UNIVERSITY OF TORONTO

VECTOR INSTITUTE

ECE1784H: Trustworthy Machine Learning

Prof. Nicolas Papernot

nicolas.papernot@utoronto.ca

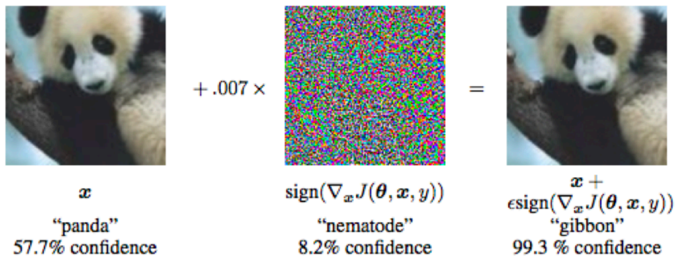
Logistics

- Course syllabus: papernot.fr/teaching/f19-trustworthy-ml
  - Schedule (will be updated)
  - Assigned reading (will be updated)
  - Assignment description
  - Grading information
  - Ethics statement
- Class: Mondays 1-3pm
- Office hours: Wednesdays 1.30-3.30pm
- Office location: Pratt 484E

What is this class?

<u>This is not a ML course</u>

What do I mean by trustworthy ML?

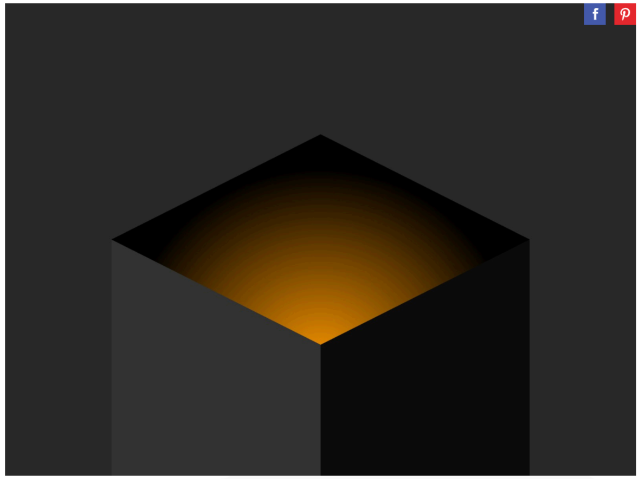


Security

Privacy

ANDY GREENBERG SECURITY 09.30.16 11:06 AM

HOW TO STEAL AN AI



Confidentiality

The New York Times

Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

Fairness & Ethics

Safety

Again, this is not a ML course.

- Exam
  - Questions will test ML background
  - No studying is necessary if you have taken a ML course
  - Friday September 13 from 4PM to 5PM
  - BA1170
  - Let me know by email before tomorrow (Tuesday) noon if you have a conflict for this time so I can arrange a time for you to take the exam in my office
  - 30% of grade
  - If you are unable to answer questions, I strongly recommend dropping the course and taking a ML class first (e.g., ECE1513 in the Winter)

Format for weeks 2-10

- 1h30mn presentation of reading materials
  - Research papers
  - One team will present and lead the discussion
  - Interactive discussion (everyone should do the reading ahead of class)
  - One team will take notes and synthetize the discussion
- 30mn work on research projects
- Deadlines:
  - Thursday before class: presenting team shares slides by 6.00pm
  - Wednesday following class: notes team should turn in notes by 6.00pm

Before class: 1-page reading summary

- Read all papers posted on website
- Summarize your reading through 1 page summary:
  - what did the papers do well?
  - where did the papers fall short?
  - what did you learn from these papers?
  - what questions do you have about the papers?
- Typeset report in LaTeX (https://www.latex-project.org/)
  - First report not in LaTeX issued a warning
  - All following reports assigned 0

During class: notes + discussion

- All: ask questions from your 1-page summary
- Presenting team:
  - May choose an appropriate format
    - Slides
    - interactive demos
    - code tutorials
  - Should involve class
  - Should cover (at least) the papers assigned for reading
  - Time the presentation to last 1h30
- Notes team:
  - Takes notes to prepare report

Presentation rubric

- papernot.fr/teaching/rubric.pdf
- Technical:
  - Depth of content
  - Accuracy of content
  - Paper criticism
  - Discussion lead
- Soft presentation skills:
  - Time management
  - Responsiveness to audience
  - Organization
  - Presentation aids

After class: notes

- Notes team:
  - Synthesize both the presentation and questions / discussions
  - Report written collectively as a team
  - Typeset notes in LaTeX
    - recommended min 4 pages in default LaTeX article style
    - Include references

Lateness policy

- Paper presentations:
  - Deadline: slides must be turned in by 6pm on Thursday before the class
  - 10% per-day late penalty
  - up to a max of 2 days
- Paper summaries:
  - Deadline: beginning of each class
  - late assignments not accepted
  - 0 for the week if physical copy not turned in at the beginning of class
- Class notes:
  - Deadline: 6pm on Wednesday following the class
  - 10% per-day late penalty
  - up to a max of 4 days

Grading scheme

- 30% exam
- 20% paper presentation
- 10% paper summaries
- 10% class notes
- 30% research project

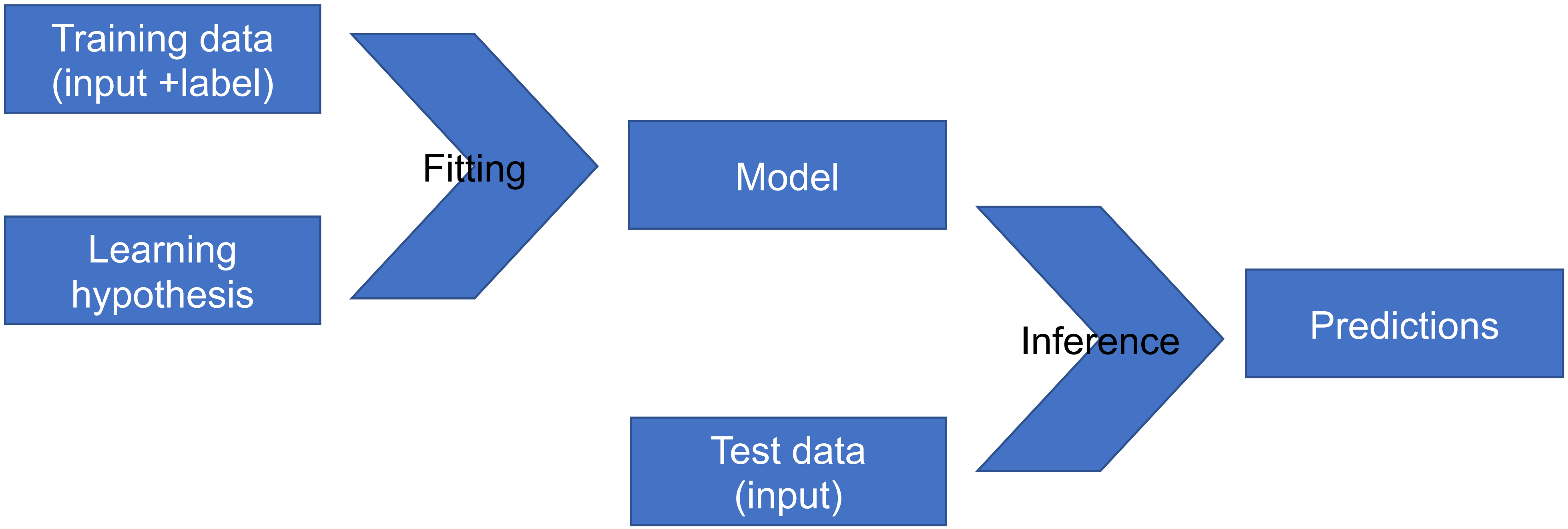
Integrity

Any instance of sharing or plagiarism, copying, cheating, or other disallowed behavior will constitute a breach of ethics. Students are responsible for reporting any violation of these rules by other students, and failure to constitutes an ethical violation that carries with it similar penalties.

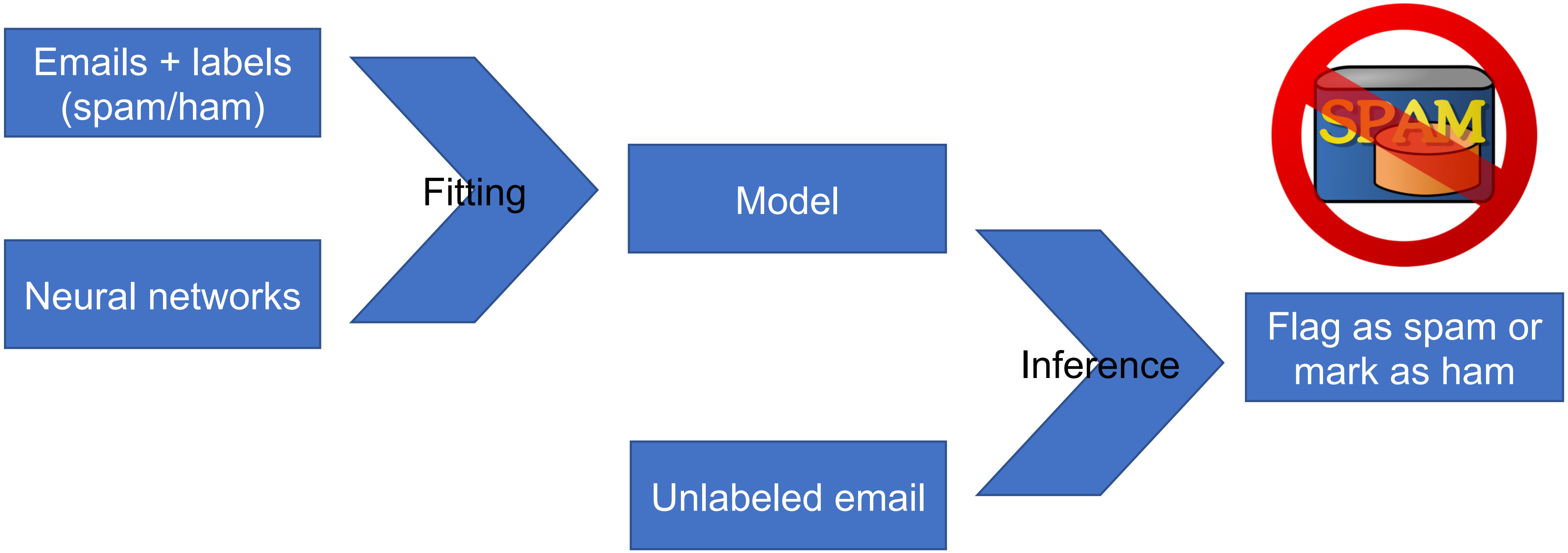
Ethics

This course covers topics in personal and public privacy and security. As part of this investigation we will explore technologies whose abuse may infringe on the rights of others. As an instructor, I rely on the ethical use of these technologies. Unethical use may include circumvention of existing security or privacy measurements for any purpose, or the dissemination, promotion, or exploitation of vulnerabilities of these services. Exceptions to these guidelines may occur in the process of reporting vulnerabilities through public and authoritative channels. Any activity outside the letter or spirit of these guidelines will be reported to the proper authorities and may result in dismissal from the class. When in doubt, please contact the course professor for advice. Do not undertake any action which could be perceived as technology misuse anywhere and/or under any circumstances unless you have received explicit permission from the instructor.

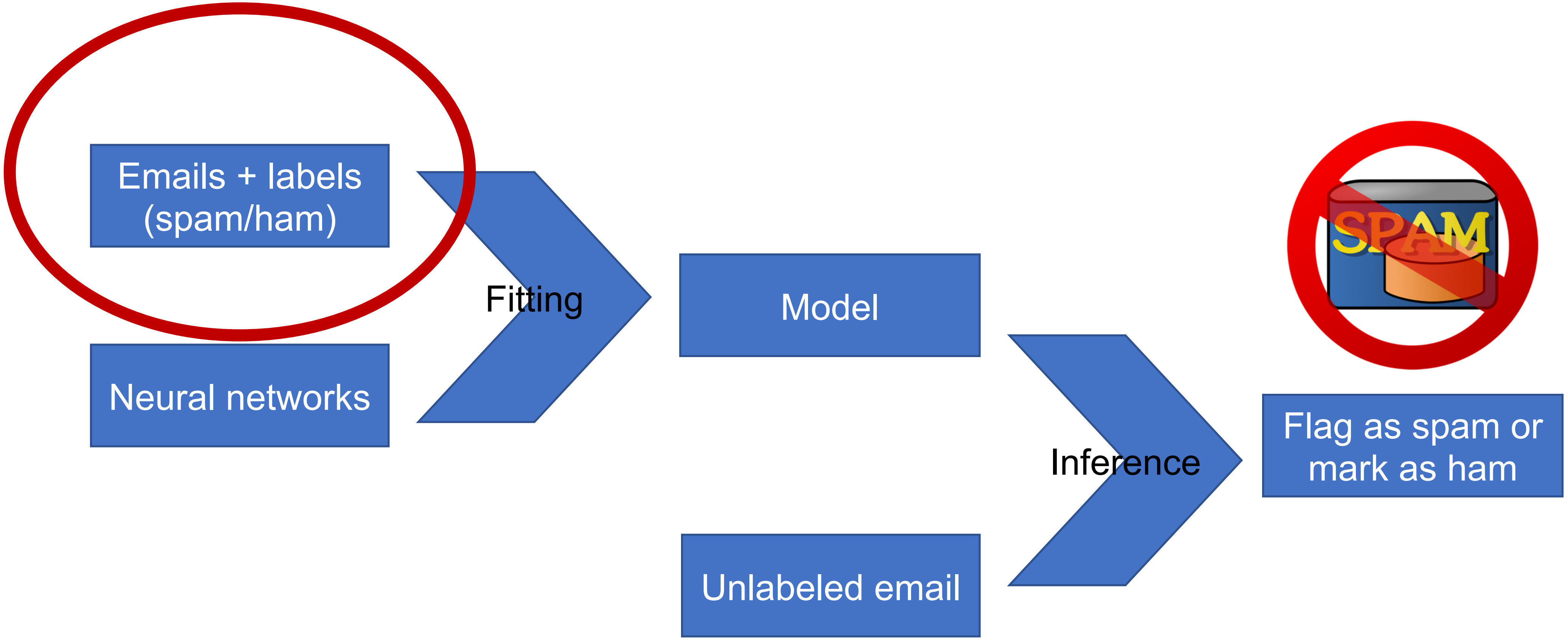
Machine learning paradigm



ML for spam detection



ML paradigm in adversarial settings



<u>Poisoning:</u> adversary inserts emails that contain spam but removes them from the spam folder back to inbox

ML paradigm in adversarial settings

Emails + labels (spam/ham)

Neural networks

Fitting

Model

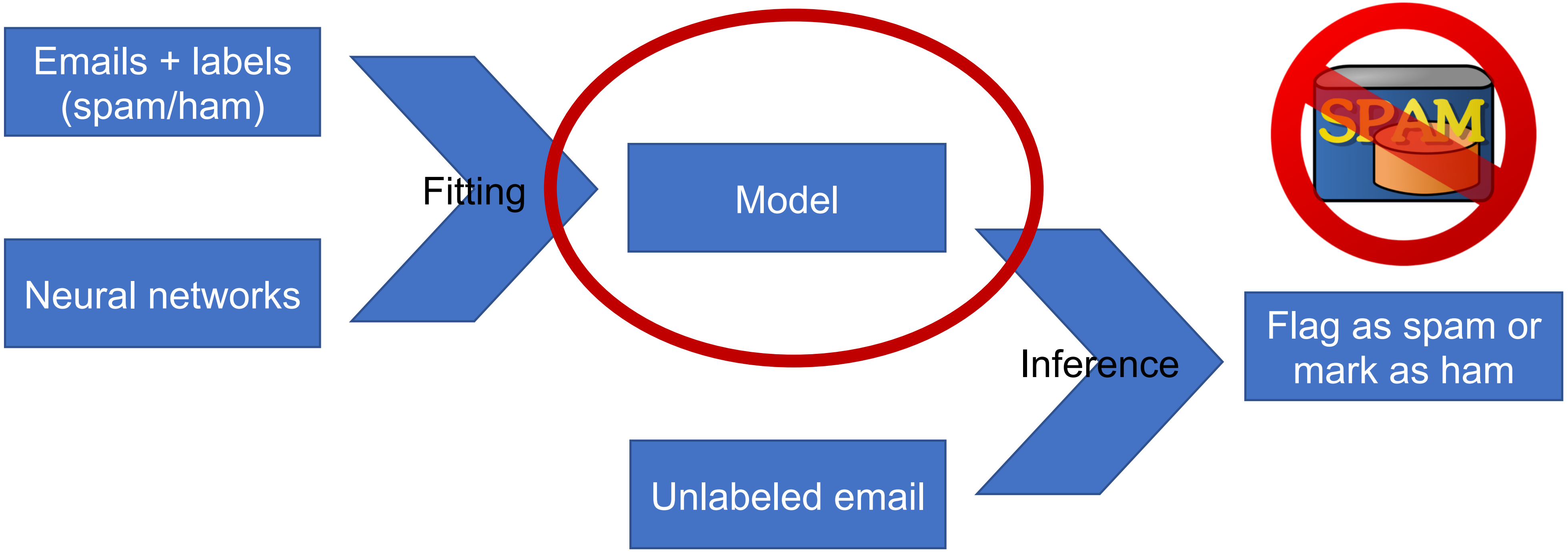
Unlabeled email

Inference

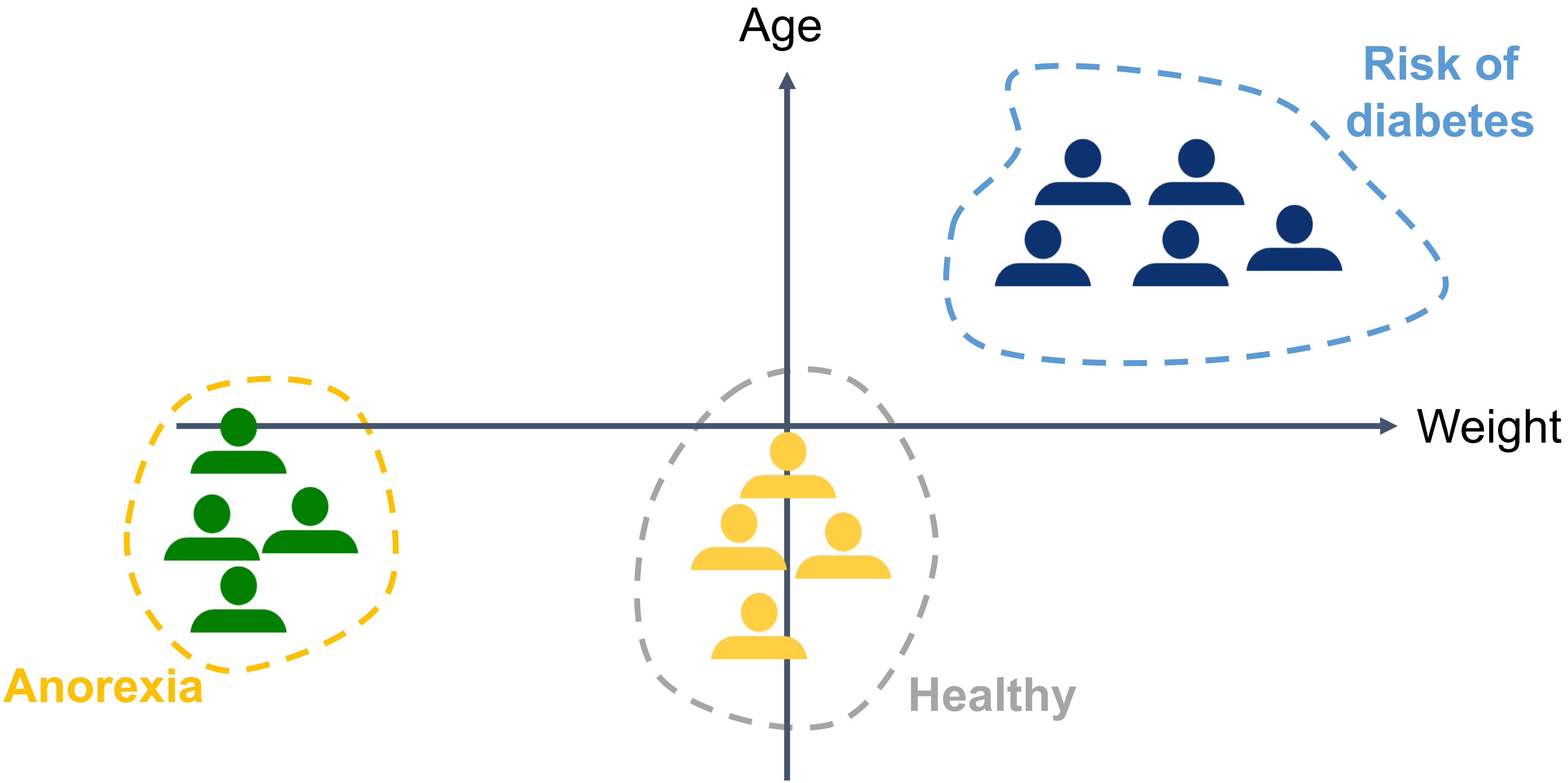
Flag as spam or mark as ham

Evasion: adversary crafts adversarial example that evades detection (spam email instantly marked as ham)

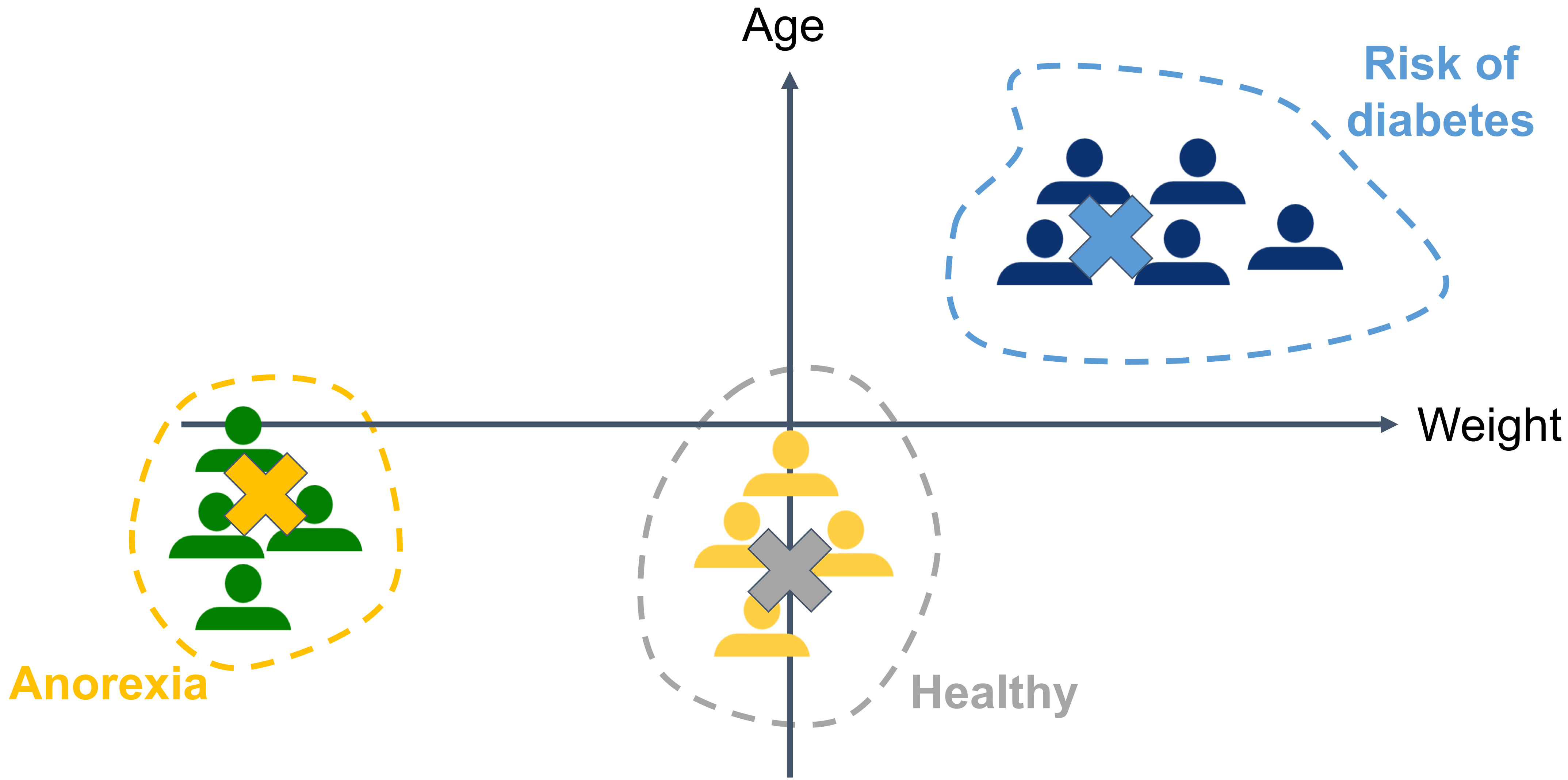
ML paradigm in adversarial settings

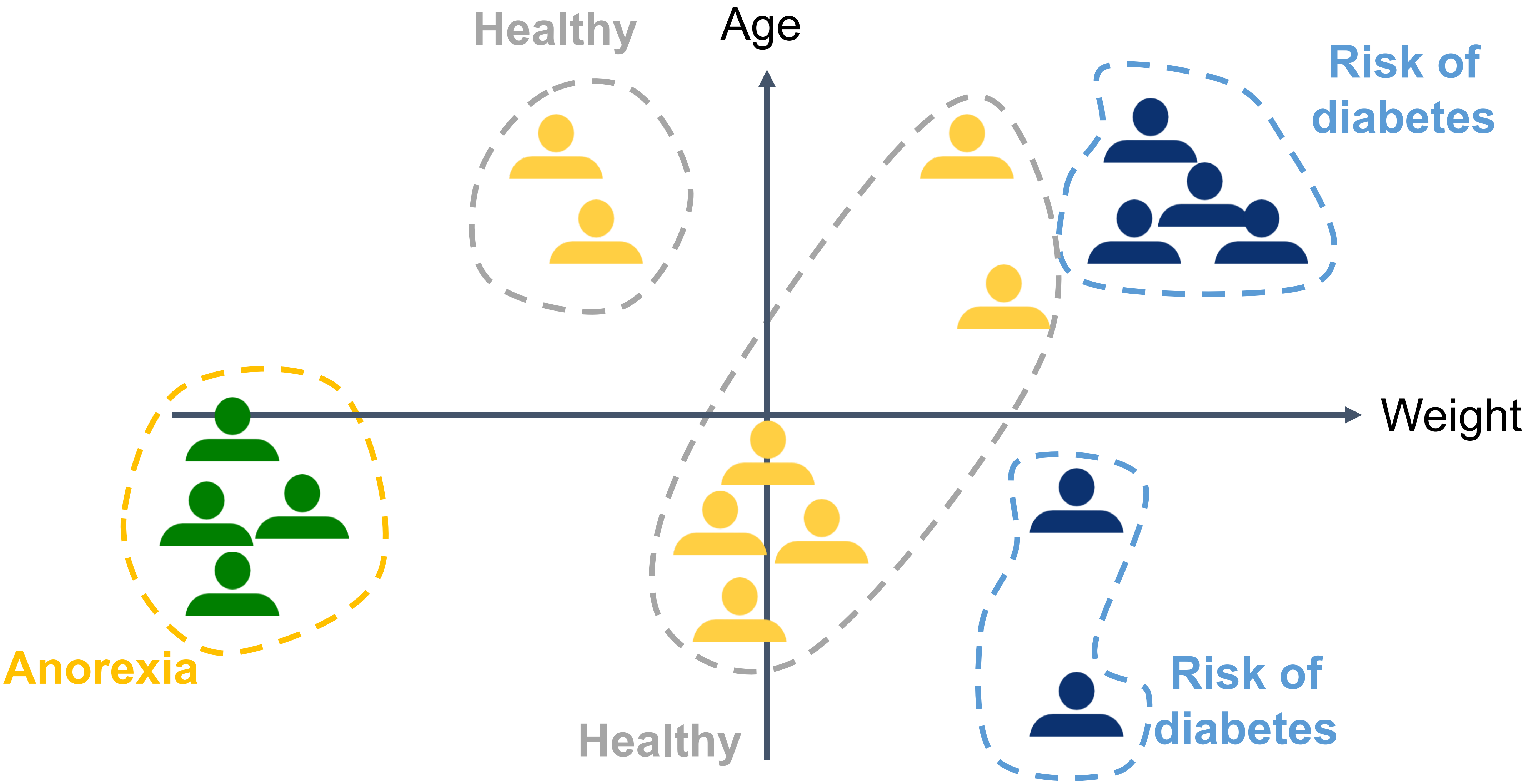


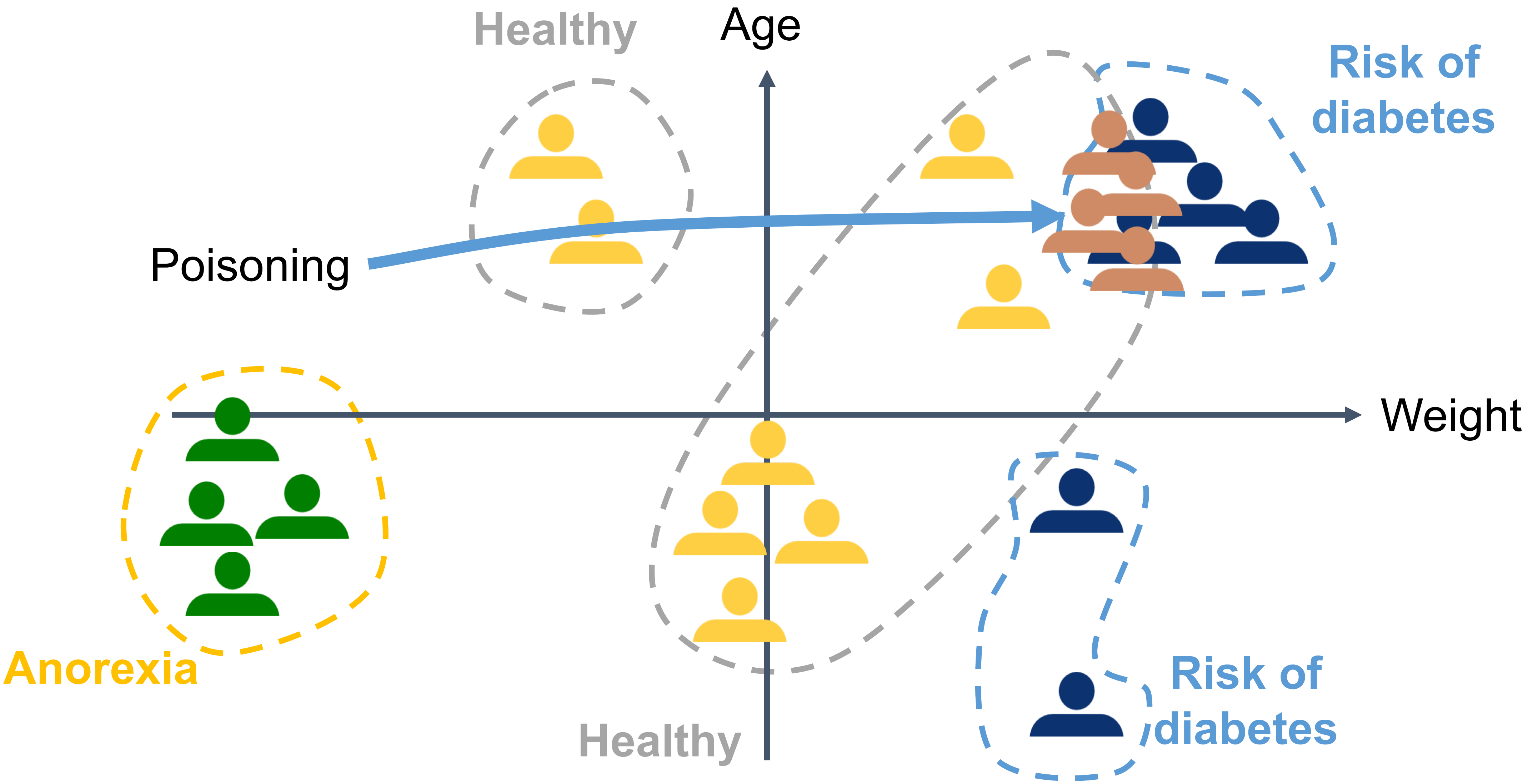
<u>Membership inference:</u> adversary inspects model to test whether an email was used to train it (privacy violation)



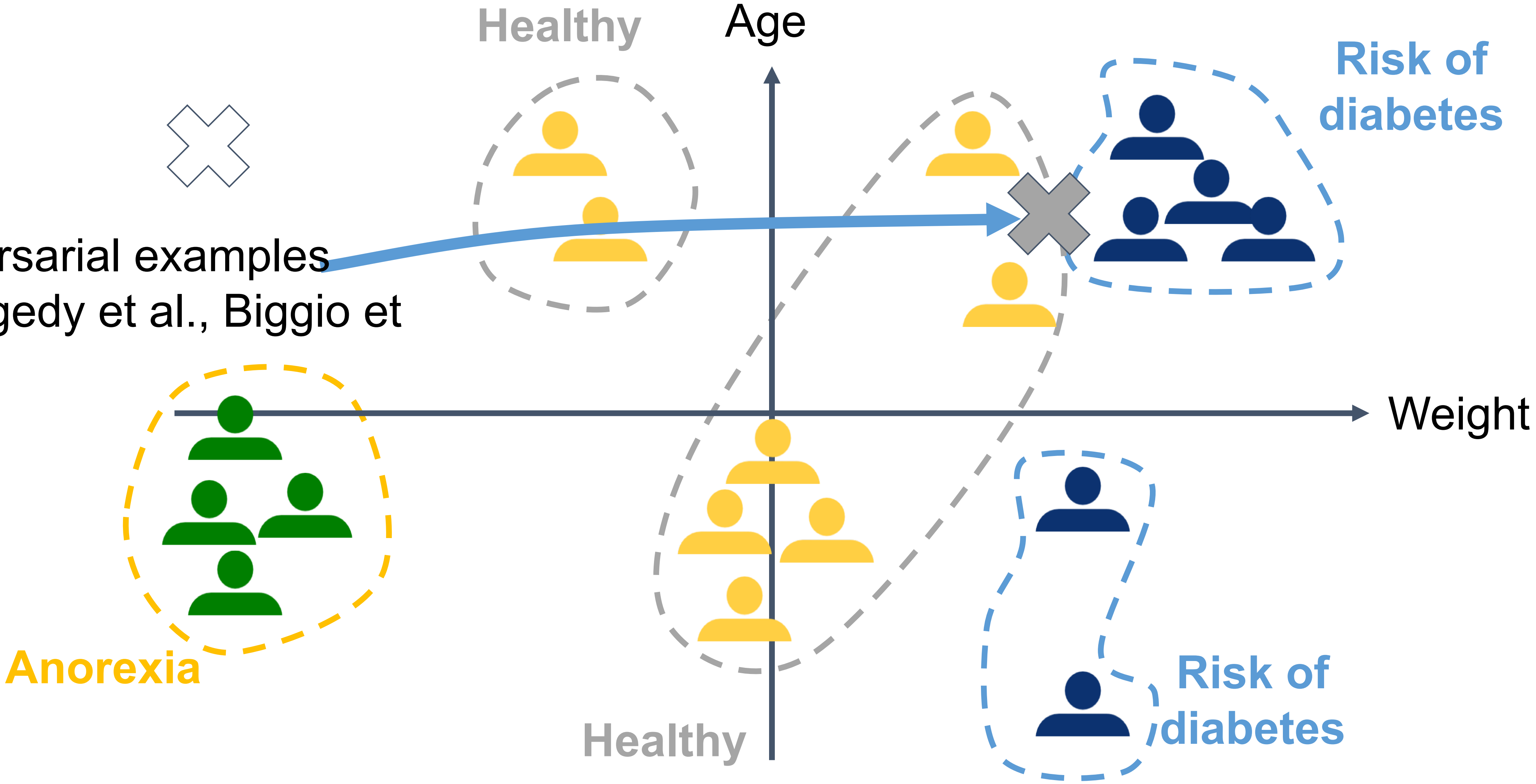
Age
Risk of
diabetes
Weight
Anorexia
Healthy

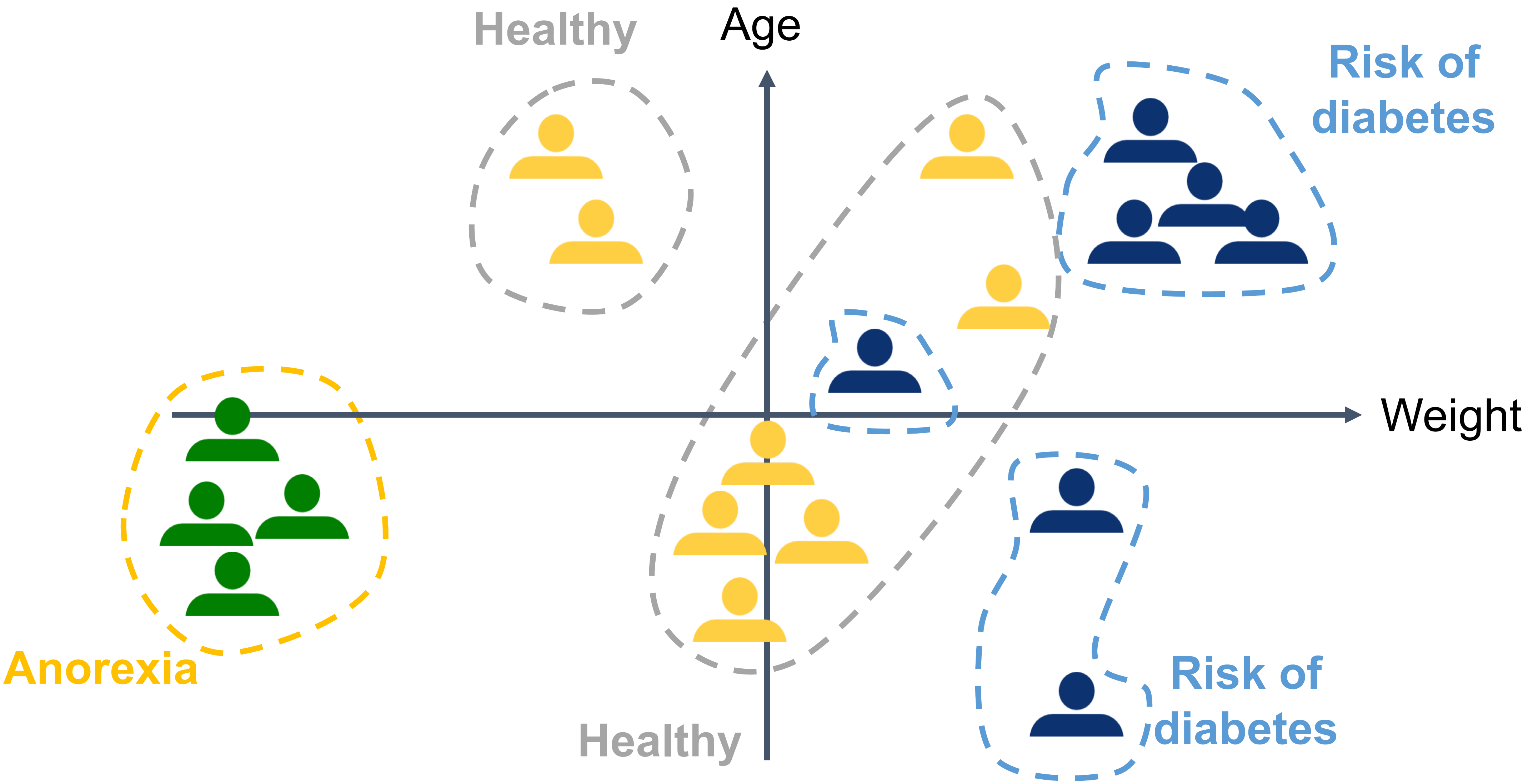

Age
Risk of diabetes
Weight
Anorexia
Healthy

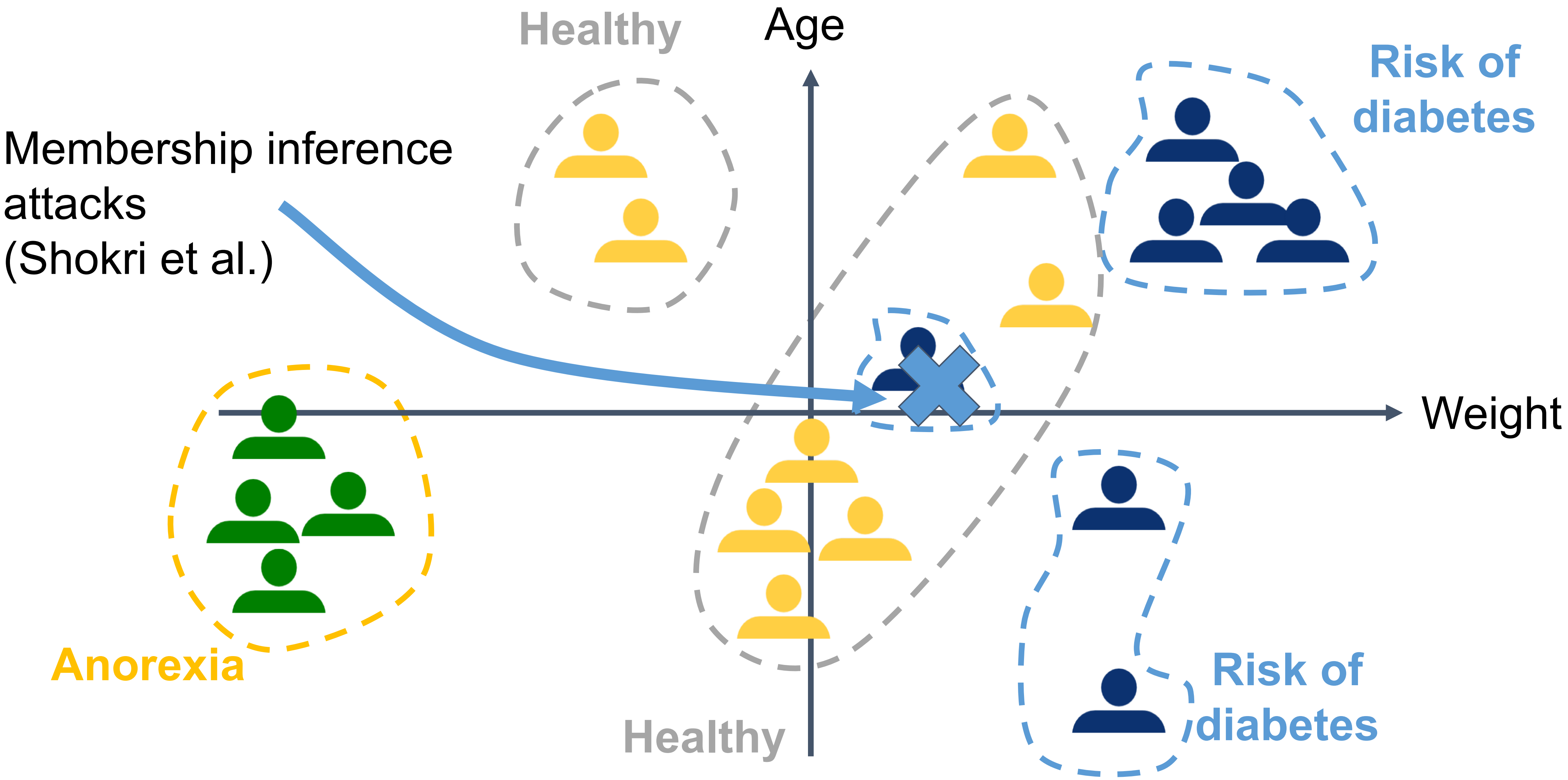

Healthy
Age
Risk of diabetes
Weight
Anorexia
Healthy
Risk of diabetes


Healthy
Age
Risk of diabetes
Poisoning
Weight
Anorexia
Healthy
Risk of diabetes

Adversarial examples (Szegedy et al., Biggio et al.)

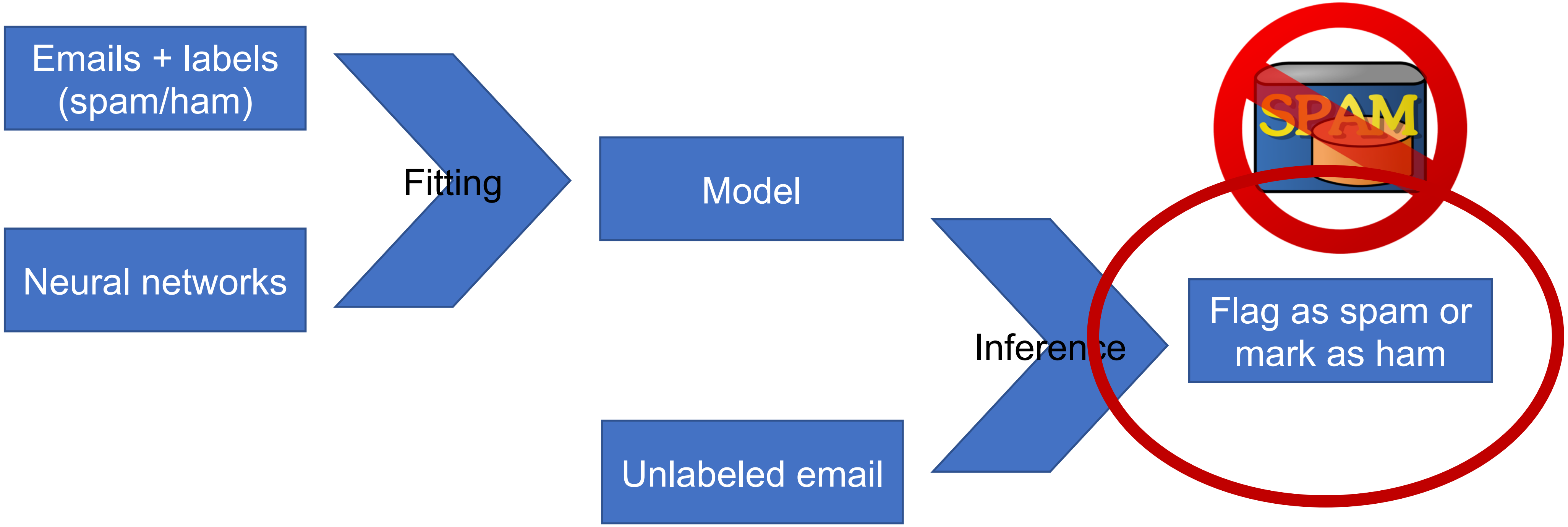



Healthy
Age
Risk of diabetes
Weight
Anorexia
Healthy
Risk of diabetes



Membership inference
attacks
(Shokri et al.)
Healthy
Age
Risk of
diabetes
Weight
Anorexia
Healthy
Risk of
diabetes

ML paradigm in adversarial settings



<u>Model extraction:</u> adversary observes predictions and reconstructs model locally

Societal aspects of the ML paradigm

Driving data

Neural networks

Fitting

Autonomous Driving Model

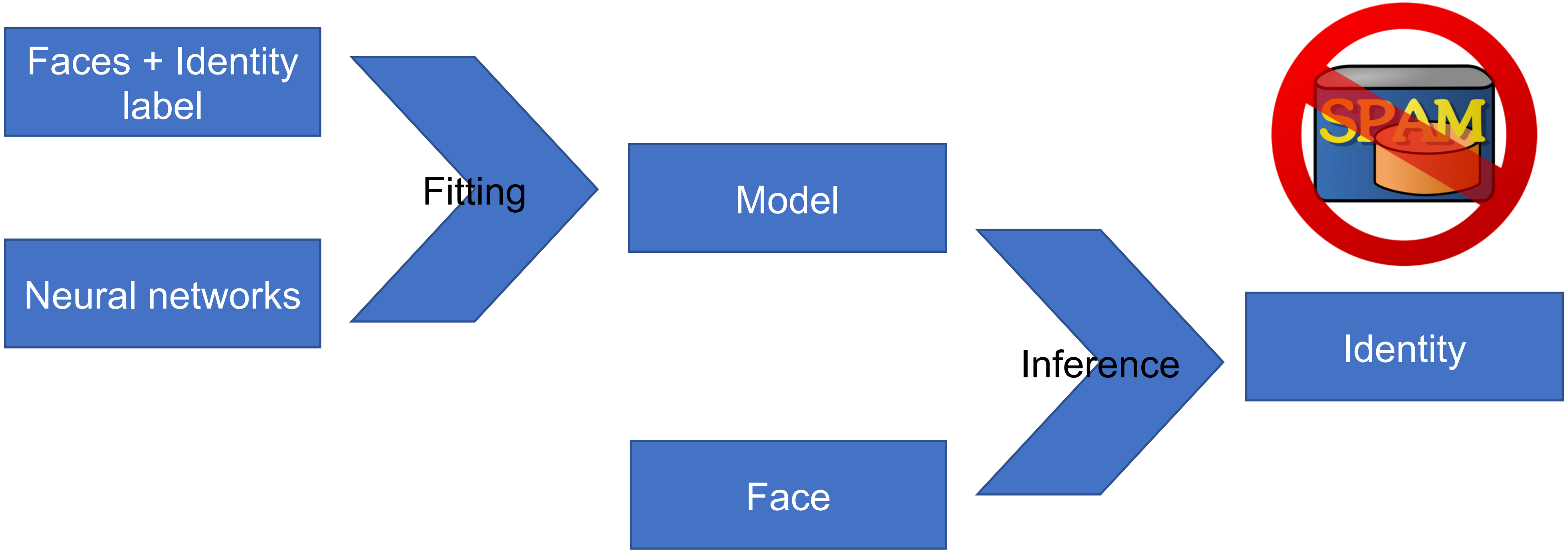
New road condition

Inference

Turn wheel

<u>Safety:</u> if training data is not comprehensive, driveless car may not take appropriate action

Societal aspects of the ML paradigm



<u>Fairness:</u> if training data does not contain enough faces from a minority, accuracy at inference suffers (model does not build relevant features)

#	Date	Topic	Reading / Assignment
1	Sep 09	Overview & motivation	•Reading: Saltzer and Schroeder. The Protection of Information in Computer Systems.
-	Sep 13	Exam	-
2	Sep 16	Training-time integrity (attacks & defenses)	•Reading: Nelson et al.. Exploiting Machine Learning to Subvert Your Spam Filter. •Rubinstein et al.. ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors. •Koh and Liang. Understanding Black-box Predictions via Influence Functions. •Jagielski et al.. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. •Shafahi et al.. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. •Wang et al.. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks.
3	Sep 23	Test-time integrity (attacks)	•Lowd and Meek. Good Word Attacks on Statistical Spam Filters. •Szegedy et al.. Intriguing properties of neural networks. •Biggio et al.. Evasion Attacks against Machine Learning at Test Time •Papernot et al.. Practical Black-Box Attacks against Machine Learning. •Xu et al.. Automatically Evading Classifiers. •Hong et al.. Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks.
4	Sep 30	Test-time integrity (defenses)	•Reading: Dalvi et al.. Adversarial Classification. •Tramer et al.. Ensemble Adversarial Training: Attacks and Defenses. •Wong and Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. •Lecuyer et al.. Certified Robustness to Adversarial Examples with Differential Privacy. •Geirhos et al.. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.
5	Oct 07	Confidentiality (of the model)	•Reading: Lowd and Meek. Adversarial Learning. •Tramer et al.. Stealing Machine Learning Models via Prediction APIs. •Chandrasekaran et al.. Model Extraction and Active Learning •Batina et al.. CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel. •Jagielski et al.. High-Fidelity Extraction of Neural Network Models.
	Oct 14	Thanksgiving	
6	Oct 21	Privacy attacks	•Reading: Narayanan and Shmatikov. Robust De-anonymization of Large Sparse Datasets. •Shokri et al.. Membership Inference Attacks against Machine Learning Models. •Carlini et al.. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. •Song and Shmatikov. Overlearning Reveals Sensitive Attributes.
7	Oct 28	Differential privacy	•Reading: Dwork et al.. Calibrating Noise to Sensitivity in Private Dat Analysis. •Abadi et al.. Deep Learning with Differential Privacy. •Papernot et al.. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data.
	Nov 04	Reading week	
8	Nov 11	Confidentiality (of the data)	•Reading: Ohrimenko et al.. Oblivious Multi-Party Machine Learning on Trusted Processors. •Dowlin et al.. CryptoNets: Applying Neural Networks to Encrypted Datawith High Throughput and Accuracy. •McMahan et al.. Communication-Efficient Learning of Deep Networks from Decentralized Data. •Nasr et al.. Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White[illegible]ence Attacks.
9	Nov 18	Safety	•Reading: Tsutomu Matsumoto. Impact of Artificial "Gummy" Fingers on Fingerprint Systems. •Amodei et al.. Concrete Problems in AI Safety. •Kurakin et al.. Adversarial Examples in the Physical World. •Gu et al.. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. •Pei et al.. DeepXplore: Automated Whitebox Testing of Deep Learning Systems.
10	Nov 25	Fairness & Ethics	•Reading: Dwork et al.. Fairness Through Awareness. •Caliskan et al.. Semantics derived automatically from language corpora contain human-like biases. •Speicher et al.. Potential for Discrimination in Online Targeted Advertising. •Kumar et al.. Law and Adversarial Machine Learning.
11	Dec 02	Poster session	

Security

Societal

Trustworthy

Saltzer and Schroeder’s principles

Economy of mechanism.
Keep the design of security mechanisms simple.

Fail-safe defaults.
Base access decisions on permission rather than exclusion.

Complete mediation.
Every access to an object is checked for authority.

Open design.
The design of security mechanisms should not be secret.

Separation of privilege.
A protection mechanism that requires two keys to unlock is more robust and flexible.

Least privilege.
Every user operates with least privileges necessary.

Least common mechanism.
Minimize mechanisms depended on by all users.

Psychological acceptability.
Human interface designed for ease of use.

Work factor.
Balance cost of circumventing the mechanism with known attacker resources.

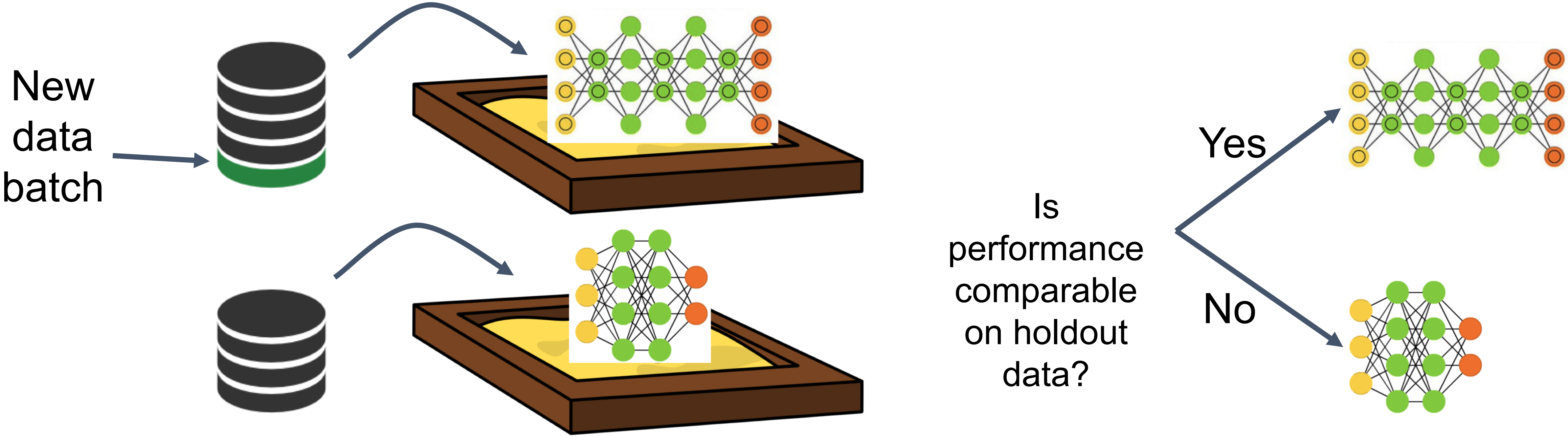
Compromise recording.
Mechanisms that reliably record compromises can be used in place of mechanisms that prevent loss.

Fail-safe defaults

Example 1: do not output low-confidence predictions at test time

Example 2: mitigate data poisoning resulting in a distribution drift

Attacker: submits poisoned points to gradually change a model's decision boundary
Defender: compares accuracy on holdout validation set ***before*** applying gradients

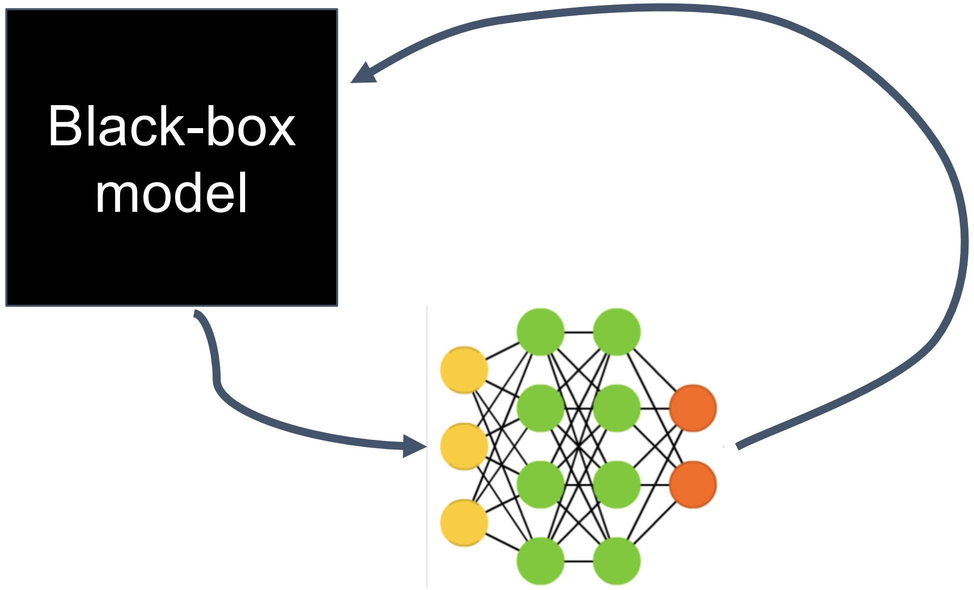


Open design

Example 1: black-box attacks are not particularly more difficult than white-box attacks

Insider leaks model

Reverse engineering

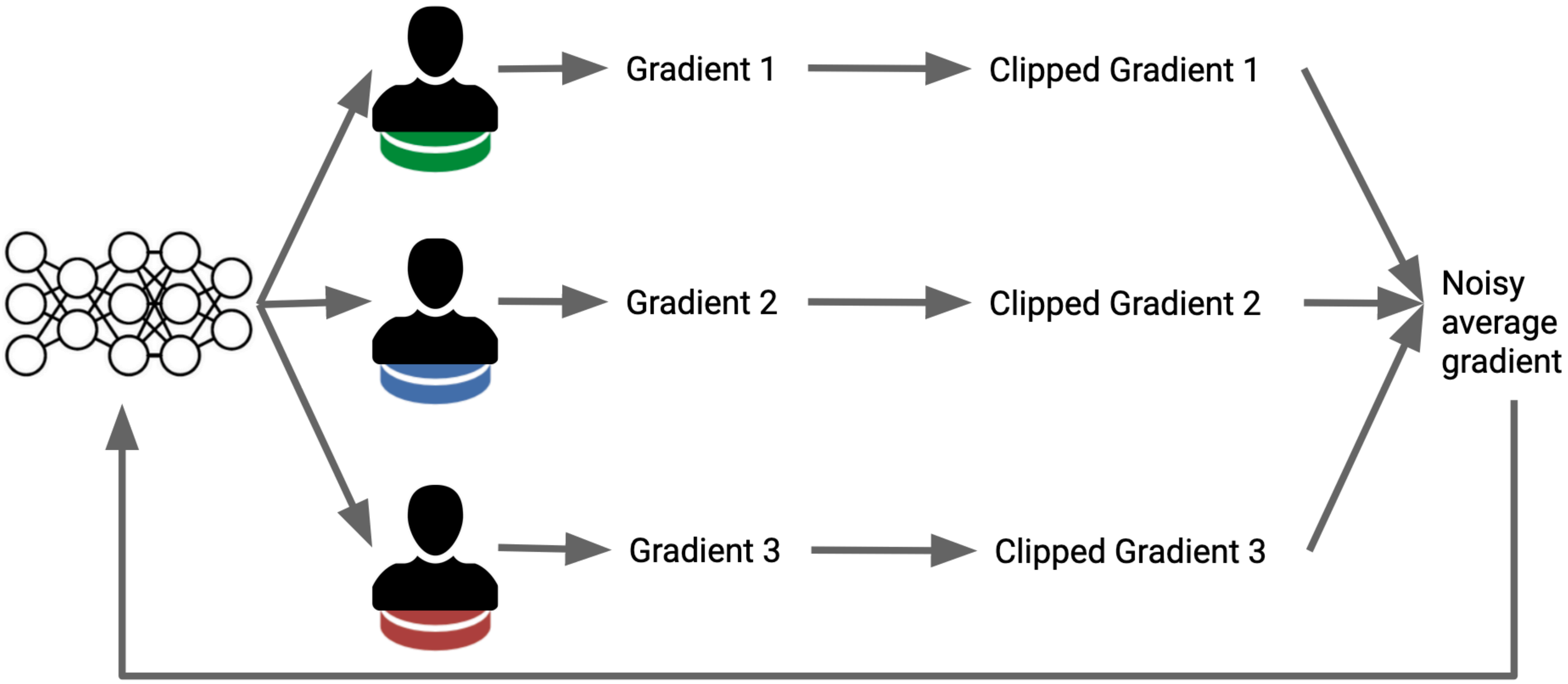


Model extraction

Transferability

ACM:2650798 (Šrndic and Laskov); arXiv:1602.02697 (Papernot et al.)

Separation of privilege



Saltzer and Schroeder’s principles

Economy of mechanism.
Keep the design of security mechanisms simple.

Fail-safe defaults.
Base access decisions on permission rather than exclusion.

Complete mediation.
Every access to an object is checked for authority.

Open design.
The design of security mechanisms should not be secret.

Separation of privilege.
A protection mechanism that requires two keys to unlock is more robust and flexible.

Least privilege.
Every user operates with least privileges necessary.

Least common mechanism.
Minimize mechanisms depended on by all users.

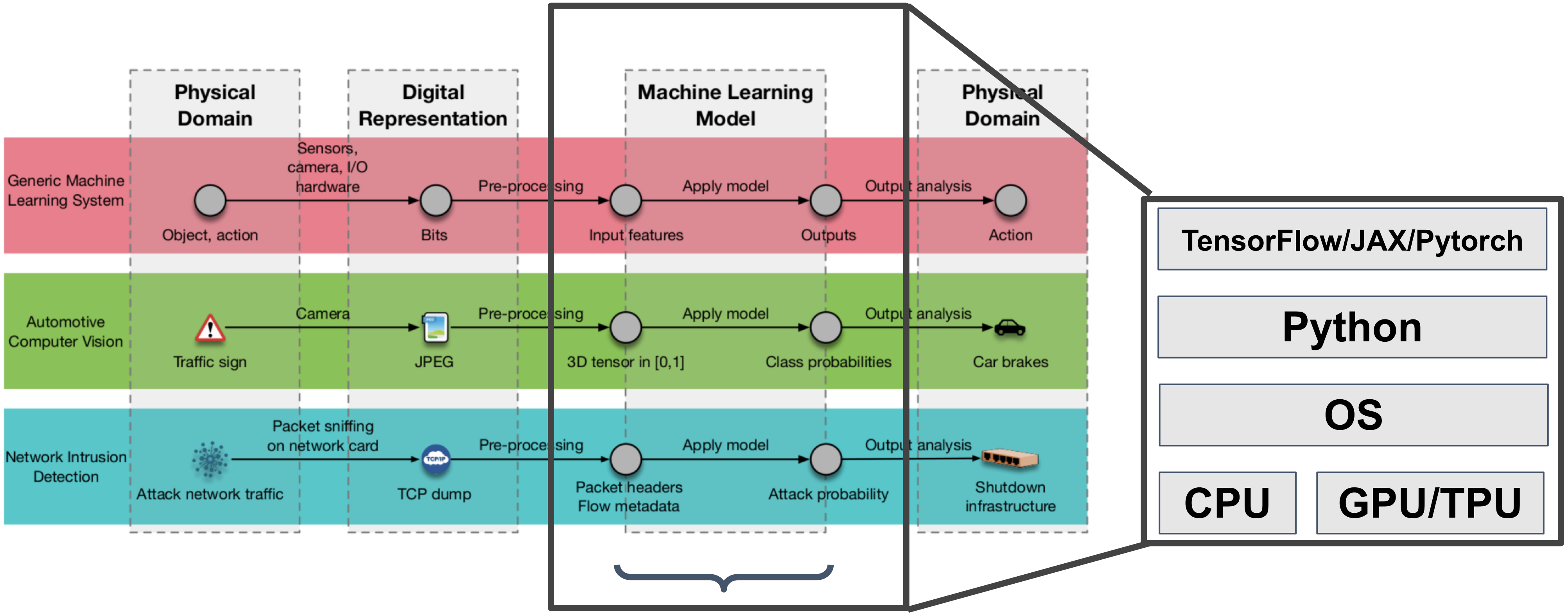
Psychological acceptability.
Human interface designed for ease of use.

Work factor.
Balance cost of circumventing the mechanism with known attacker resources.

Compromise recording.
Mechanisms that reliably record compromises can be used in place of mechanisms that prevent loss.

https://www.cs.virginia.edu/~evans/cs551/saltzer/

Trusted Computing Base?



Research project

- 30mn in class each week, plus work outside class
- Take a look at topics and papers covered in the syllabus
- Identify two areas of interest
- Formulate a project proposal (1/2 page, due by next week)
  - Proposed title
  - Proposed team (optional)
  - Proposed problem
  - Proposed methodology (optional)
  - Alternative topic you would be interested in
- I will help form teams after September 23rd if you do not already have teammates by then

- Exam (30% of grade):
  - Friday September 13 from 4PM to 5PM
  - BA1170

- [ASSIGNMENT] 1 page summary of all papers assigned for reading is due at the beginning of each class (bring a physical copy)
- [ASSIGNMENT] Project proposal (1/2 page)
- I will reach out to a group of students to prepare the presentation for next class.

- Syllabus: papernot.fr/teaching/f19-trustworthy-ml
- Office hours: Wednesdays 1.30-3.30pm (Pratt 484E)
- Email: nicolas.papernot@utoronto.ca

- EXIT FORM
  - Write down name + 2 things you hope to learn this semester