



Artificial Intelligence for Society
since 1991



Emerging threats in Digital Identity: from basic attacks to deepfakes and synthetic identities

Prof Sébastien Marcel (marcel@idiap.ch)

*Interim management team / Direction Ad Interim
Head of group **Biometrics Security & Privacy**
Idiap Research Institute
Martigny, Switzerland (www.idiap.ch)*

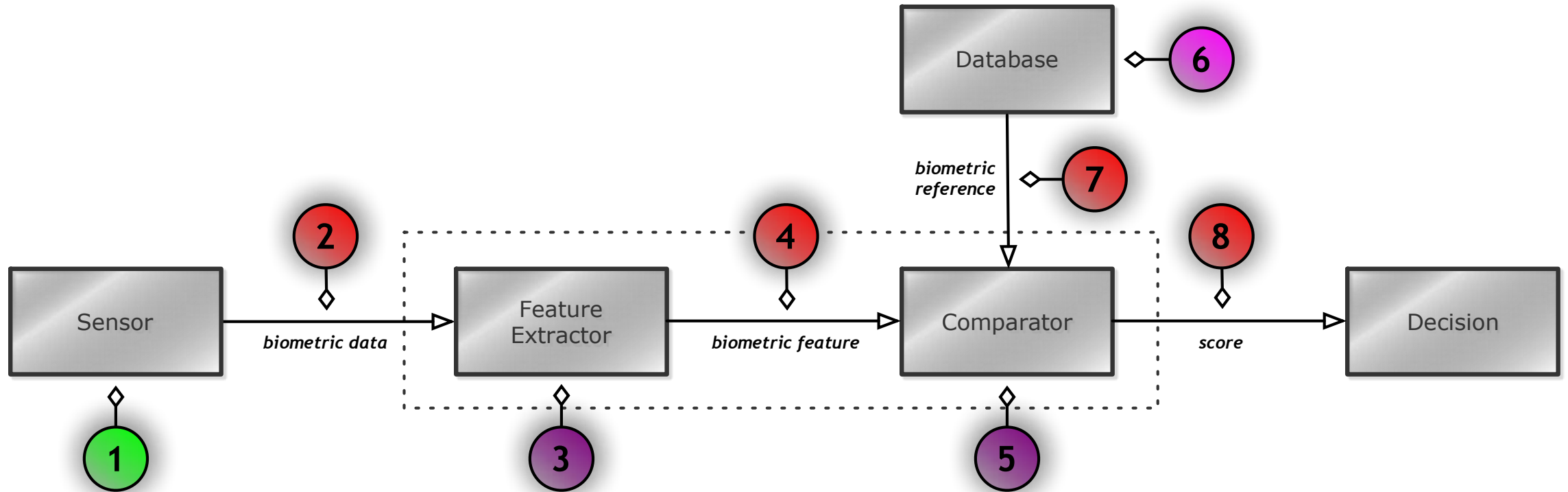
*Professor
School of Criminal Justice (ESC)
Université de Lausanne (UNIL)
Lausanne, Switzerland (www.unil.ch/esc)*

The Age of Synthetic Realities



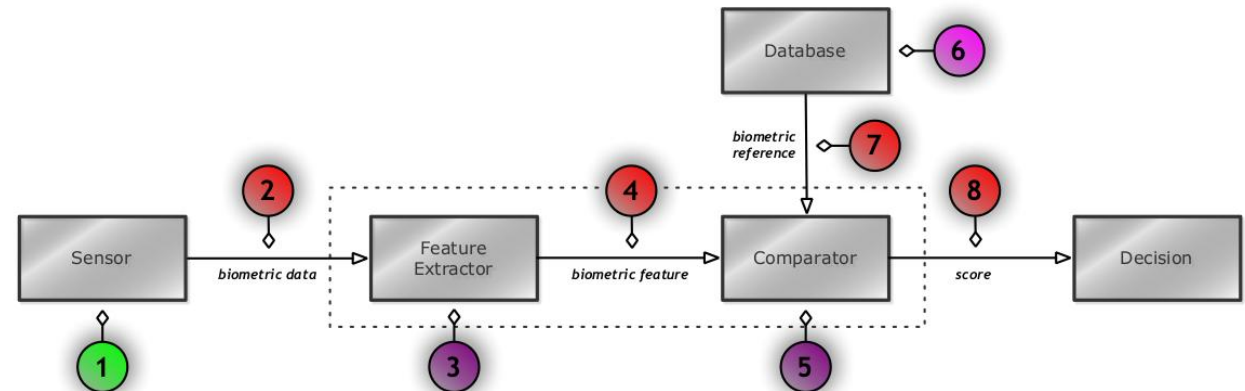
ALL THESE IMAGES ARE FAKES

Vulnerabilities of Digital Identity systems



Attacks: from PAs to IAs (eg. DeepFakes)

- 1 • Presentation Attacks (**PA**)
- 2 • Injection Attacks:
 - Morphing Attacks (**MA**)
 - Deepfakes (**DF**)



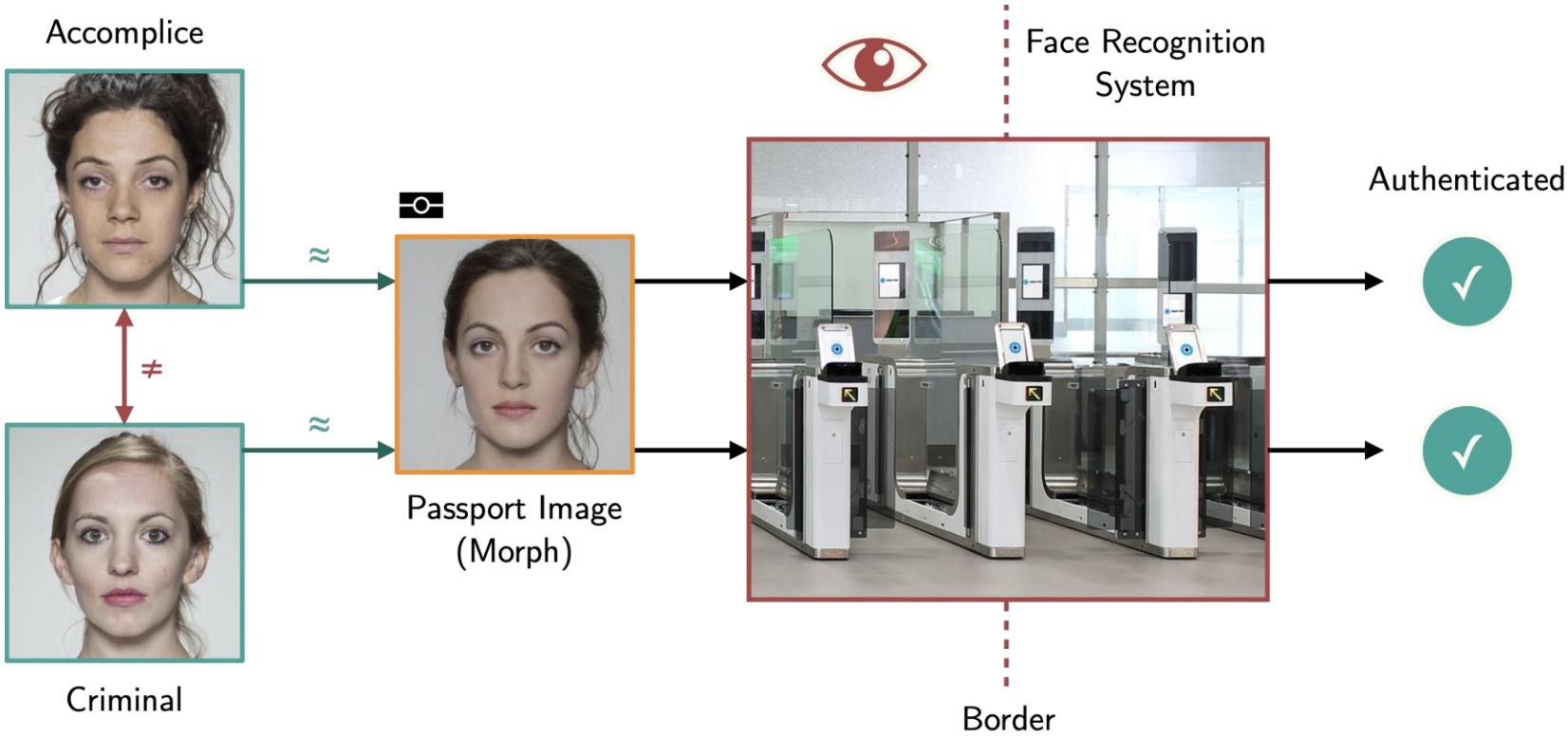
Presentation Attacks



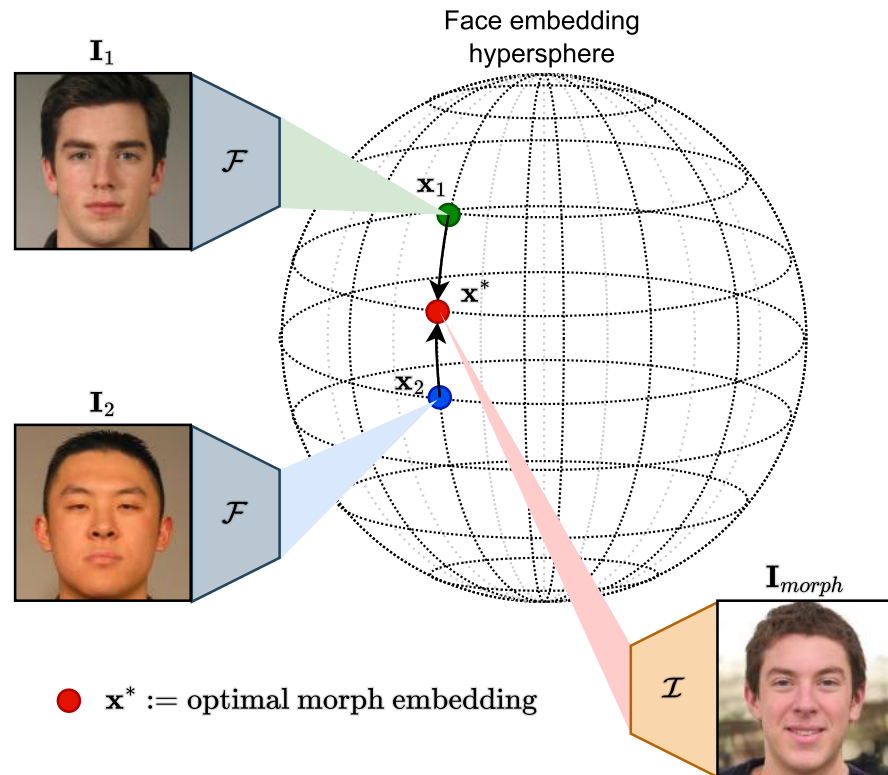
Presentation Attacks



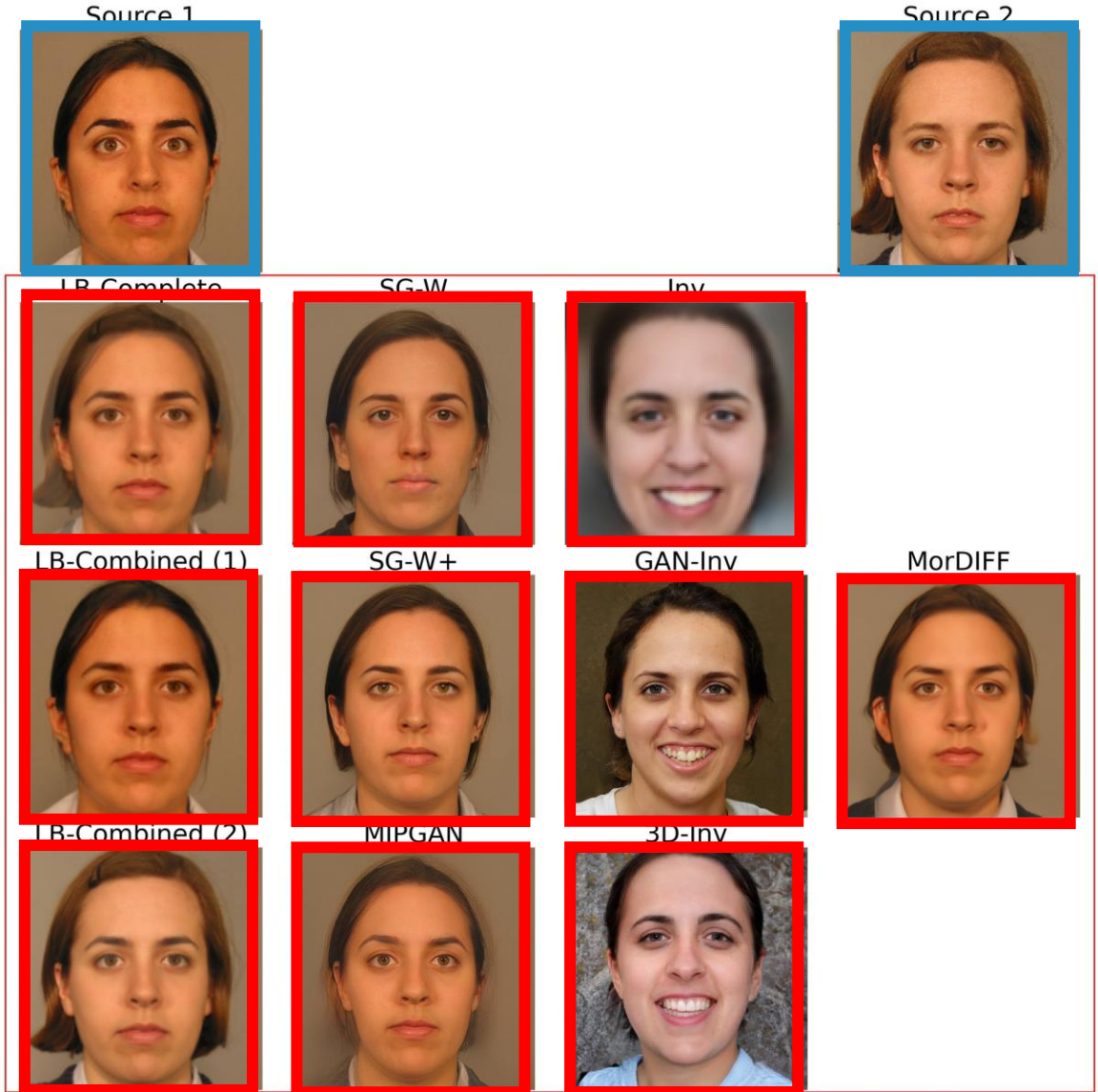
Morphing Attacks



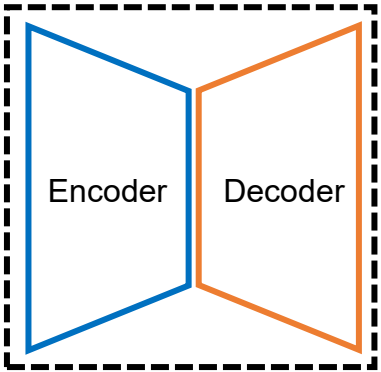
Morphing Attacks



Source images	Embedding extraction	Embedding interpolation	Embedding inversion	Morph
	$\mathbf{x}_i = \mathcal{F}(\mathbf{I}_i)$	$\mathbf{x}^* = \frac{\mathbf{x}_1 + \mathbf{x}_2}{\ \mathbf{x}_1 + \mathbf{x}_2\ }$	$\mathbf{I}_{morph} = \mathcal{I}(\mathbf{x}^*)$	→



DeepFakes: Face-swaps



Attacker



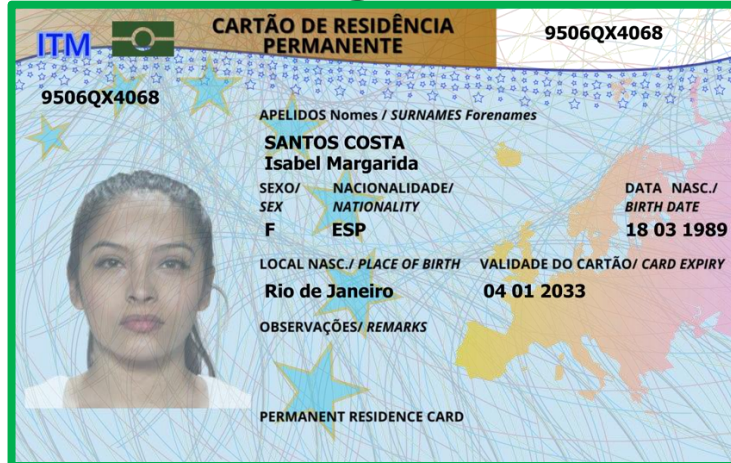
Victim



DeepFake

DeepFakes: Synthetic Identity Documents

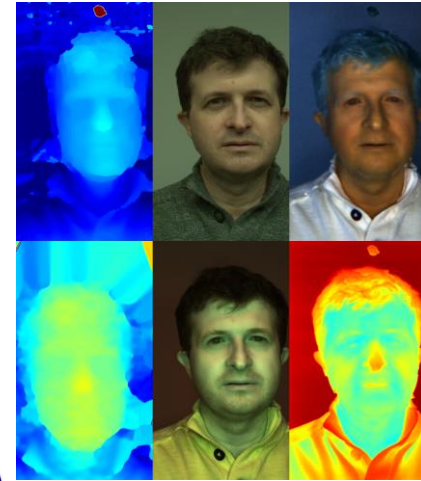
AI-generated images of ID documents (face and text)



Attack Detection is AI-based



- Presentation Attack Detection (**PAD**)
 - e.g. liveness detection, Multispectral, Multimodal (face+voice)
- Morphing Attack Detection (**MAD**)
- Deepfake Detection (**DD**)

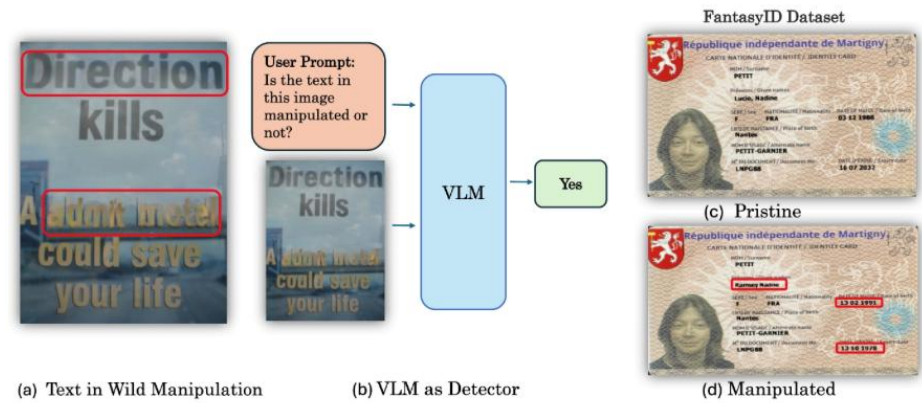


System prompt example System
 You are an AI model for detecting presentation attacks in images. Determine if an image is bonafide (real) or a presentation attack, ...

Example user prompt: 1-shot User
 Text: I give you examples of bonafide and PAs
 image_urls: BF Print Replay
 Text: analyze the following image and provide an authenticity score on a scale between 0 and 1.

Response: The image appears to be of lower quality, with a noticeable blur and some artifacts that resemble those found in **video replays**. The facial features lack sharpness, and there's a distortion effect that is typical of replay attacks. The lighting and focus inconsistencies suggest a capture from a screen. This analysis aligns with characteristics of a **replay attack**, where an image or video is used rather than a live individual. **score: 0.2**

Emerging technologies: Foundation Models for Attack Detection



How to trust a technology vendor ?



- Testing, testing and testing = **certification** is key:
 - International Standards: for definitions and metrics
e.g. «Presentation Attack Detection (PAD)» (ISO/IEC 30107)
 - Independent testing labs i.e. accredited by the FIDO Alliance to certify products





*Prof Sébastien Marcel (marcel@idiap.ch)
Biometrics Security & Privacy
Idiap Research Institute
Martigny, Switzerland (www.idiap.ch)*

*More on Thursday, 14 May from 13:20 to 14:00 Track 2.2
Segment 3 « Securing Trust Across the Identity Ecosystem »*

Q&A

Thanks for your attention



35

years XP

+200

Employees

Research

20
Education

R&D engineers

stitut

in

artificielle

16

Valais



91

Research groups

IDIAP

ceptive

intelligence

Research Institute of National Importance

Non-profit foundation

ale Mo

nnovations



2001

Torch is built at Idiap

2014

Torch becomes the
core of PyTorch
(Facebook)

 PyTorch

Today

60% of AI Research
implementations use PyTorch

+600k Repositories on Github

PyTorch is used at

OpenAI to build ChatGPT

NASA to do space research

Tesla to build self driving cars

AstraZeneca for Drug Discovery