UNIVERSITY OF TORONTO

VECTOR INSTITUTE

ECE1784H/CSC2559H: Trustworthy Machine Learning

Prof. Nicolas Papernot

nicolas.papernot@utoronto.ca

Land acknowledgment

We wish to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and most recently, the Mississaugas of the Credit River. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

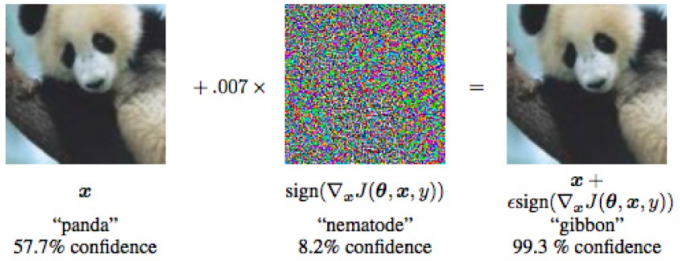
Logistics

- Course syllabus: papernot.fr/teaching/f21-trustworthy-ml
  - Schedule
  - Assigned reading
  - Assignment description
  - Grading information
  - Ethics statement
- Class: Tuesdays 3-5pm
- Office hours: Tuesdays 5-6pm (Zoom)

What is this class?

This is not a ML course

What do I mean by trustworthy ML?

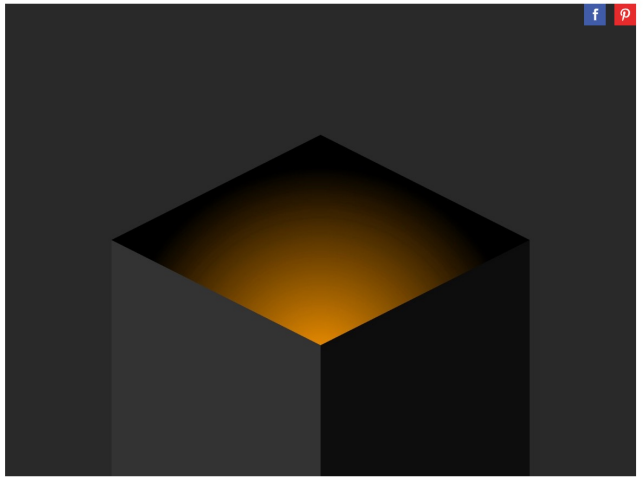


Security

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.

stochastic convex optimization, MSE, PCA, SGD, L-BFGS, TPU, label smoothing, distillation, semi-supervised learning, embeddings, ResNet, BERT, Transformer, central limit theorem, SVM, dropout, computation graph, non-IID, regularization, CNN, Newton step, generalization, expressivity

Format for weeks 2-10

- High level:
  - Research papers
  - One team will present and lead the discussion
  - Interactive discussion (everyone should do the reading ahead of class)

- 10mn: introduction to week theme
- 40mn: presentation on papers
- 40mn: discussion
- 20mn: teaching team available to provide guidance on research projects

Timeline

- d-14 (Tuesday): presenter team hands in draft of slides
- d-7 (Tuesday): slides are released to class, all non-presenting students comment on slides while reading papers
- d-1 (Monday): non-presenting students submitted discussion questions
- d (Tuesday): presenter team lectures, everyone participates in discussion
- d+3 (Friday): presenter team submits final slide deck

During class: discussion

- All: ask questions
- Presenting team:
  - May choose an appropriate format
    - Slides
    - interactive demos
    - code tutorials
  - Should involve class
  - Should cover (at least) the papers assigned for reading

Rubrics

- See syllabus. For presentation:
- Technical:
  - Depth of content
  - Accuracy of content
  - Paper criticism
  - Discussion lead
- Soft presentation skills:
  - Time management
  - Responsiveness to audience
  - Organization
  - Presentation aids

Lateness policy

- Slide deck commenting and questions submissions assigned each week will not be accepted late
- All other assignments (i.e., presentation slides and project reports) will be assessed
  - a 10% per-day late penalty
  - up to a maximum of 2 days
- Students with legitimate reasons who contact the professor before the deadline may apply for an extension.

Grading scheme

- 15% weekly reading questions
- 20% participation (slide deck commenting and in class discussion)
- 30% paper presentation
- 35% research project

Research project

- Teaching team available at end of class each week
- Take a look at topics and papers covered in the syllabus
- Identify two areas of interest
- Formulate a project proposal and discuss with us ahead of Oct 8
  - Proposed title
  - Proposed team (optional)
  - Proposed problem
  - Proposed methodology (optional)
  - Alternative topic you would be interested in
- If you do not find teammates within 1-2 weeks, let us know on piazza

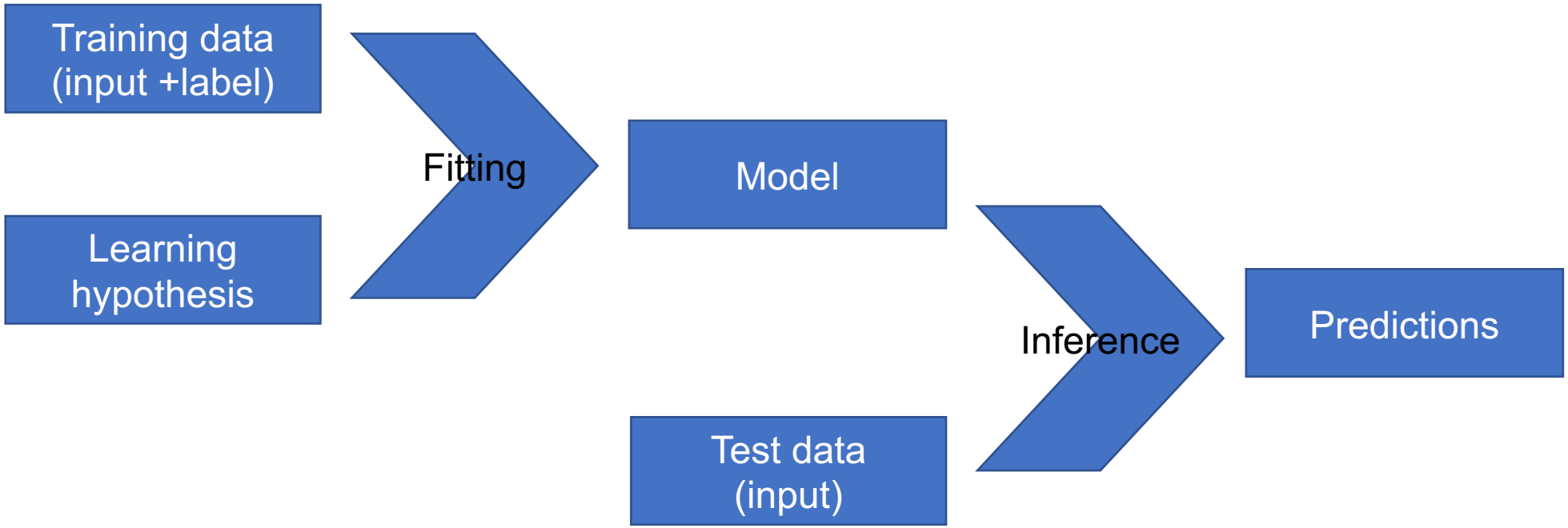
Integrity

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.

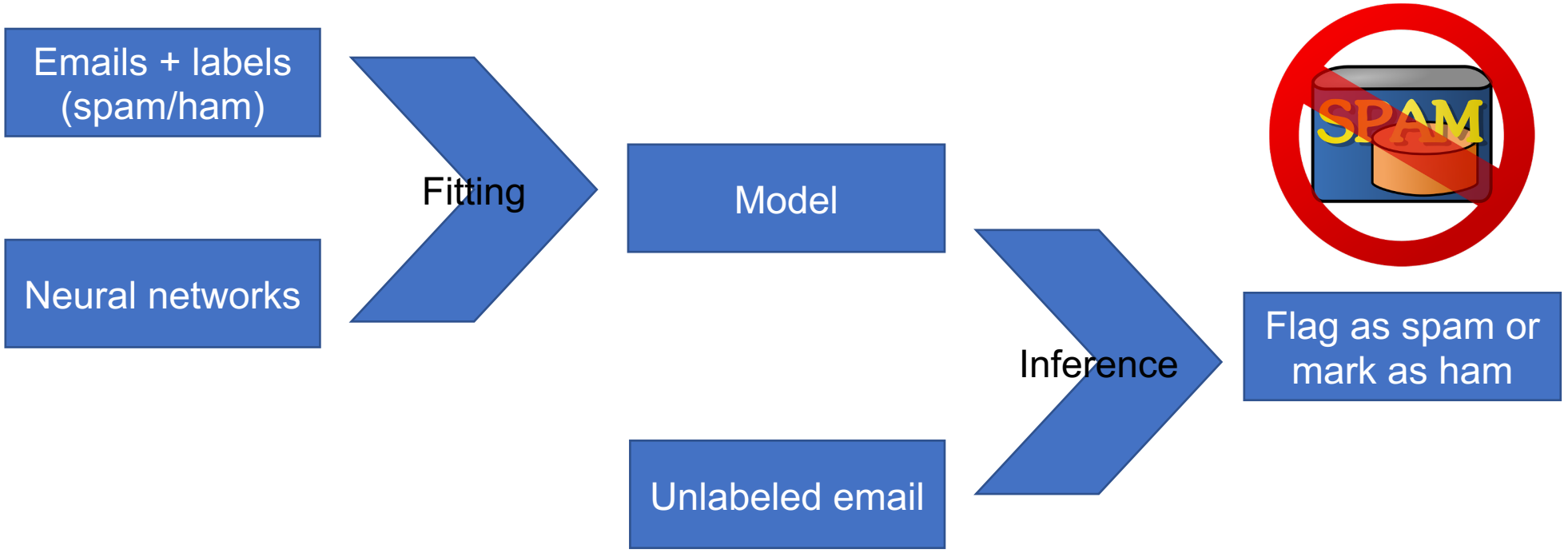
Ethics

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.

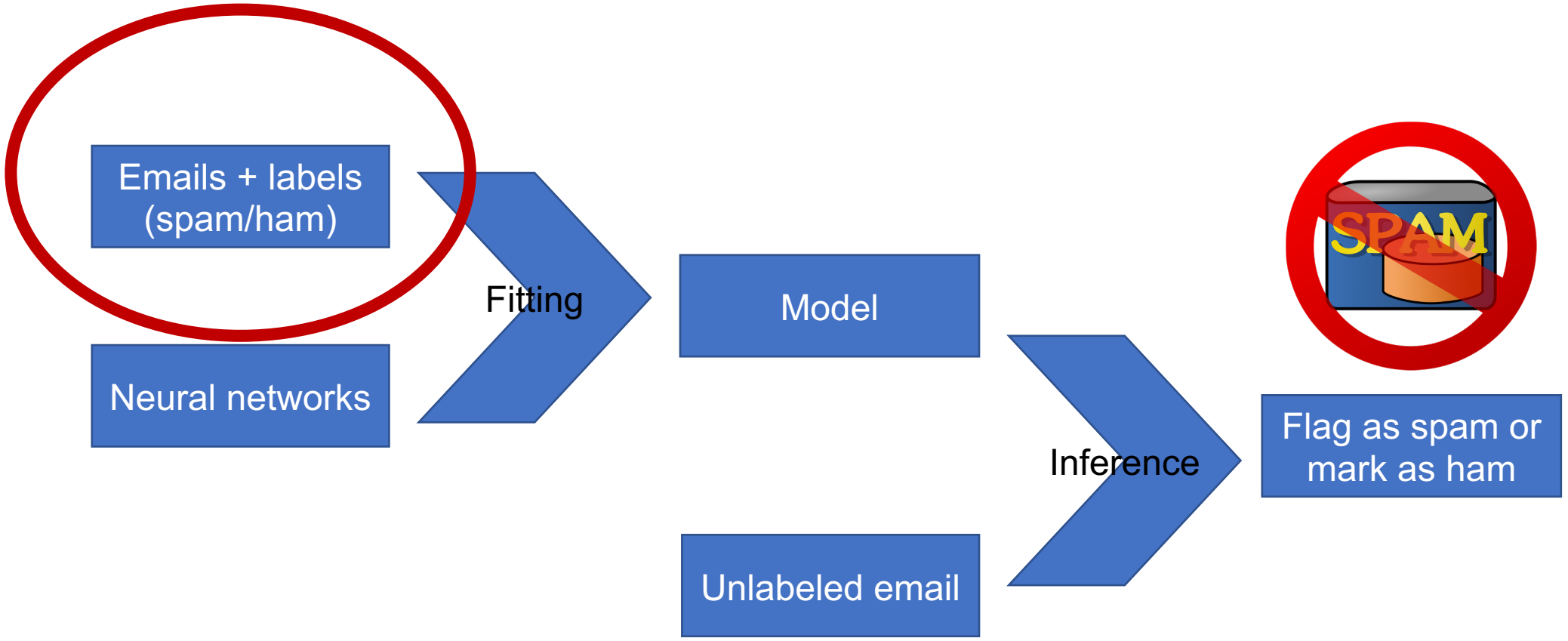
Machine learning paradigm



ML for spam detection

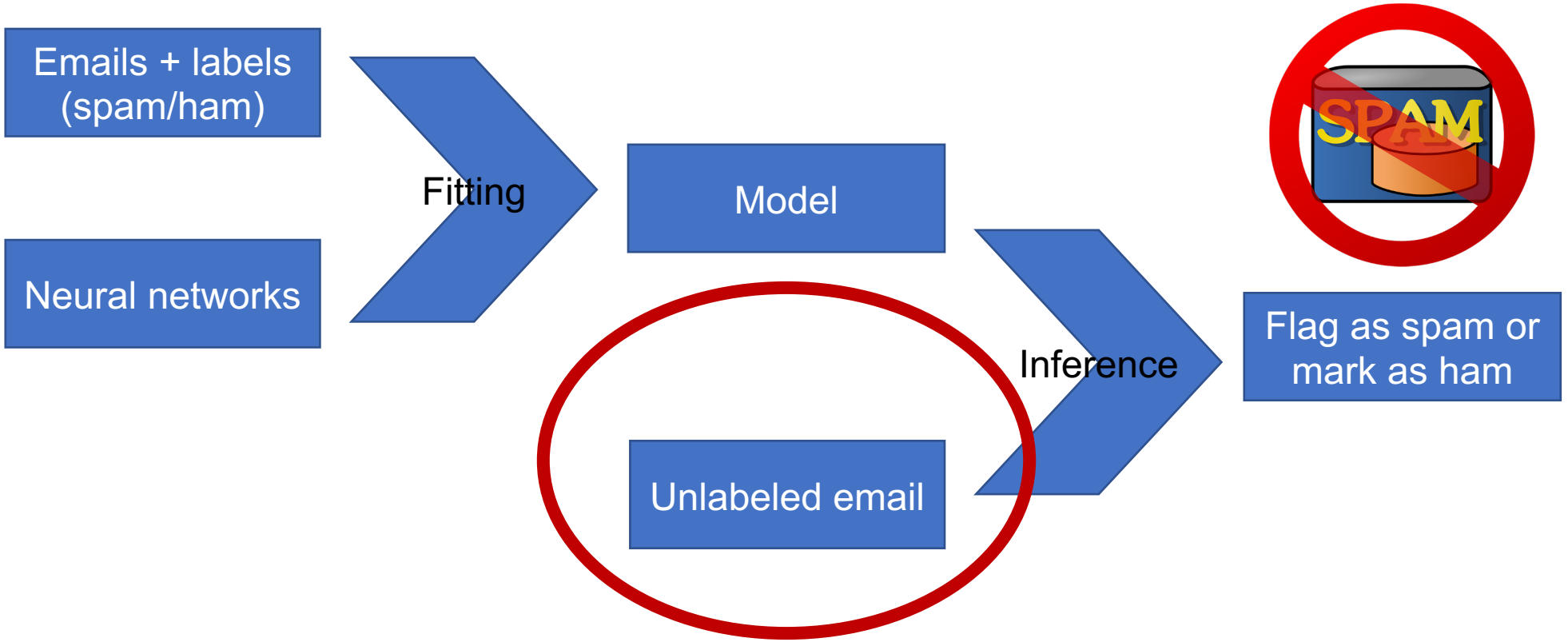


ML paradigm in adversarial settings



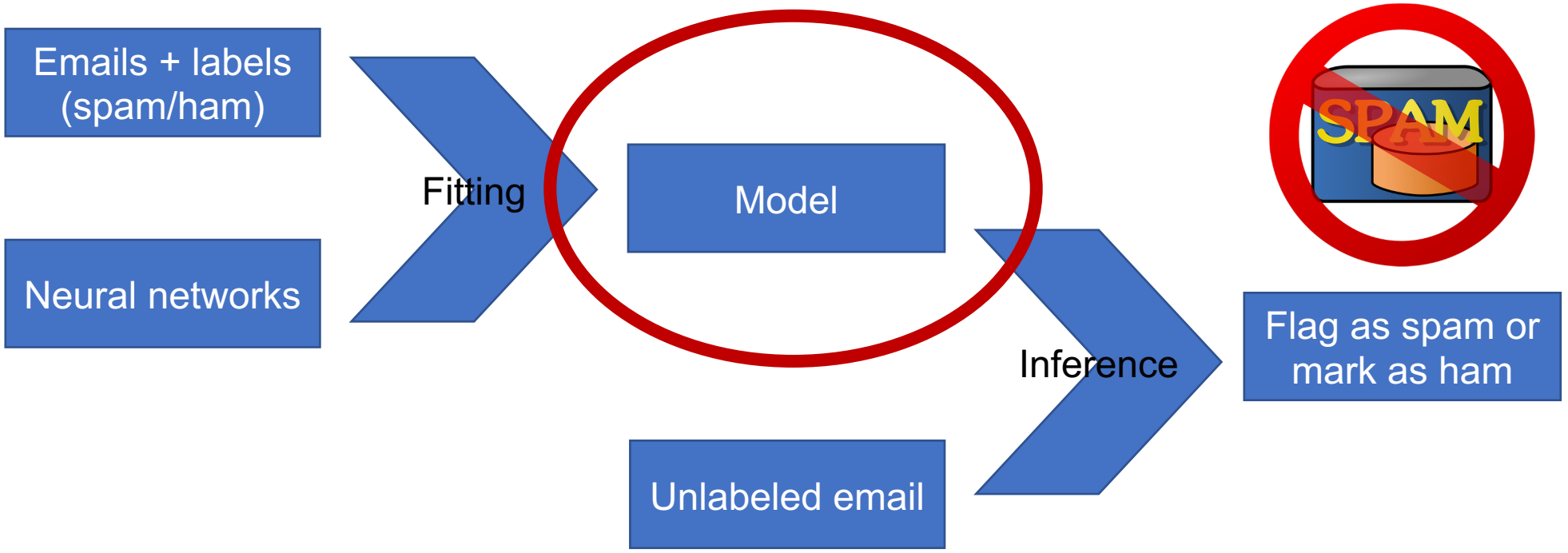
Poisoning: adversary inserts emails that contain spam but removes them from the spam folder back to inbox

ML paradigm in adversarial settings

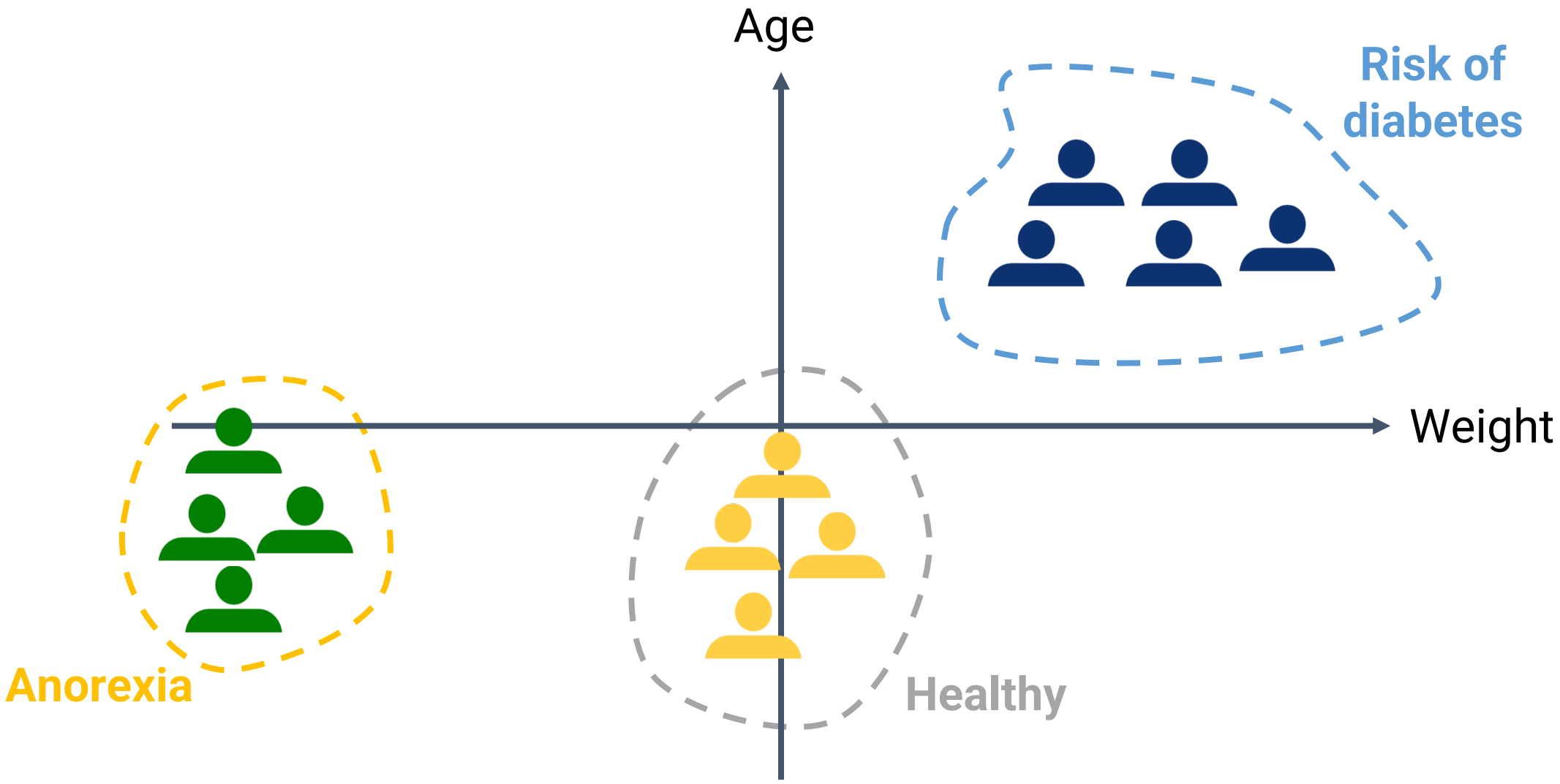


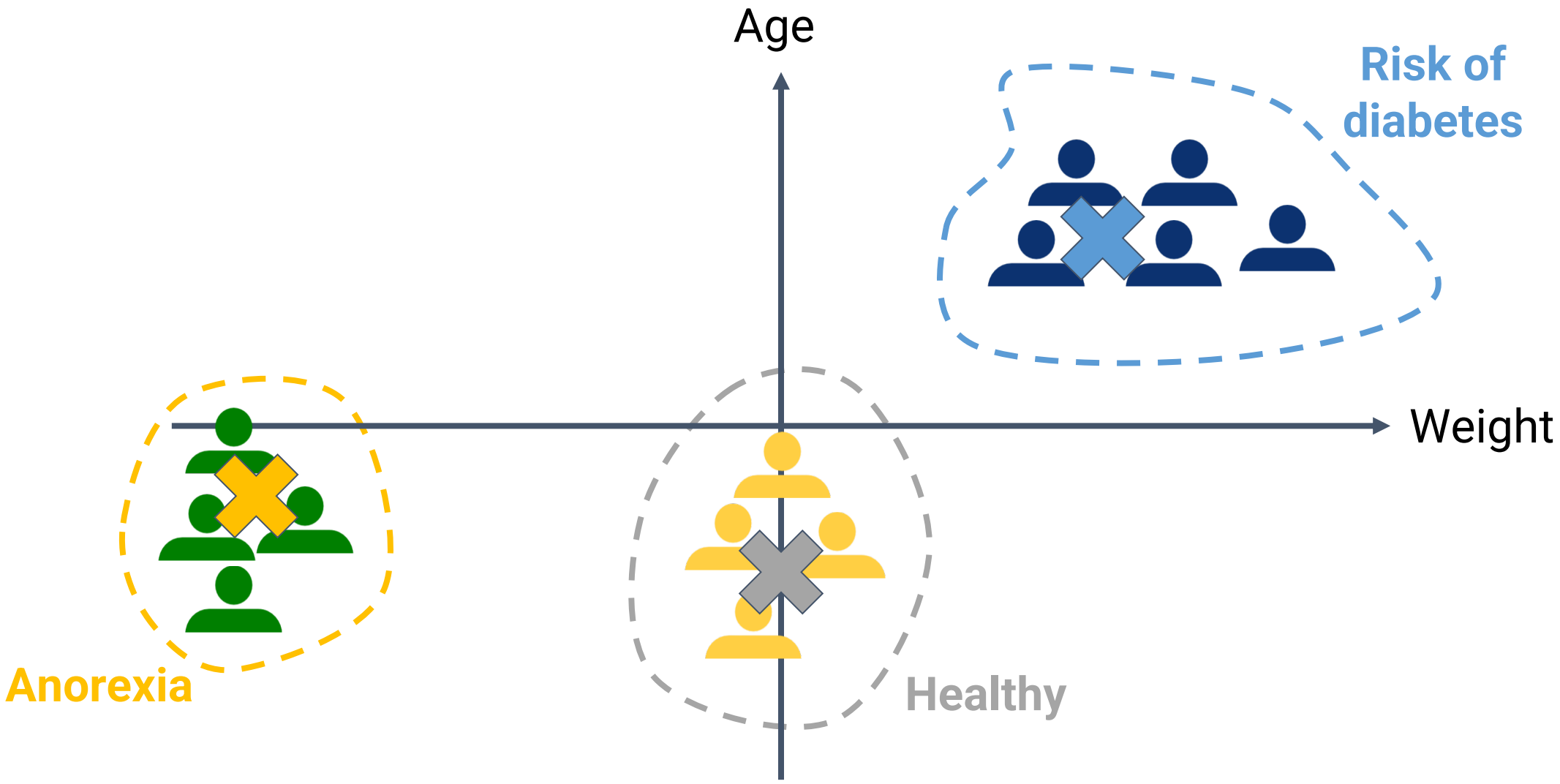
Evasion: adversary crafts adversarial example that evades detection (spam email instantly marked as ham)

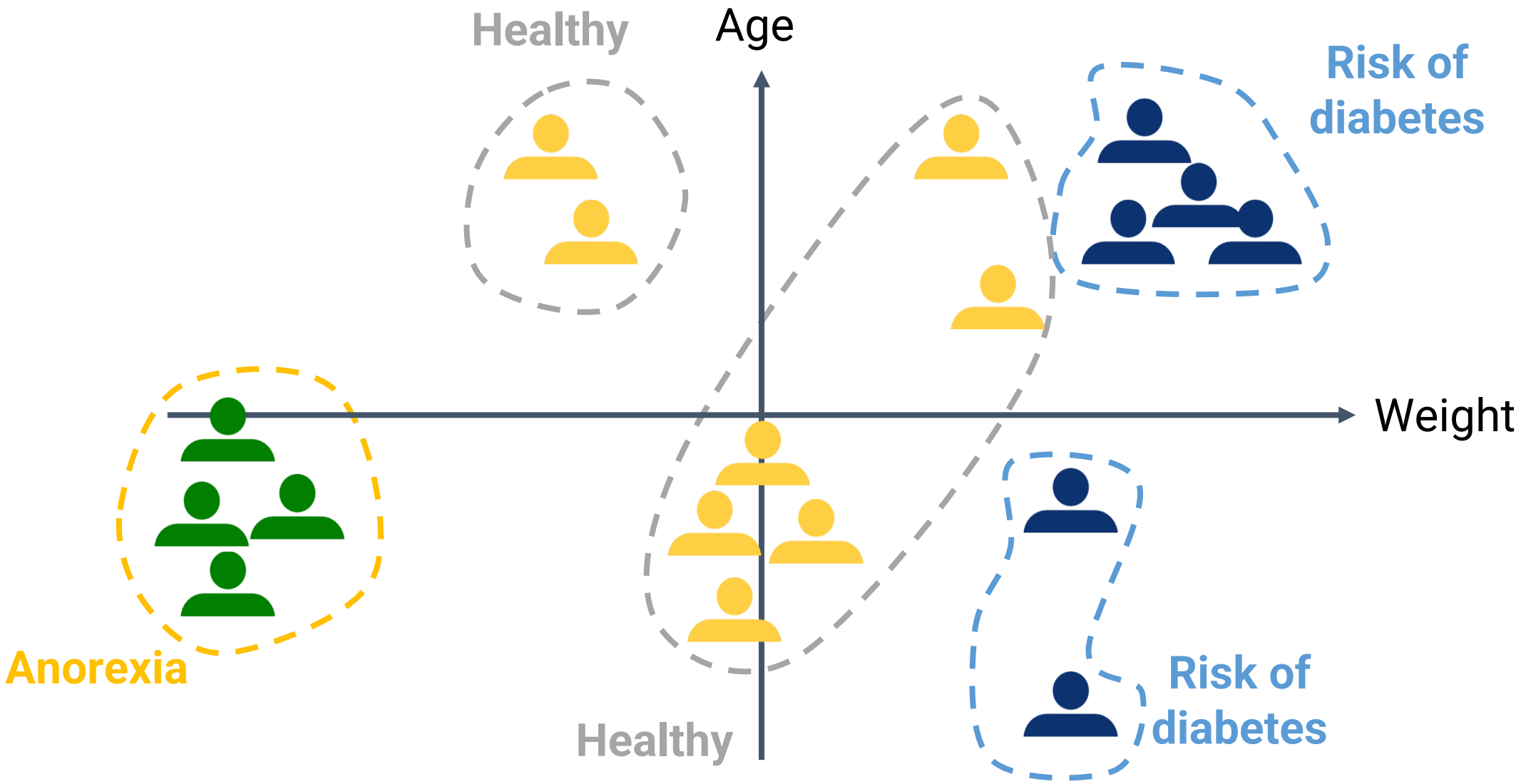
ML paradigm in adversarial settings

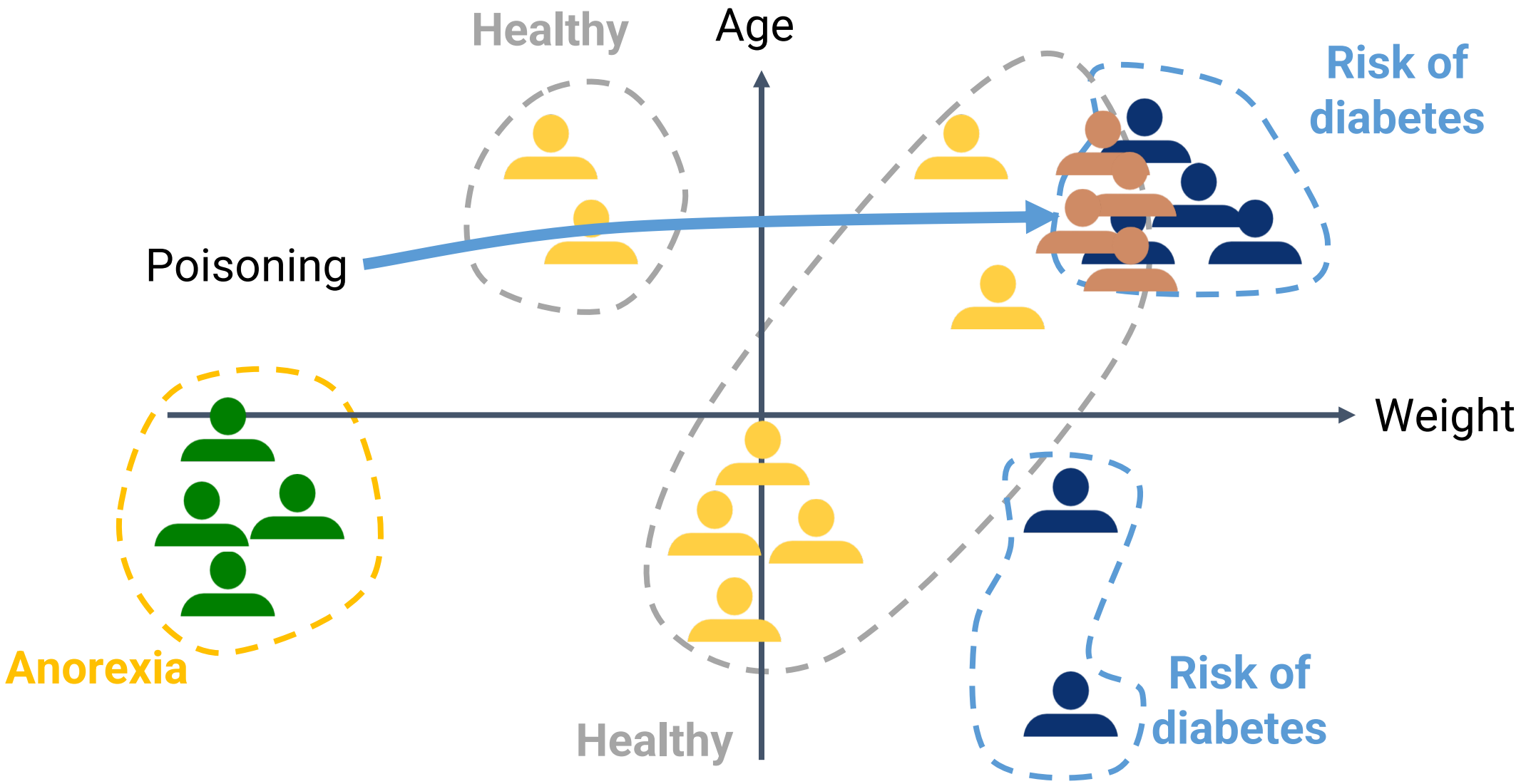


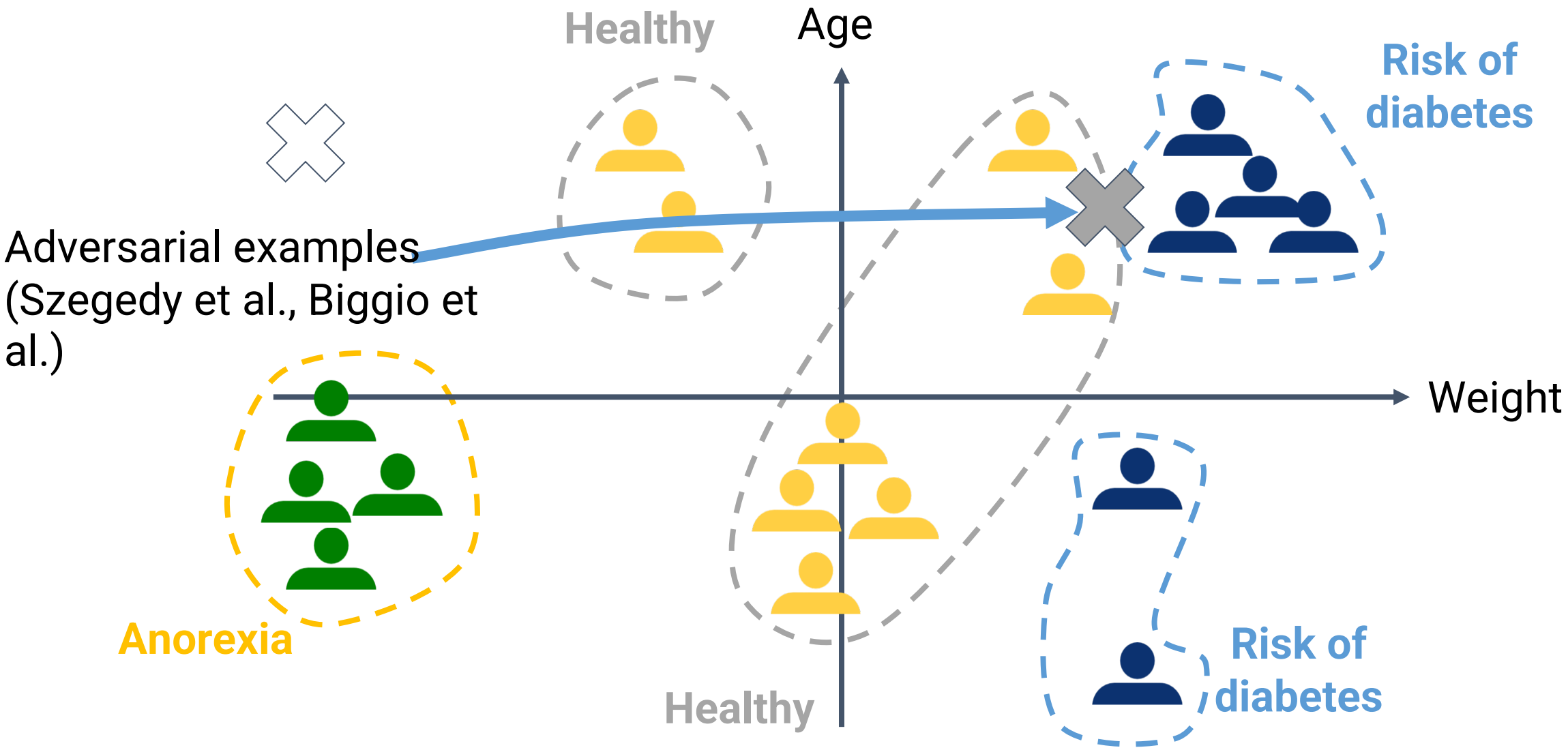
Membership inference: adversary inspects model to test whether an email was used to train it (privacy violation)

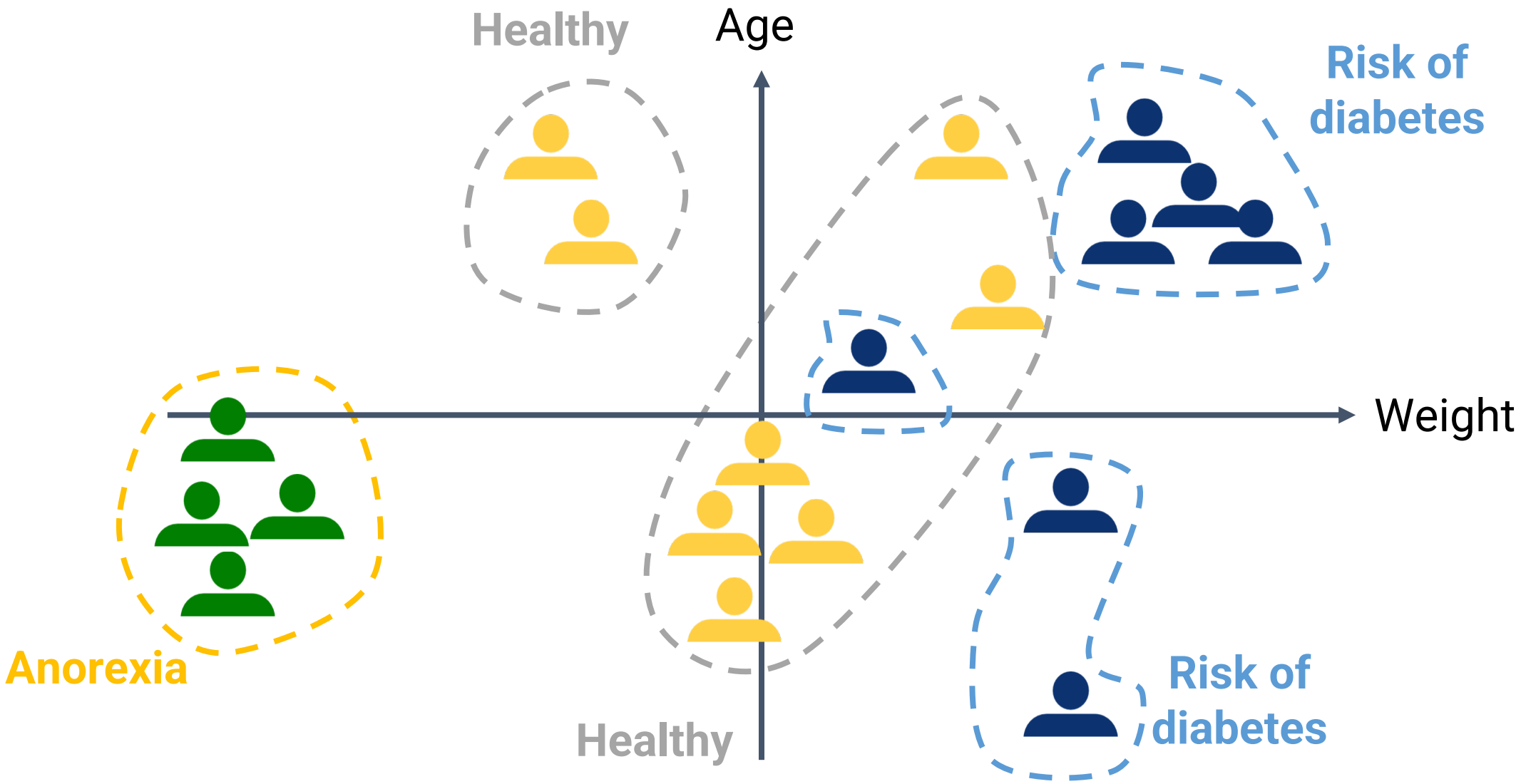

Age
Weight
Risk of diabetes
Anorexia
Healthy

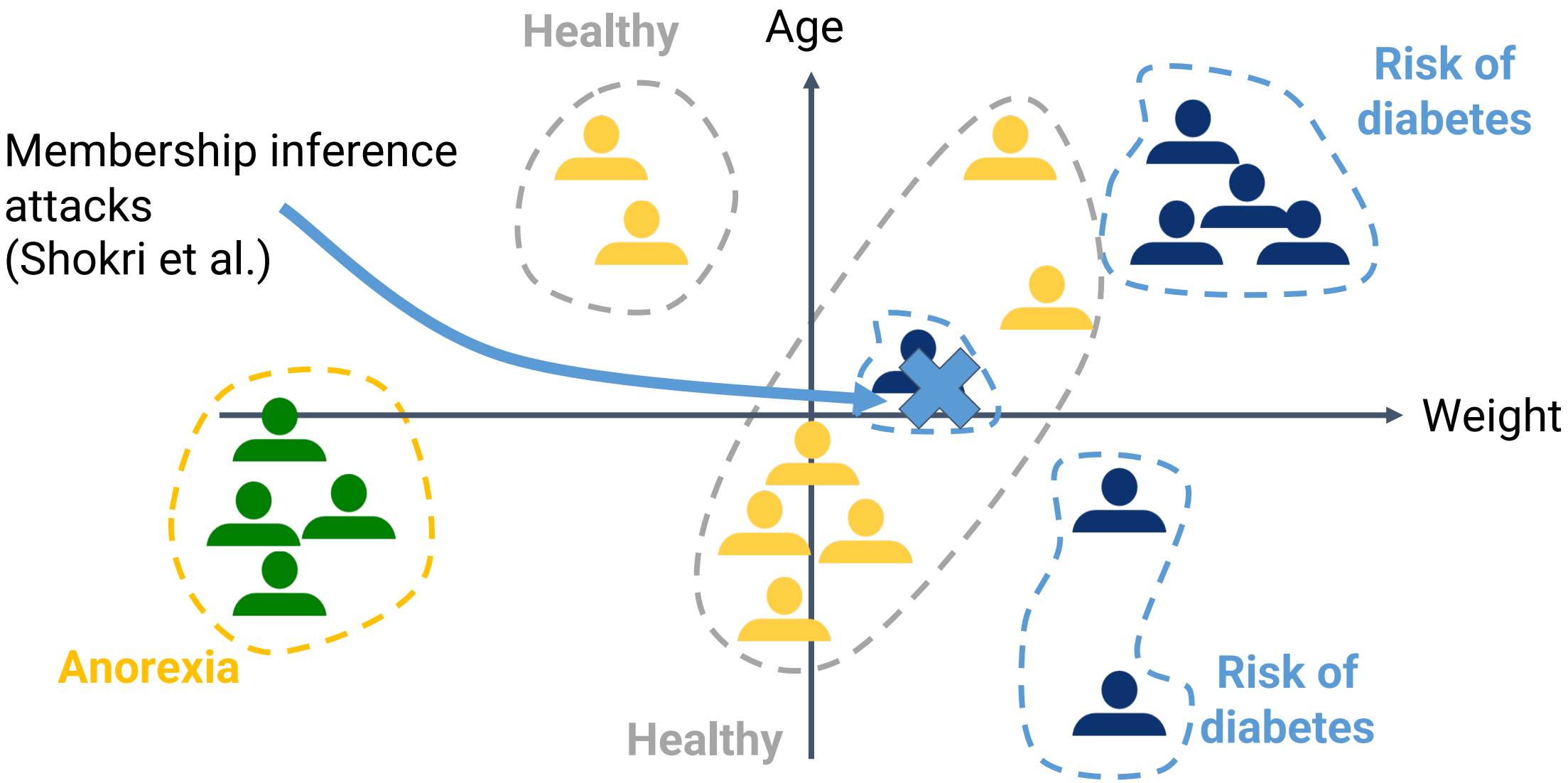

Age
Weight
Risk of diabetes
Anorexia
Healthy


Healthy
Age
Risk of diabetes
Weight
Anorexia
Healthy
Risk of diabetes

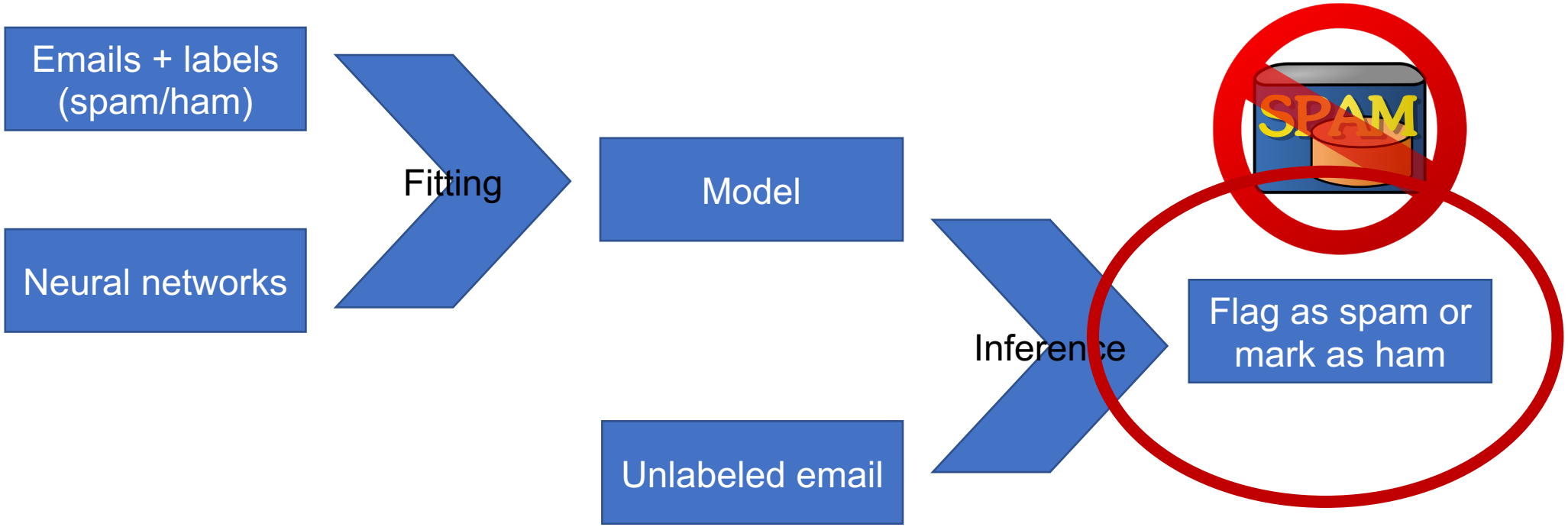

Healthy
Age
Risk of diabetes
Poisoning
Weight
Anorexia
Healthy
Risk of diabetes


Healthy
Age
Risk of diabetes
Adversarial examples (Szegedy et al., Biggio et al.)
Weight
Anorexia
Healthy
Risk of diabetes


Healthy
Age
Risk of diabetes
Weight
Anorexia
Healthy
Risk of diabetes

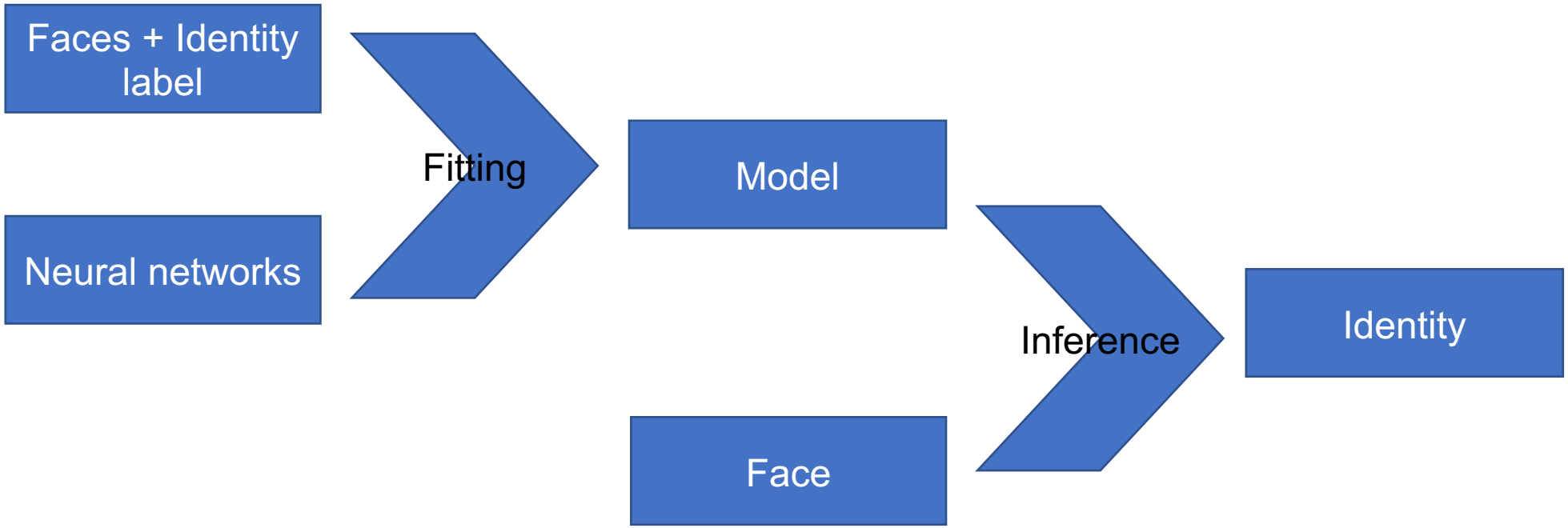

Membership inference attacks (Shokri et al.)
Healthy
Age
Risk of diabetes
Weight
Anorexia
Healthy
Risk of diabetes

ML paradigm in adversarial settings



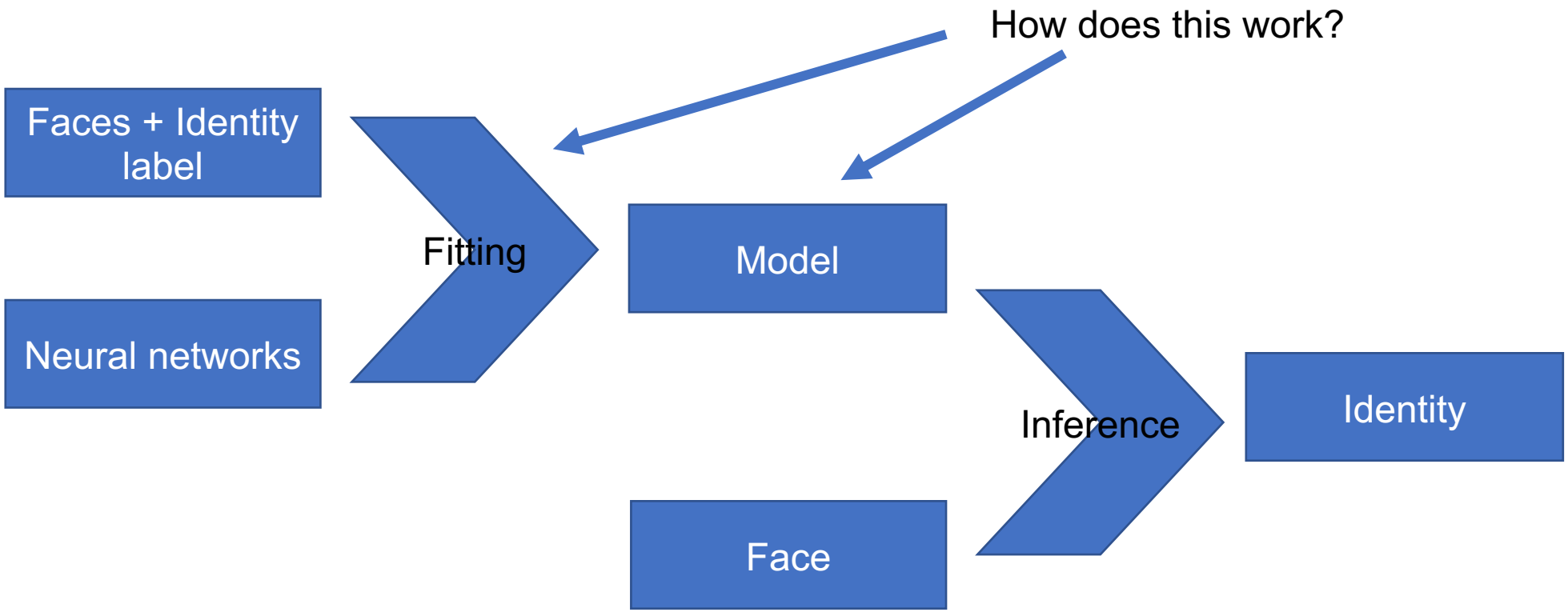
Model extraction: adversary observes predictions and reconstructs model locally

Societal aspects of the ML paradigm



Fairness: if training data does not contain enough faces from a minority or wrong training objective is used, accuracy at inference suffers (model does not build relevant features)

Societal aspects of the ML paradigm



Interpretability: how do we explain a ML algorithm to a human?

#	Date	Topic	Slides	Reading / Assignment
1	Sep 14	Overview & motivation		Reading: 1. Saltzer and Schroeder, The Protection of Information in Computer Systems.
2	Sep 21	Poisoning		Reading: 1. Rubinstein et al., ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors. 2. Jagielski et al., Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. 3. Diakonikolas et al., Sever: A Robust Meta-Algorithm for Stochastic Optimization.
3	Sep 28	Adversarial examples		Reading: 1. Szegedy et al., Intriguing properties of neural networks. 2. Papernot et al., Practical Black-Box Attacks against Machine Learning. 3. Cohen et al., Certified Adversarial Robustness via Randomized Smoothing.
4	Oct 5	Availability		Reading: 1. Rakin et al., Bit-Flip Attack: Crushing Neural Network with Progressive Bit Search. 2. Shumailov et al., Sponge Examples: Energy-Latency Attacks on Neural Networks. 3. Shumailov et al., Manipulating SGD with Data Ordering Attacks.
-	Oct 8	Research project problem statement due		
5	Oct 12	Model stealing		Reading: 1. Tramer et al., Stealing Machine Learning Models via Prediction APIs. 2. Jia et al., Entangled Watermarks as a Defense against Model Extraction. 3. Maini et al., Dataset Inference: Ownership Resolution in Machine Learning.
6	Oct 19	Verification in ML		Reading: 1. Ohrimenko et al., Oblivious Multi-Party Machine Learning on Trusted Processors. 2. Juvekar et al., GAZELLE: A Low Latency Framework for Secure Neural Network Inference. 3. Jia et al., Proof-of-Learning: Definitions and Practice.
7	Oct 26	Data privacy		Reading: 1. Narayanan and Shmatikov, Robust De-anonymization of Large Sparse Datasets. 2. Abadi et al., Deep Learning with Differential Privacy. 3. Choquette-Choo et al., Label-Only Membership Inference Attacks.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Security + Societal = Trustworthy

8	Nov 2	Distributed learning	Reading: 1. McMahan et al., Communication-Efficient Learning of Deep Networks from Decentralized Data. 2. Nasr et al., Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks. 3. Choquette-Choo et al., CaPC Learning: Confidential and Private Collaborative Learning.
	Nov 9	Reading Week	
9	Nov 16	Unlearning	Reading: 1. Song and Shmatikov, Overlearning Reveals Sensitive Attributes. 2. Bourtoule et al., Machine Unlearning. 3. Gupta et al., Adaptive Machine Unlearning.
10	Nov 23	Fairness	Reading: 1. Dwork et al., Fairness Through Awareness. 2. Zemel et al., Learning Fair Representations. 3. Hardt et al., Equality of Opportunity in Supervised Learning.
11	Nov 30	Interpretability	Reading: 1. Zhang et al., Understanding deep learning requires rethinking generalization. 2. Koh and Liang, Understanding Black-box Predictions via Influence Functions. 3. Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.

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.

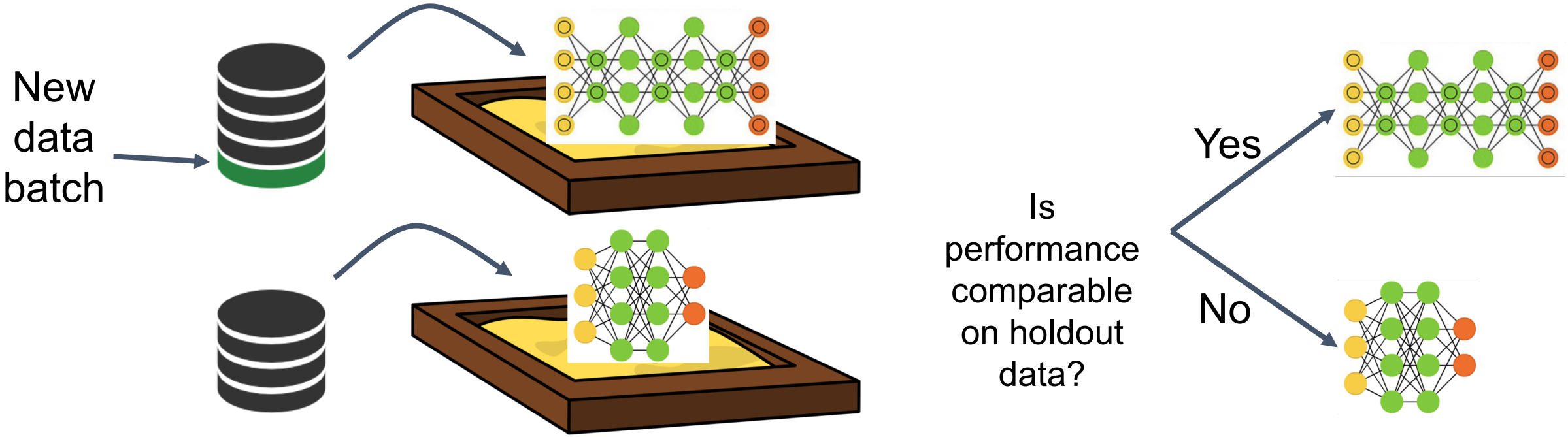
Compromise recording.
Mechanisms that reliably record compromises can be used in place of mechanisms that prevent loss.

Fail-safe defaults

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

Example 2: mitigate data poisoning resulting in a distribution drift

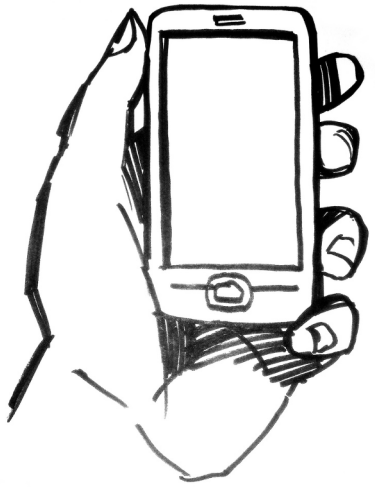
Attacker: submits poisoned points to gradually change a model's decision boundary
Defender: compares accuracy on holdout validation set ***before*** applying gradients



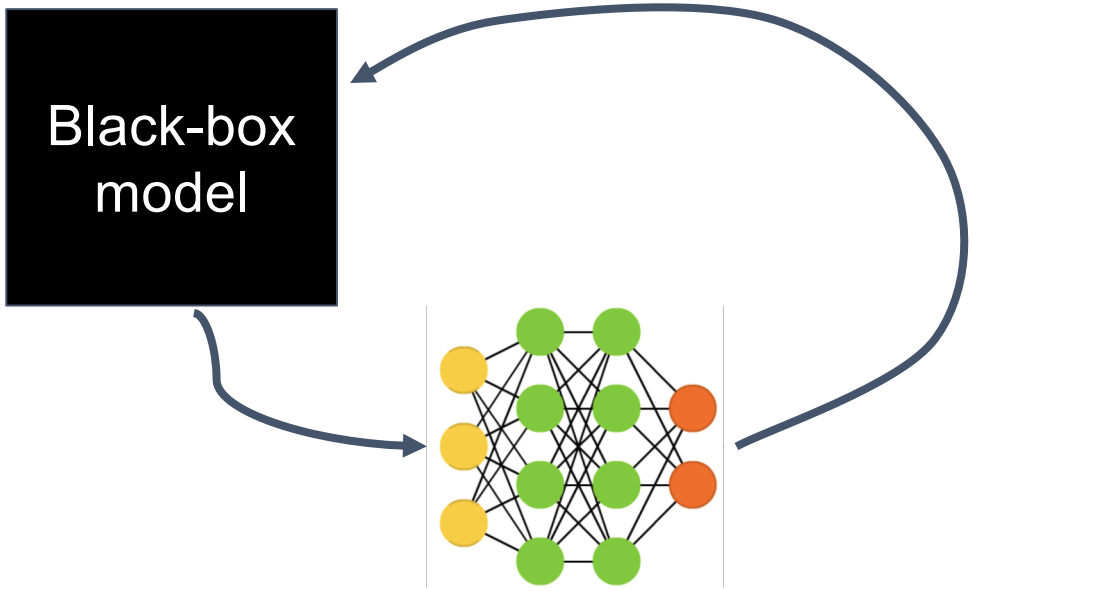
Open design

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

Insider leaks model



Reverse engineering

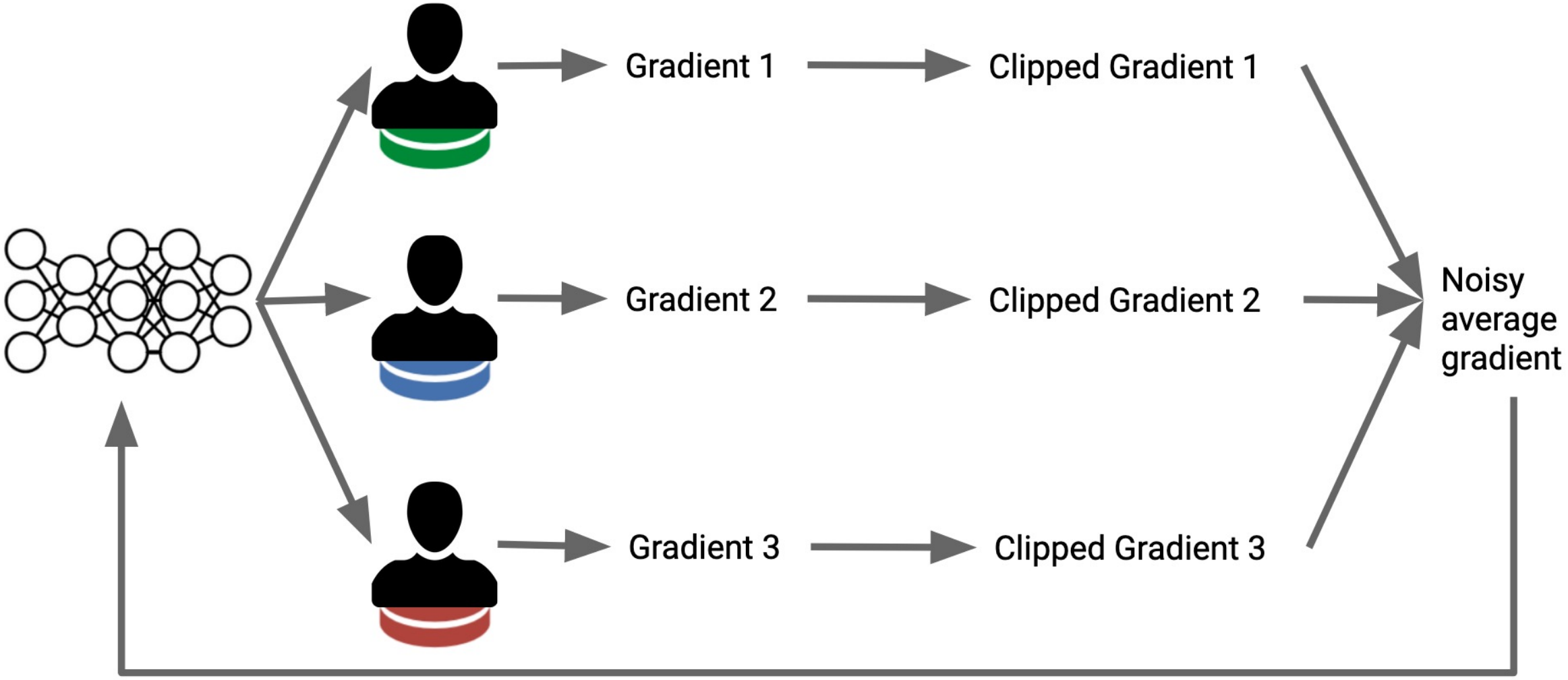


Model extraction

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.

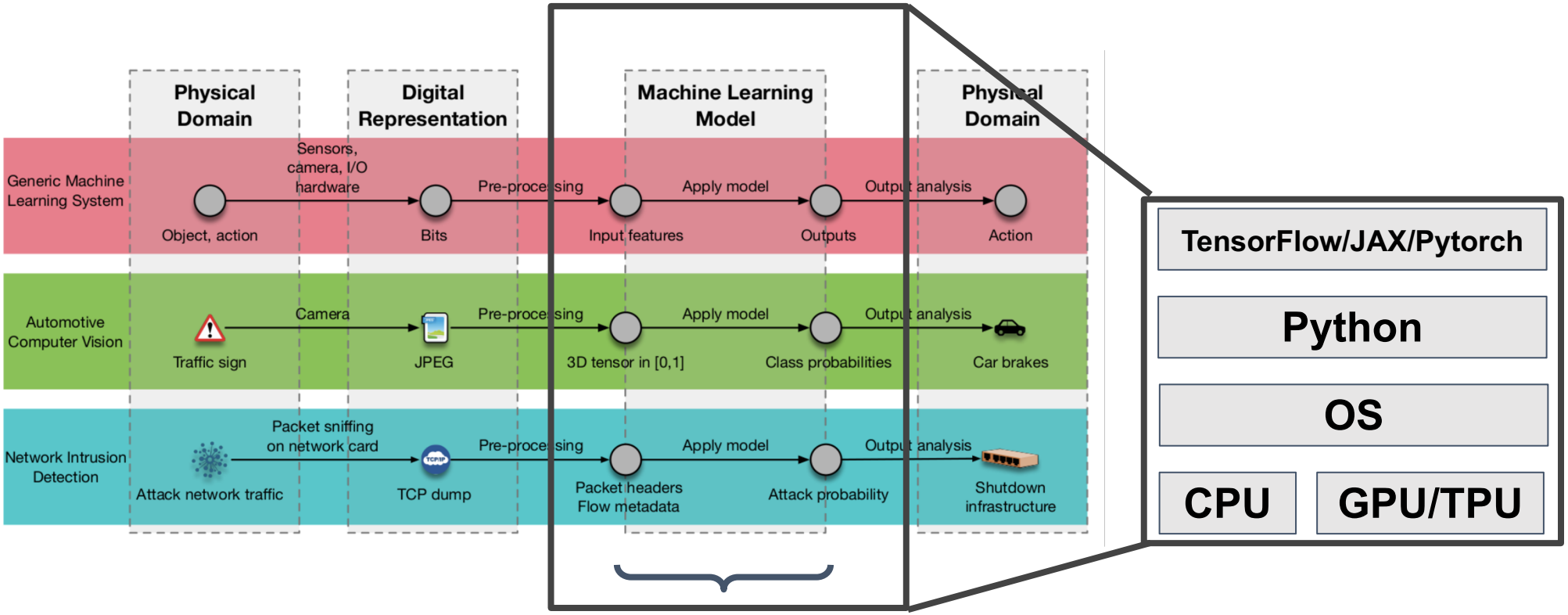
Psychological acceptability.
Human interface designed for ease of use.

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?



- Syllabus: papernot.fr/teaching/f21-trustworthy-ml
- Email: nicolas.papernot@utoronto.ca