Biometric Backdoors: A Poisoning Attack Against Unsupervised Template Updating

Giulio Lovisotto, Simon Eberz and Ivan Martinovic

giulio.lovisotto@cs.ox.ac.uk

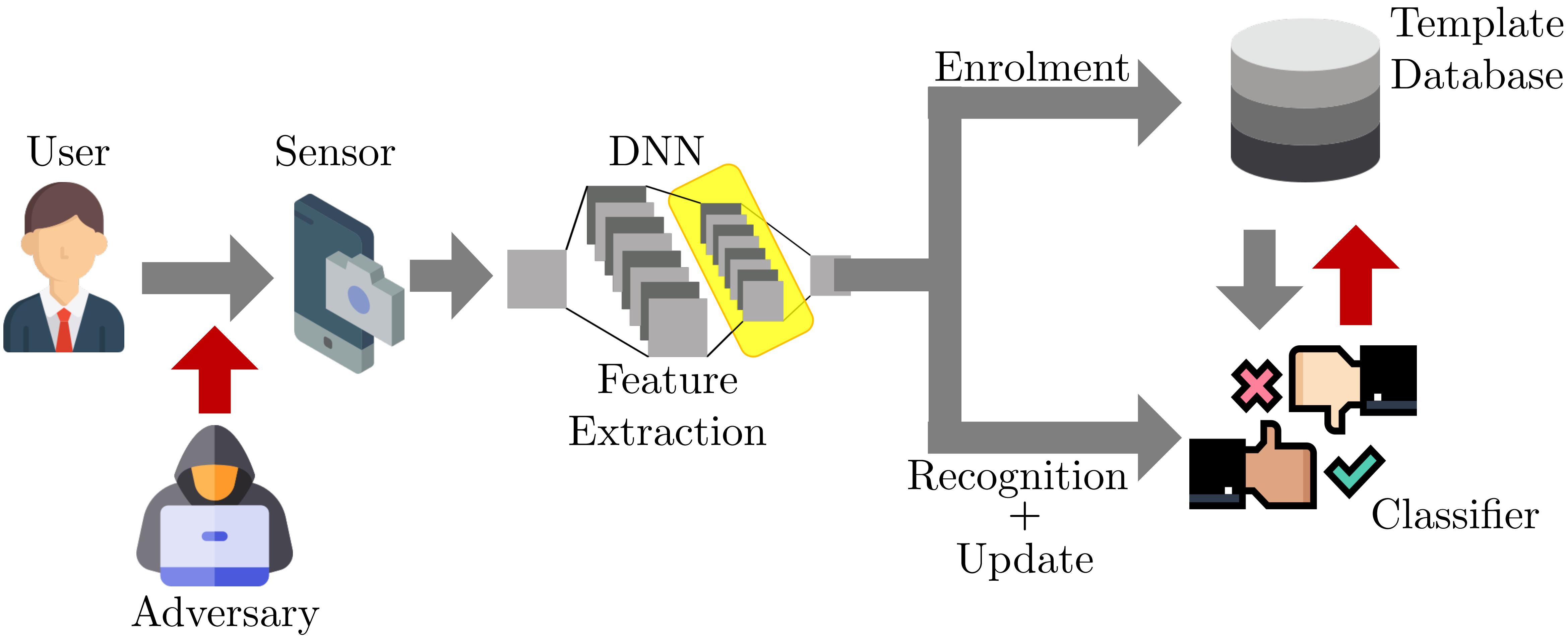
University of Oxford, UK

Biometric Backdoor?

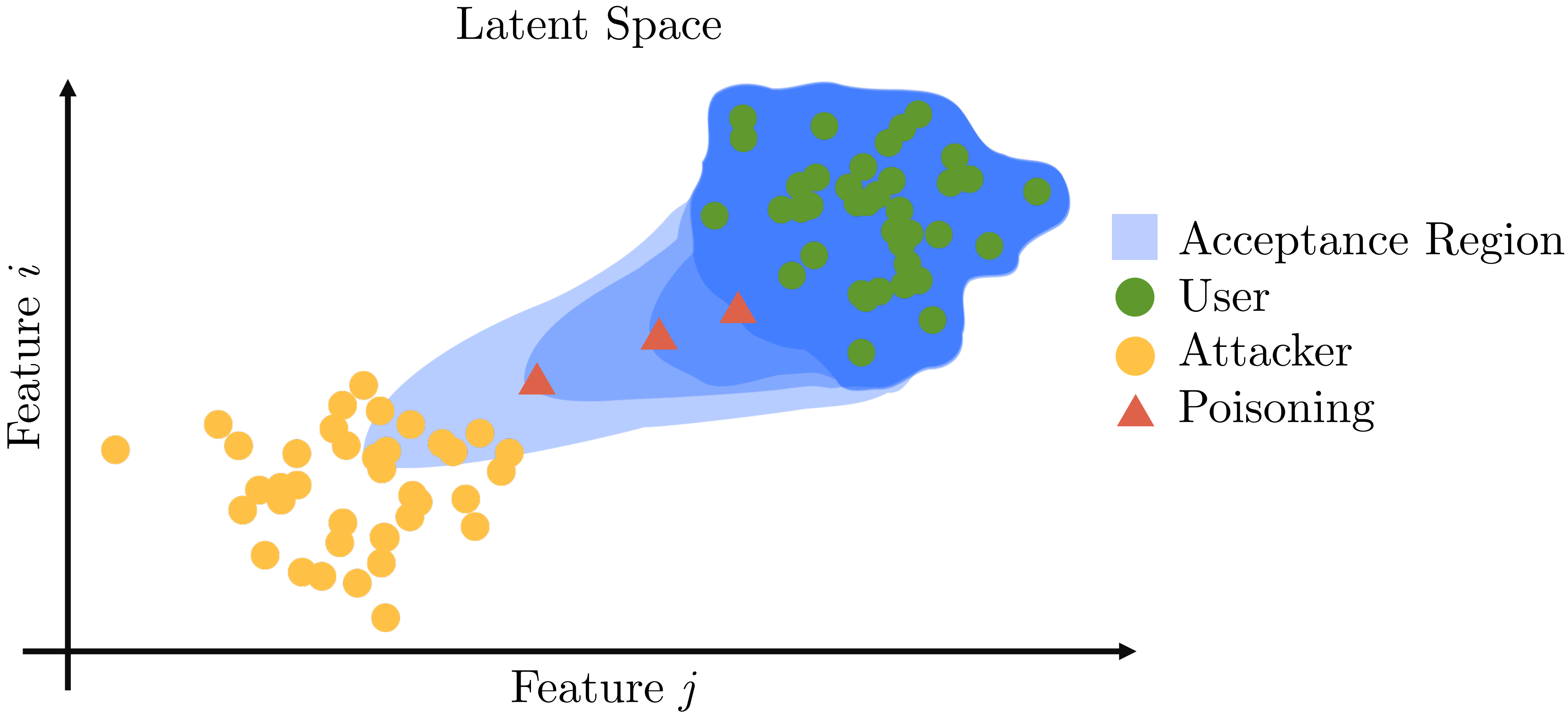
- **Accessories** for impersonation:
  - Fashionable ✓
  - Physically realizable ✓
  - Suspicious ✗
  - One-shot ✗

- Can we design an attack that grants (i) **long-term** and (ii) **inconspicuous** impersonation?

In Consumer Biometric Recognition



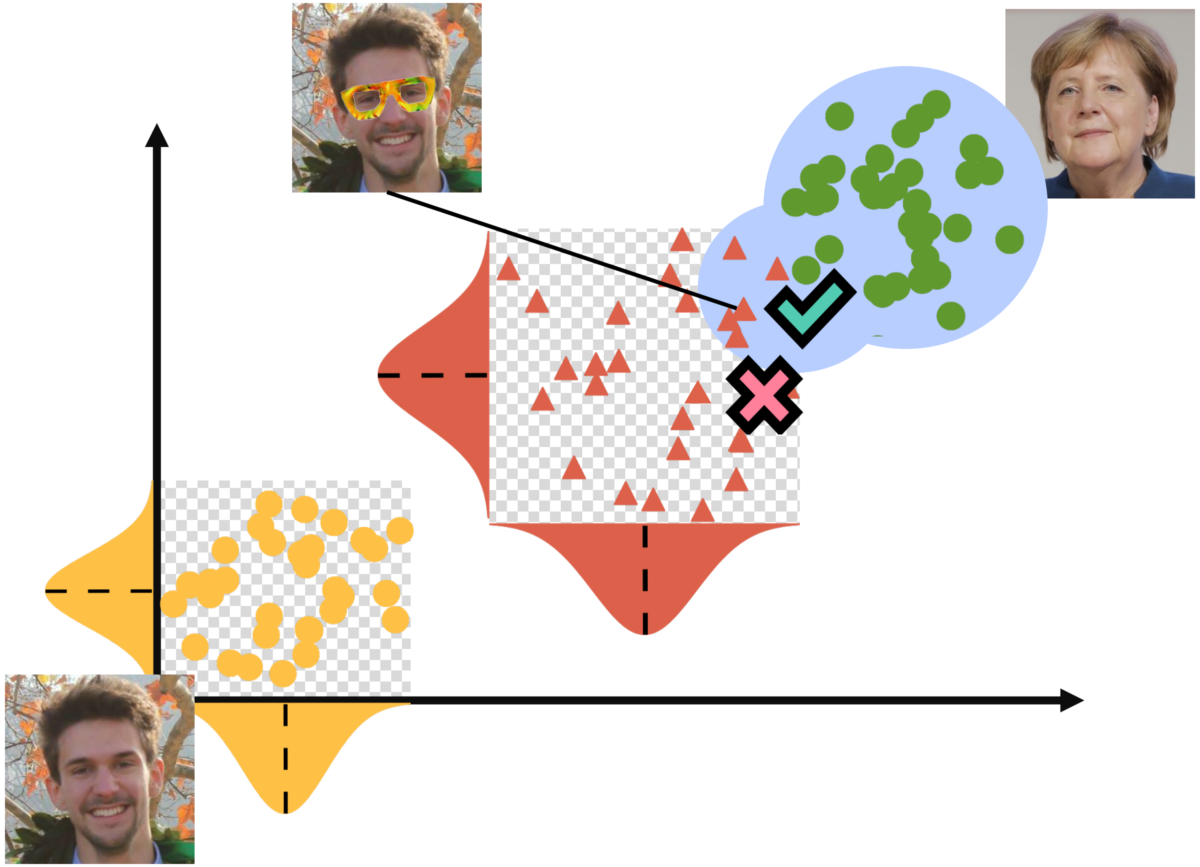
Backdoor (or Poisoning) Outline



Latent Space

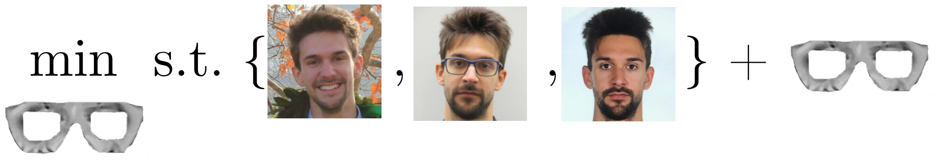
Challenges

- Crafting malicious inputs ▲
- Control input variability
- Limit # of attempts



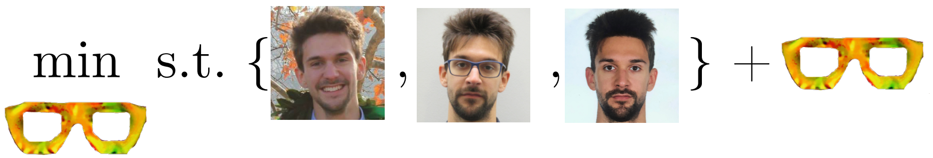
Method (briefly)

- Optimize **starting** glasses



min s.t. { , , } + = ● centroid

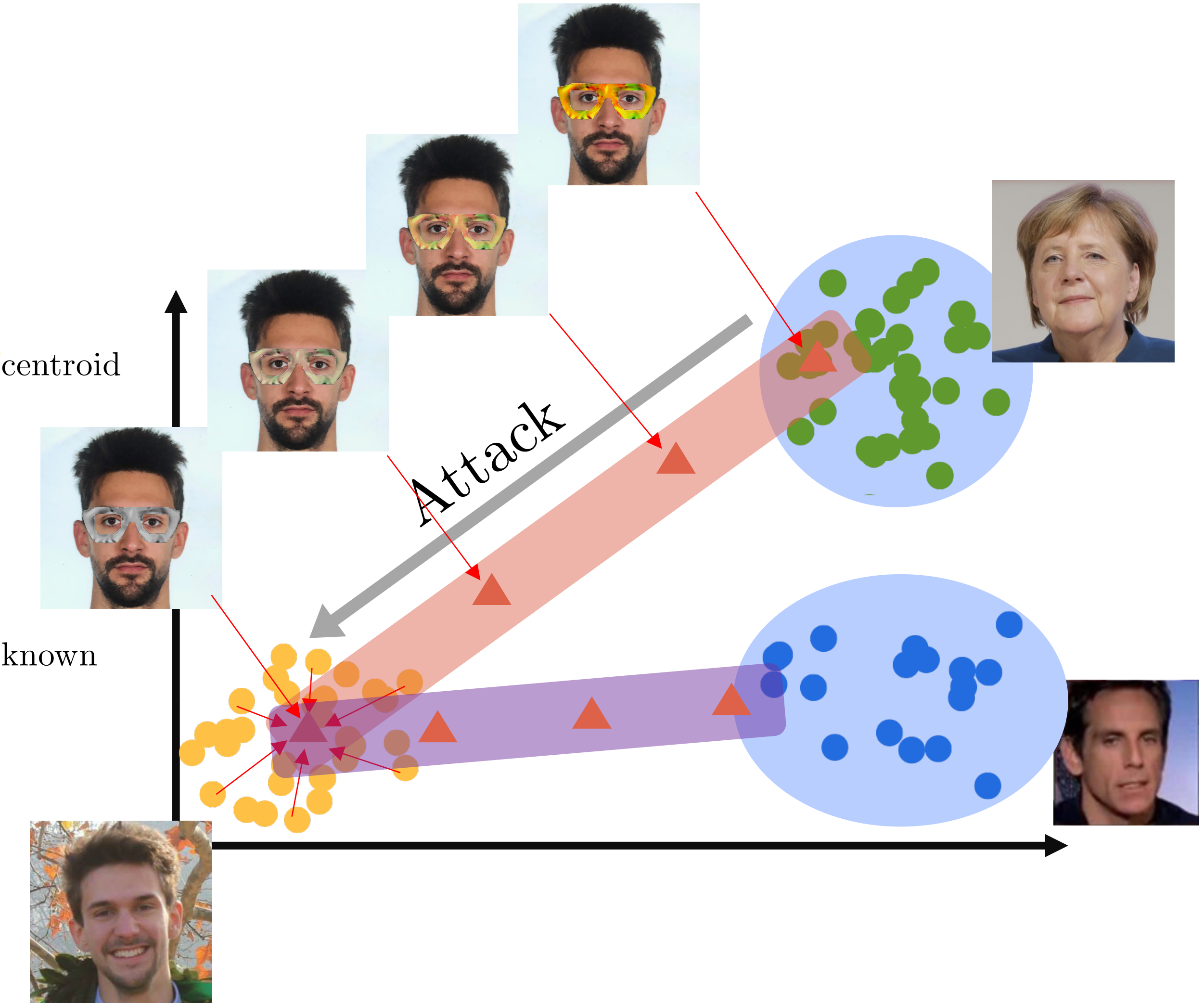
- Generate **all** poisoning glasses



min s.t. { , , } + = ● known

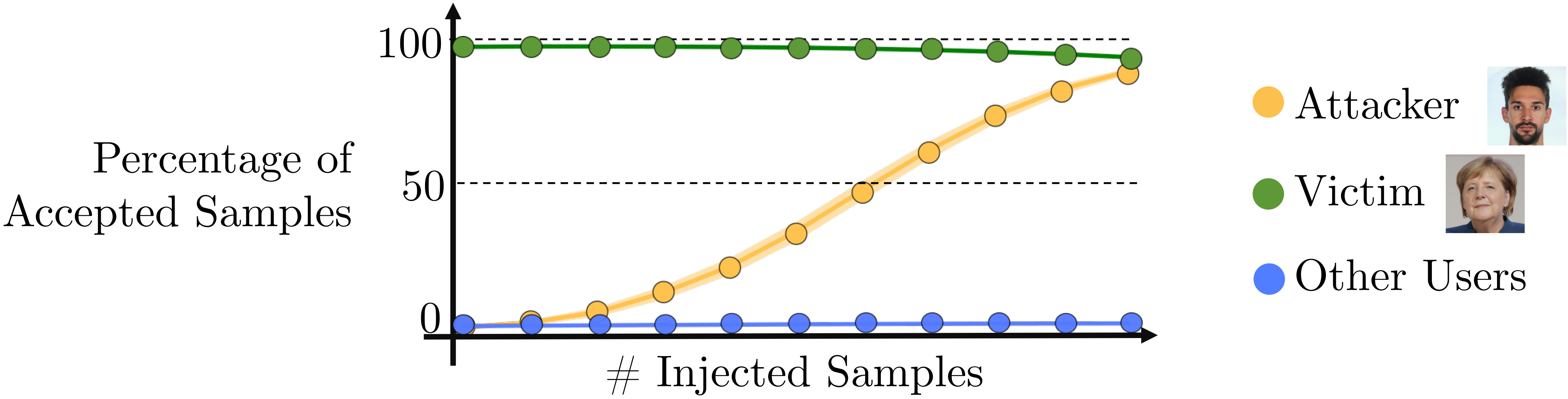
- Estimate sample to inject with **population** data

🟥 ≈ 🟪

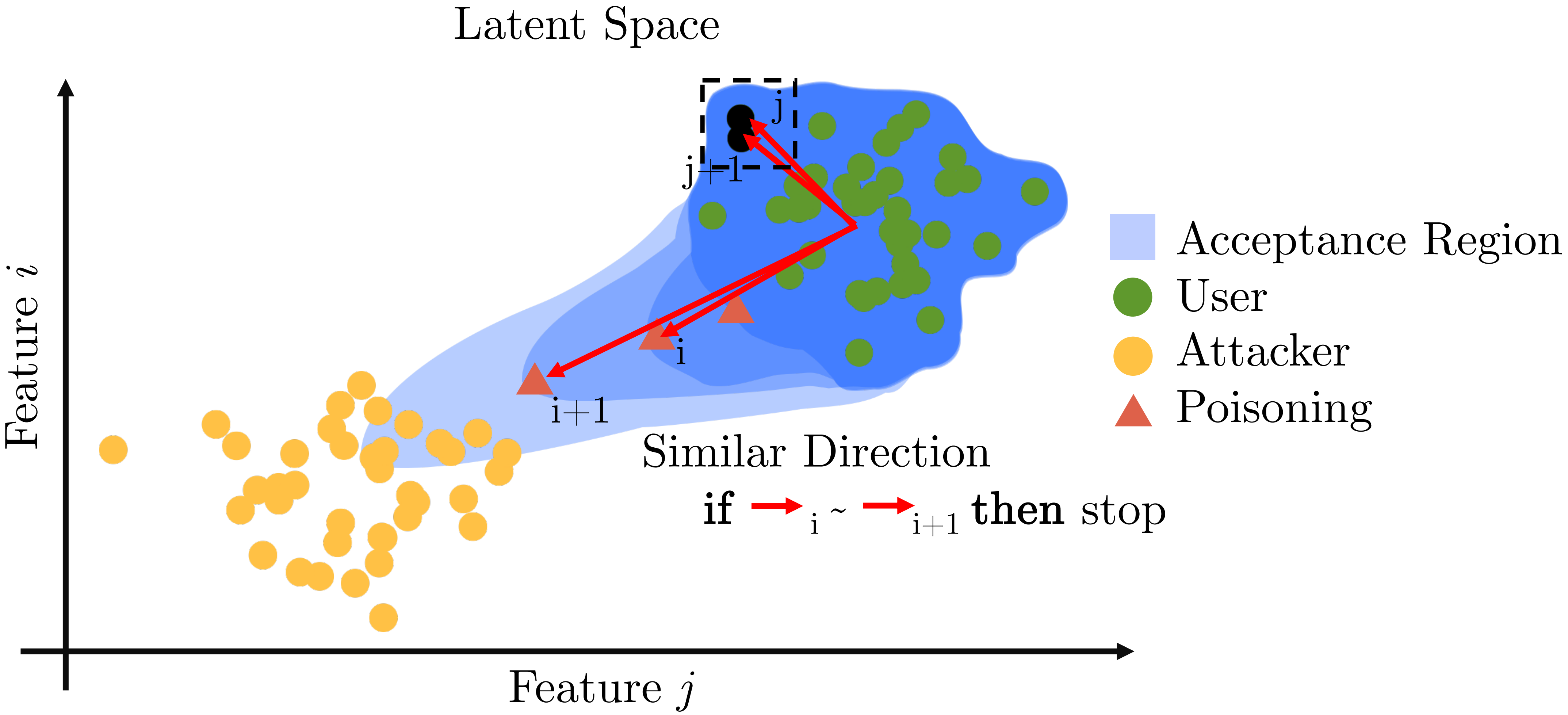


Results Takeaways

- Few injected samples suffice for the adversary to impersonate
- Victim can still authenticate with barely any performance degradation
- The system can still reject non-legitimate “other” users



Countermeasure

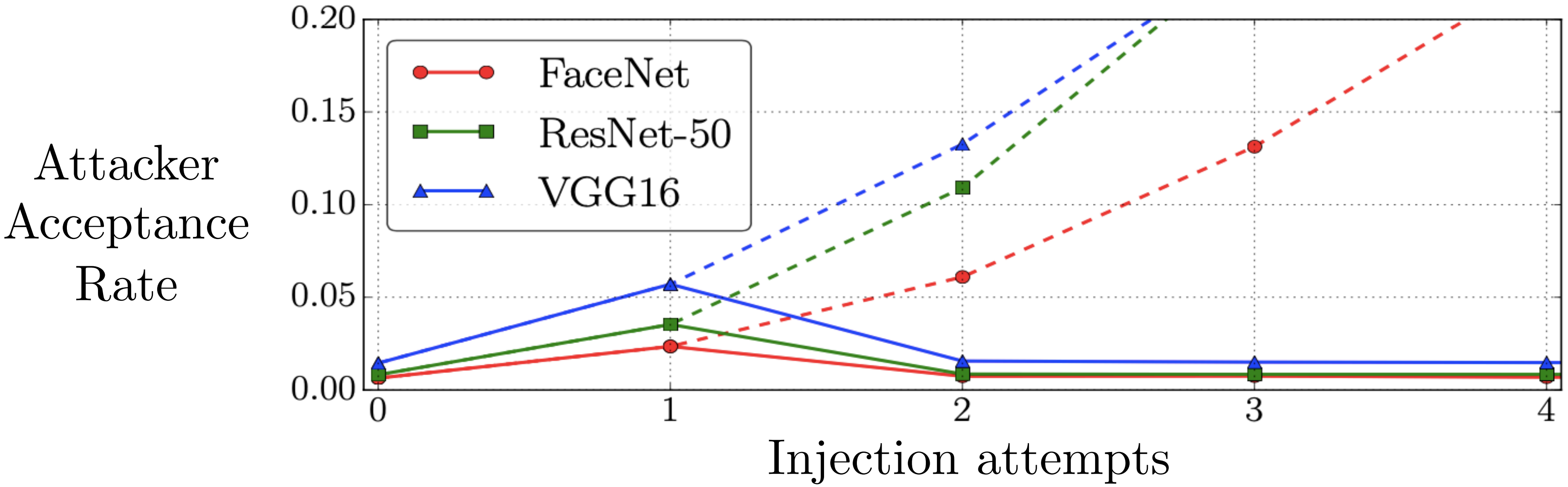


Similar Direction

if $\rightarrow_i \sim \rightarrow_{i+1}$ **then** stop

Countermeasure

- **Insight**: intra-user variation factors may generate consistent directional updates
- **Evaluation**:
    1. Select legitimate sequences of updates which generate directional shifts
    2. Test the detection with a binary threshold on the cosine similarity
- **Result**: 93% detection rate (@EER) on whether a pair of updates is malicious.



Conclusion

- Introduced a **backdoor attack** by exploiting the unsupervised **template update** procedure:
  - The attack copes with *limited knowledge* about the victim and *limited capabilities* of injection
  - A successful attack leads to *inconspicuous* and *long-term* impersonation
  - Some classifiers are particularly vulnerable, with only *one* injected sample sufficient to allow impersonation.
- We proposed a **countermeasure** and we evaluated its detection trade-offs with legitimate template updates:
  - Our countermeasure can detect poisoning samples *93%* of times