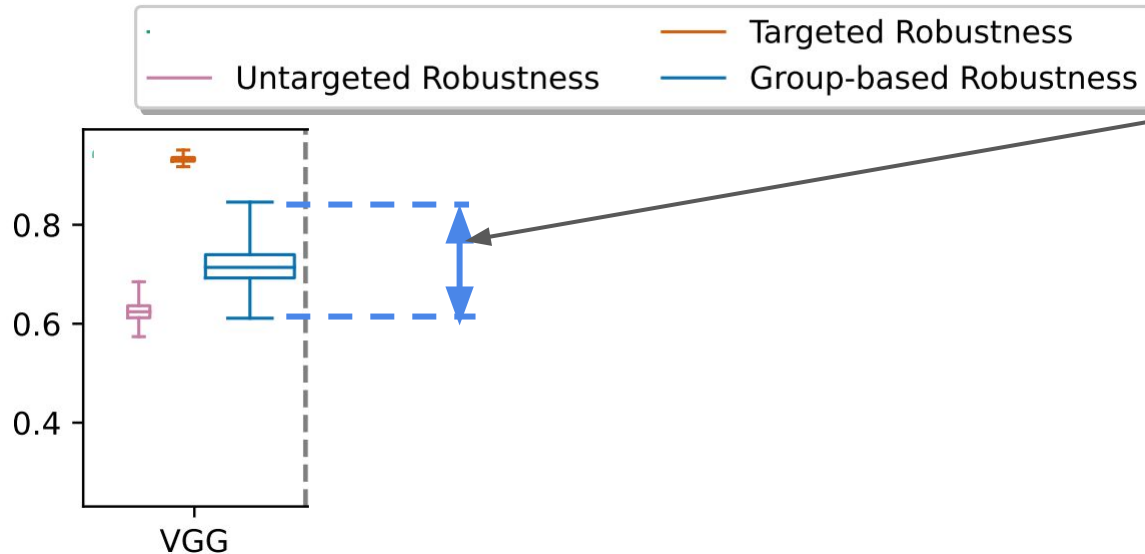
- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks
 - Similar to a cryptography game

Contribution #1: Better assessment of risk

- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks
 - Similar to a cryptography game
 - Targeted and Untargeted robustness are special cases of group-based robustness
 - i.e. when the game has specific parameters

Contribution #1: Better assessment of risk

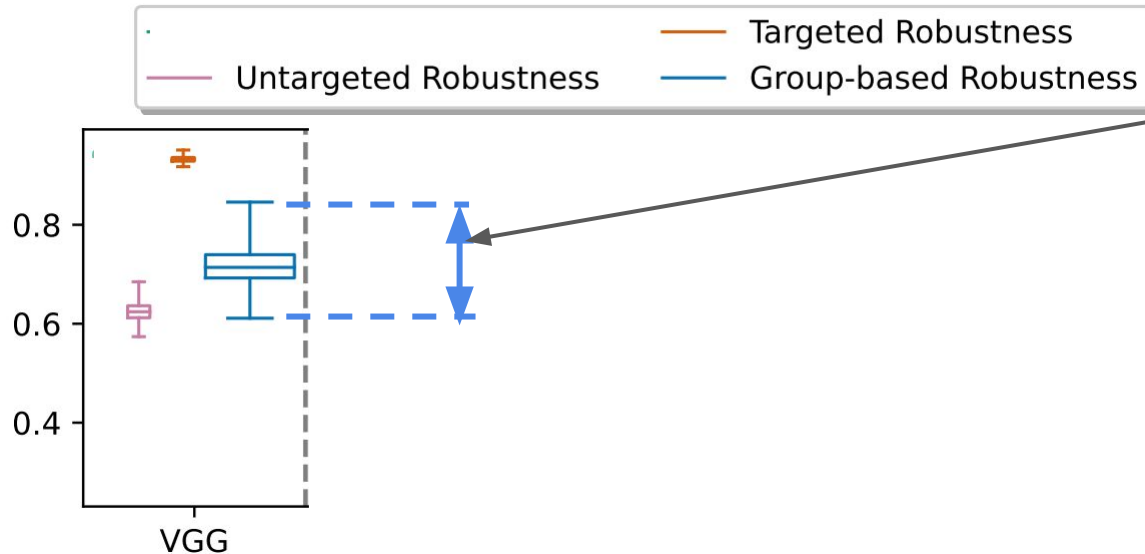
- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks



New metric reveals much bigger variance between model instances within the same architecture

Contribution #1: Better assessment of risk

- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks



New metric reveals much bigger variance between model instances within the same architecture

Our new metric suggests that models differ more than previously thought

Contribution #1: Better assessment of risk

- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks
- Group-based robustness is ~ uncorrelated with existing metrics


Contribution #1: Better assessment of risk

- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks
- Group-based robustness is ~ uncorrelated with existing metrics
 - Implication: none of existing metrics can substitute our new metric

Contribution #1: Better assessment of risk

- We formally define *group-based robustness* as a new metric that more accurately reflects true threat of attacks
- Group-based robustness is ~ uncorrelated with existing metrics
 - Implication: none of existing metrics can substitute our new metric
 - Previously believed stronger defenses might be weaker by new metric

Our contributions

- New definitions of robustness that better assess risk (#1)
- Enabled by new definitions:
 - **Faster attacks (#2)** 
 - **Better defenses (#3)**

Contribution #2: Faster attacks

- We found computationally cheaper ways to estimate group-based robustness


Contribution #2: Faster attacks

- We found computationally cheaper ways to estimate group-based robustness
- Our attacks
 - find a similar number of attacks while faster by $|T|$ (e.g. number of instructors)

Contribution #2: Faster attacks

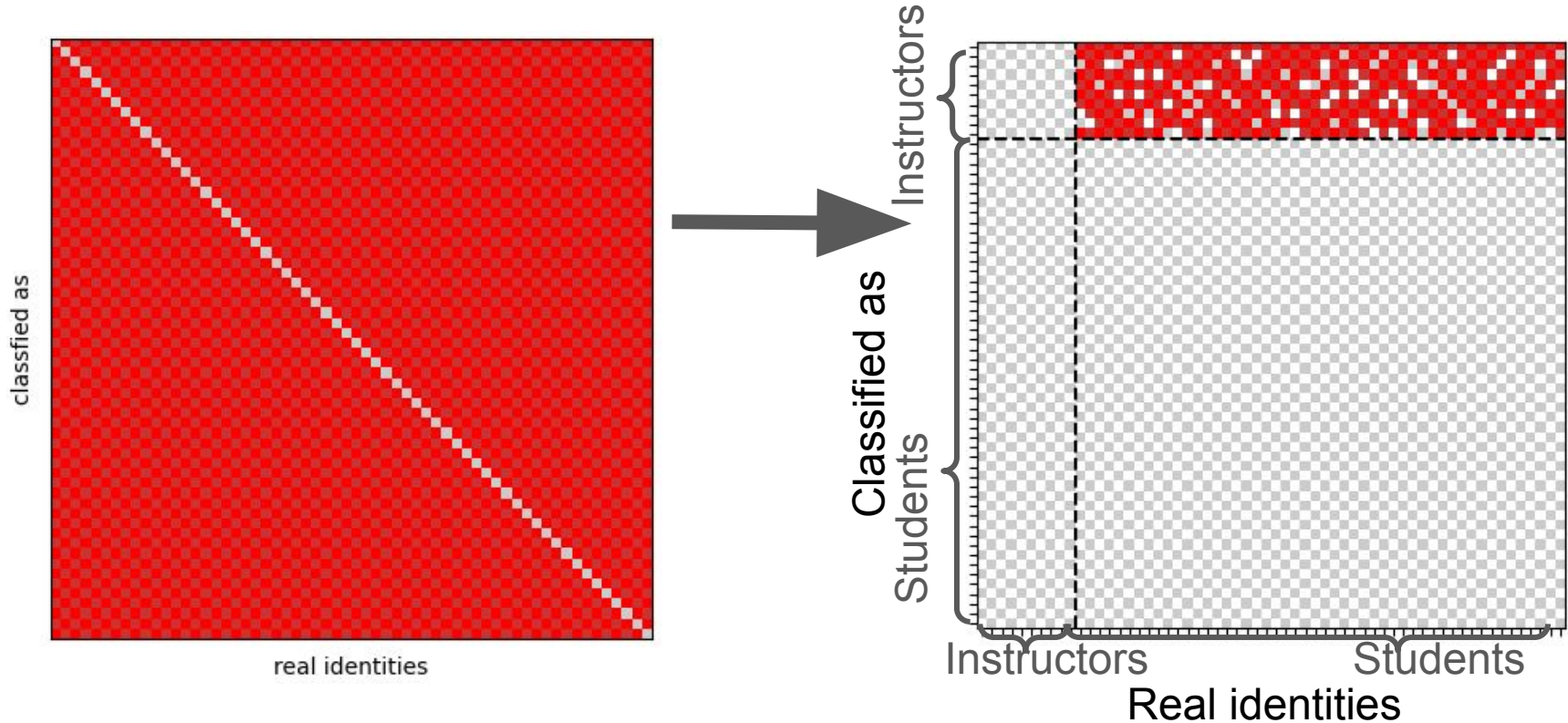
- We found computationally cheaper ways to estimate group-based robustness
- Our attacks
 - find a similar number of attacks while faster by $|T|$ (e.g. number of instructors)
 - or find many more attacks using the same amount of time

Our contributions

- New definitions of robustness that better assess risk (#1)
- Enabled by new definitions:
 - **Faster attacks** (#2)
 - **Better defenses** (#3) 

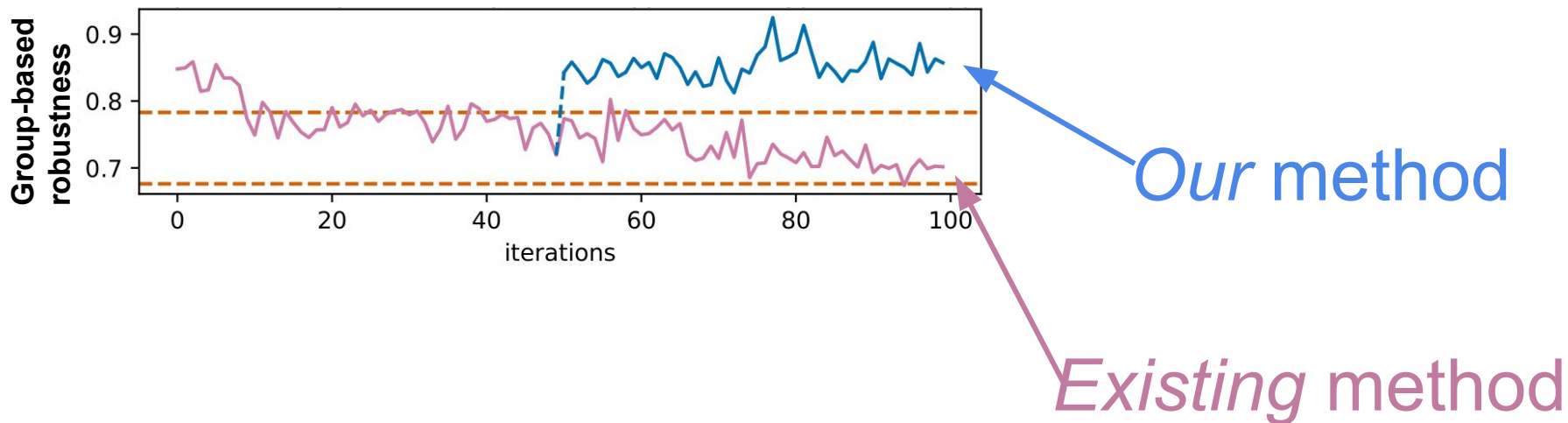
Contribution #3: Better defenses

- Awareness of threat => can train models to better avoid it



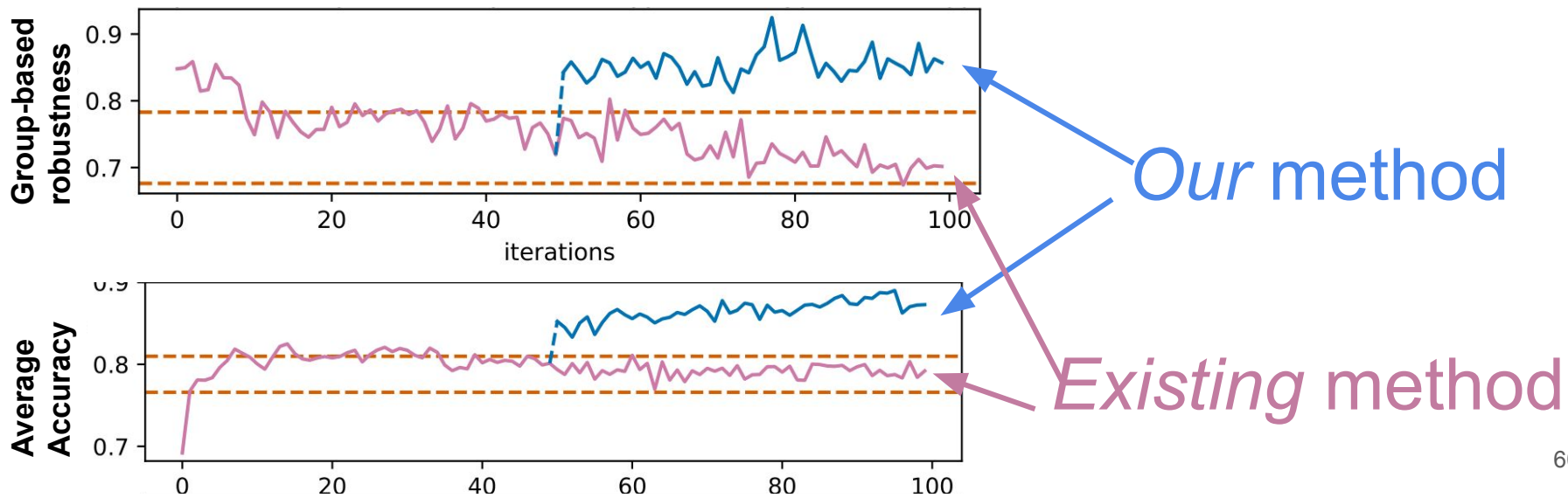
Contribution #3: Better defenses

- Awareness of threat => can train models to better avoid it
- We built defenses that outperform existing ones in:
 - Group-based robustness



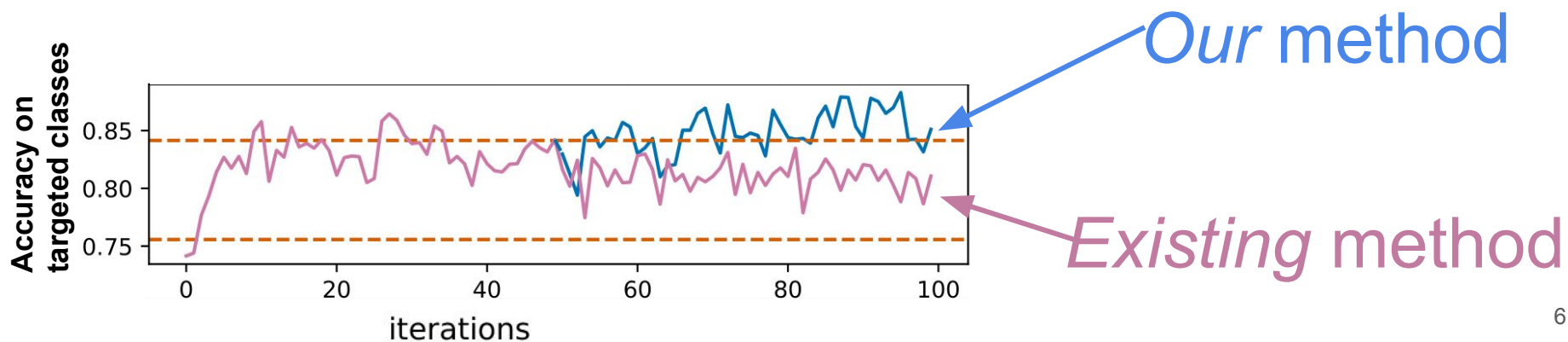
Contribution #3: Better defenses

- Awareness of threat => can train models to better avoid it
- We built defenses that outperform existing ones in:
 - Group-based robustness
 - Average accuracy



Contribution #3: Better defenses

- Awareness of threat => can train models to better avoid it
- We built defenses that outperform existing ones in:
 - Group-based robustness
 - Average accuracy
 - Accuracy on targeted classes:
 - Never predicting instructors is not a solution!



In summary

- Robustness is not always targeted or untargeted
- Robustness sometimes cannot be measured on a per-input-instance basis

In summary

- Robustness is not always targeted or untargeted
- Robustness sometimes cannot be measured on a per-input-instance basis

- **We need (and now have) better definitions of robustness**

In summary

- Robustness is not always targeted or untargeted
- Robustness sometimes cannot be measured on a per-input-instance basis

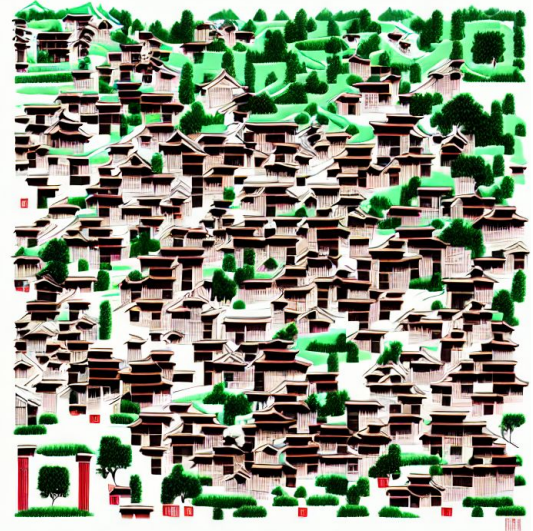
- We **need (and now have)** better definitions of robustness
- New definitions enable:
 - **Better assessment of risk**
 - **Faster attacks**
 - **Better defenses**

In summary

- Robustness is not always targeted or untargeted
- Robustness sometimes cannot be measured on a per-input-instance basis

- We need (and now have) better definitions of robustness
- New definitions enable:
 - **Better assessment of risk**
 - **Faster attacks**
 - **Better defenses**

- Check out our paper for more details!
 - Scan this



Group-based Robustness: A General Framework for Customized Robustness in the Real World



- Robustness is not always targeted or untargeted
- Robustness sometimes cannot be measured on a per-input-instance basis

- We need (and now have) better definitions of robustness
- New definitions enable:
 - **Better assessment of risk**
 - **Faster attacks**
 - **Better defenses**

- Check out our paper for more details!
 - Scan this

