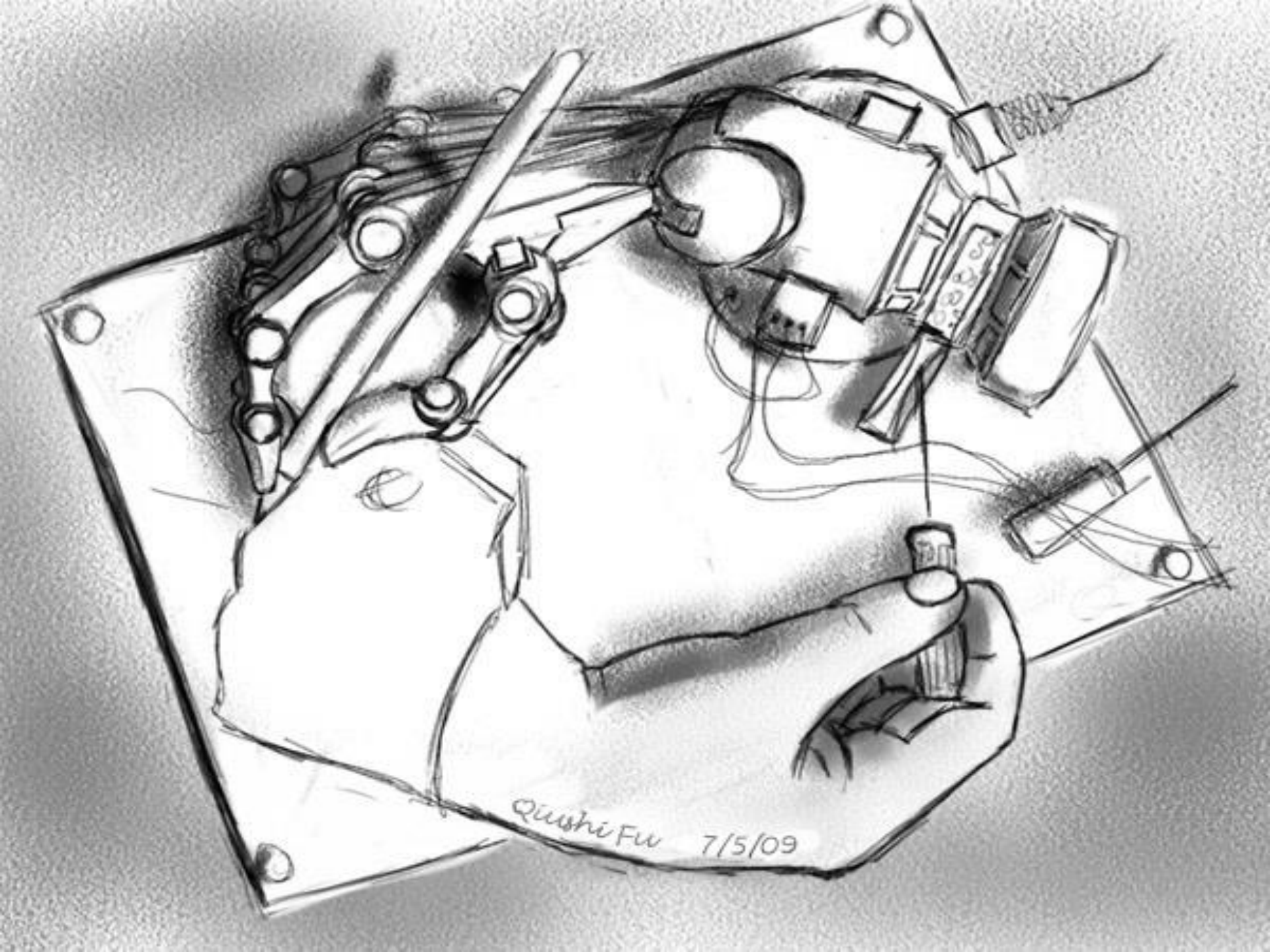
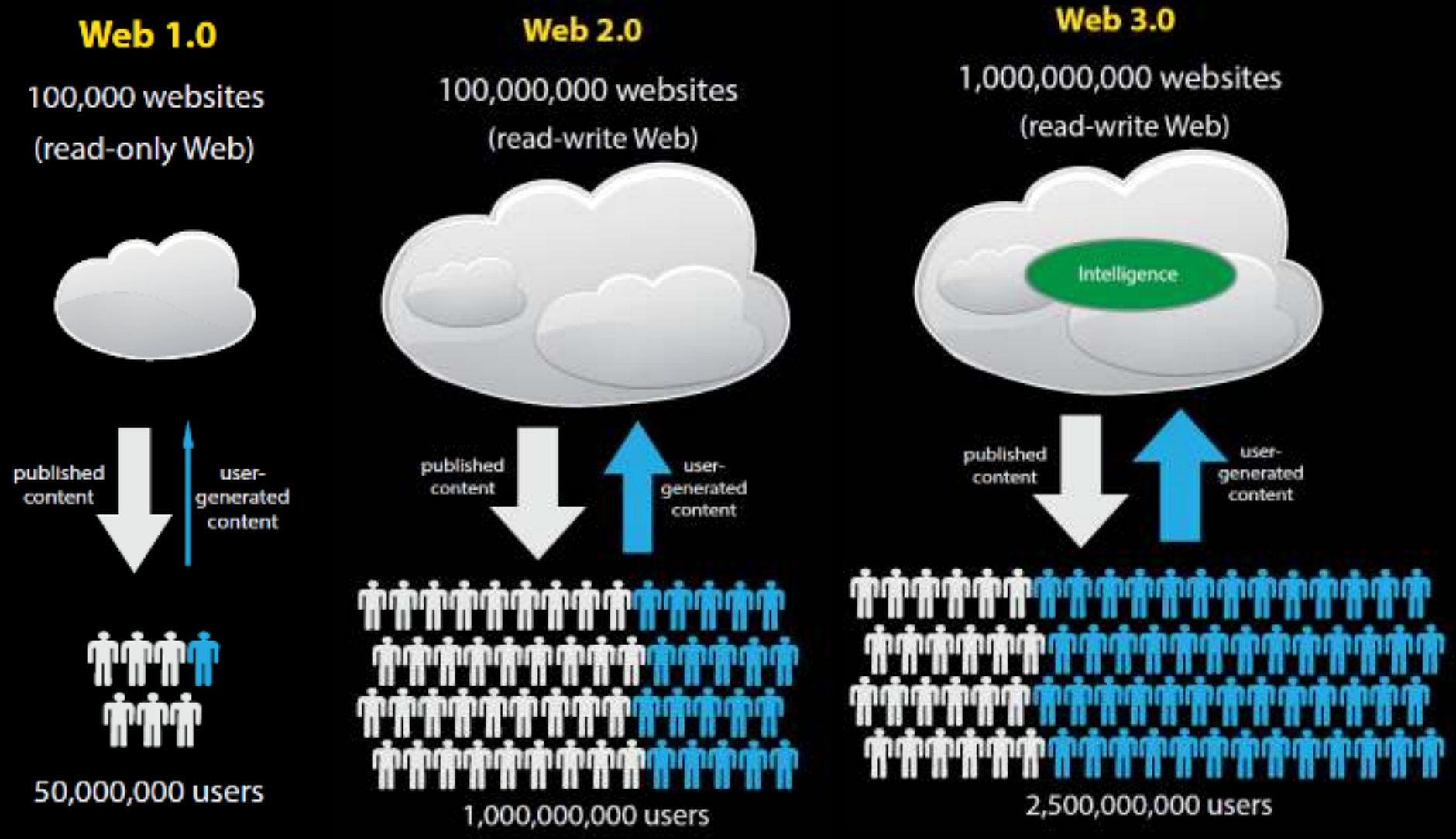
Złośliwe sterowanie ludźmi i maszynami

płk dr inż. Rafał KASPRZYK, e-mail: rafal.kasprzyk@wat.edu.pl

Wydział Cybernetyki
Wojskowa Akademia Techniczna


Qiushi Fu 7/5/09

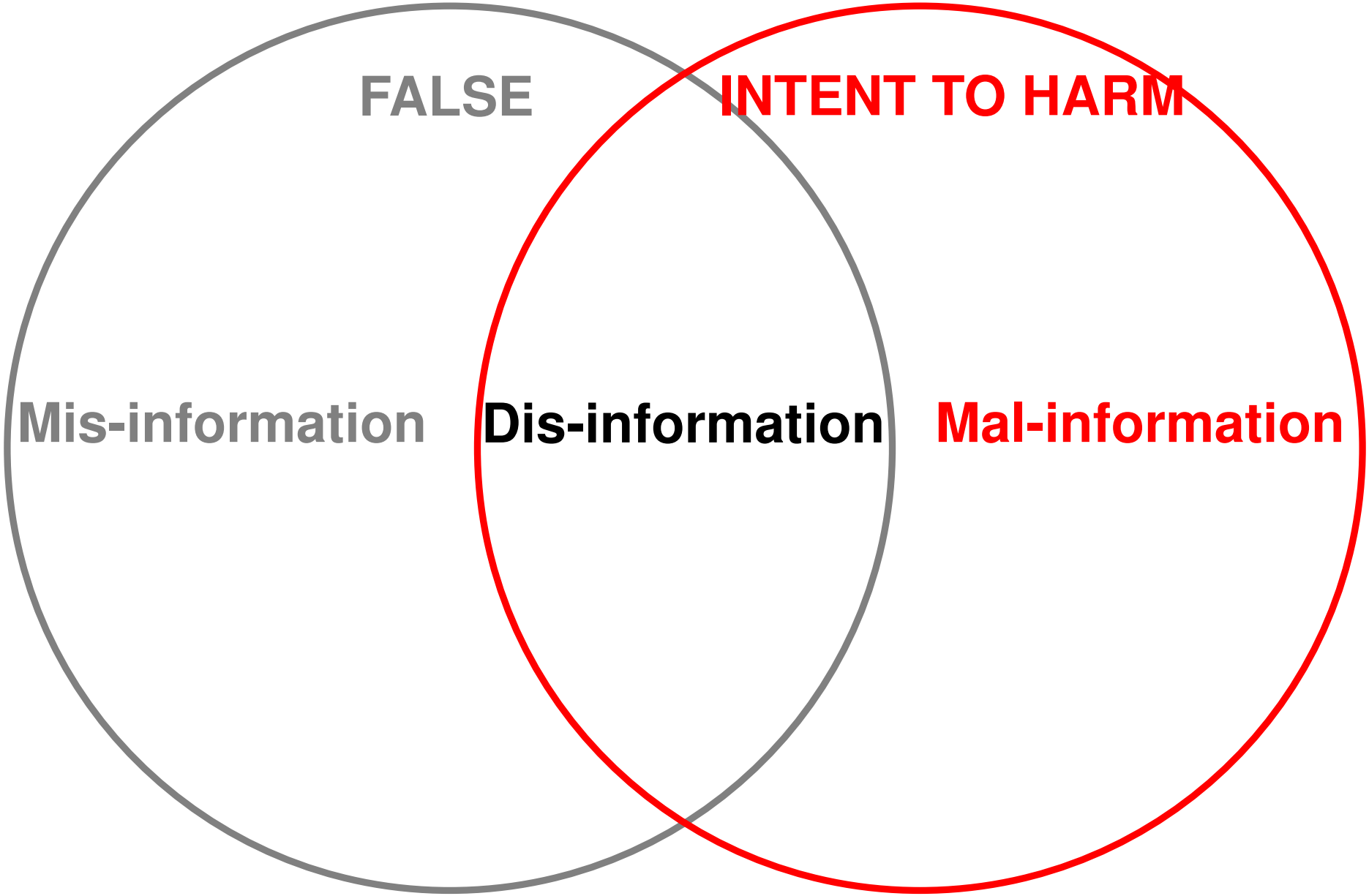
Ewolucja Internetu



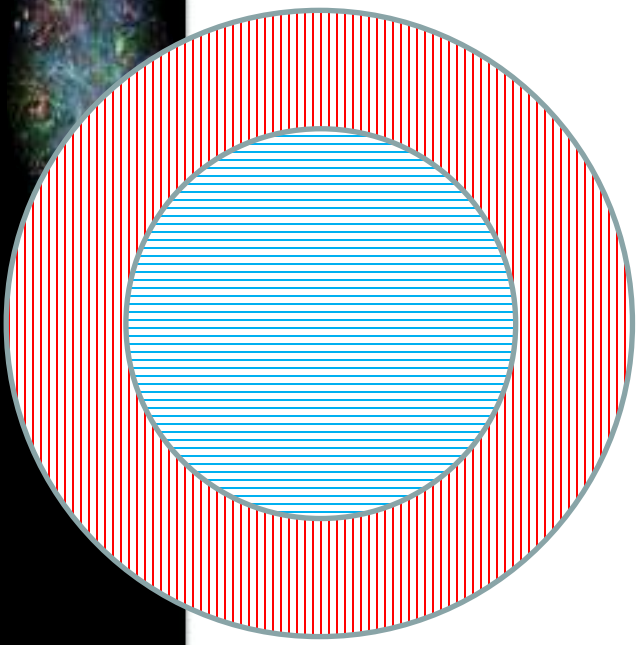
Social Media:

EMOCJE → MANIPULACJA

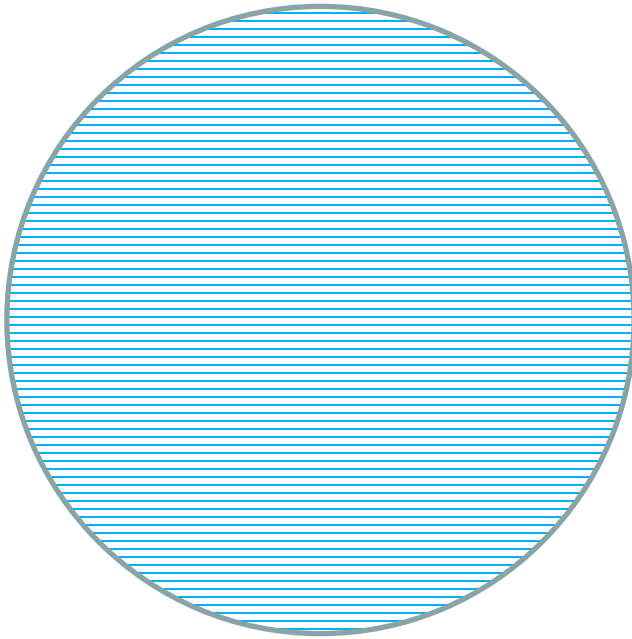
Zaburzenia (patologie) informacji



Proces złośliwego sterowania ludźmi



Obiekt X



Obiekt Y

Istota refleksji (1/2)

Refleksja pierwszego rodzaju (autorefleksja)

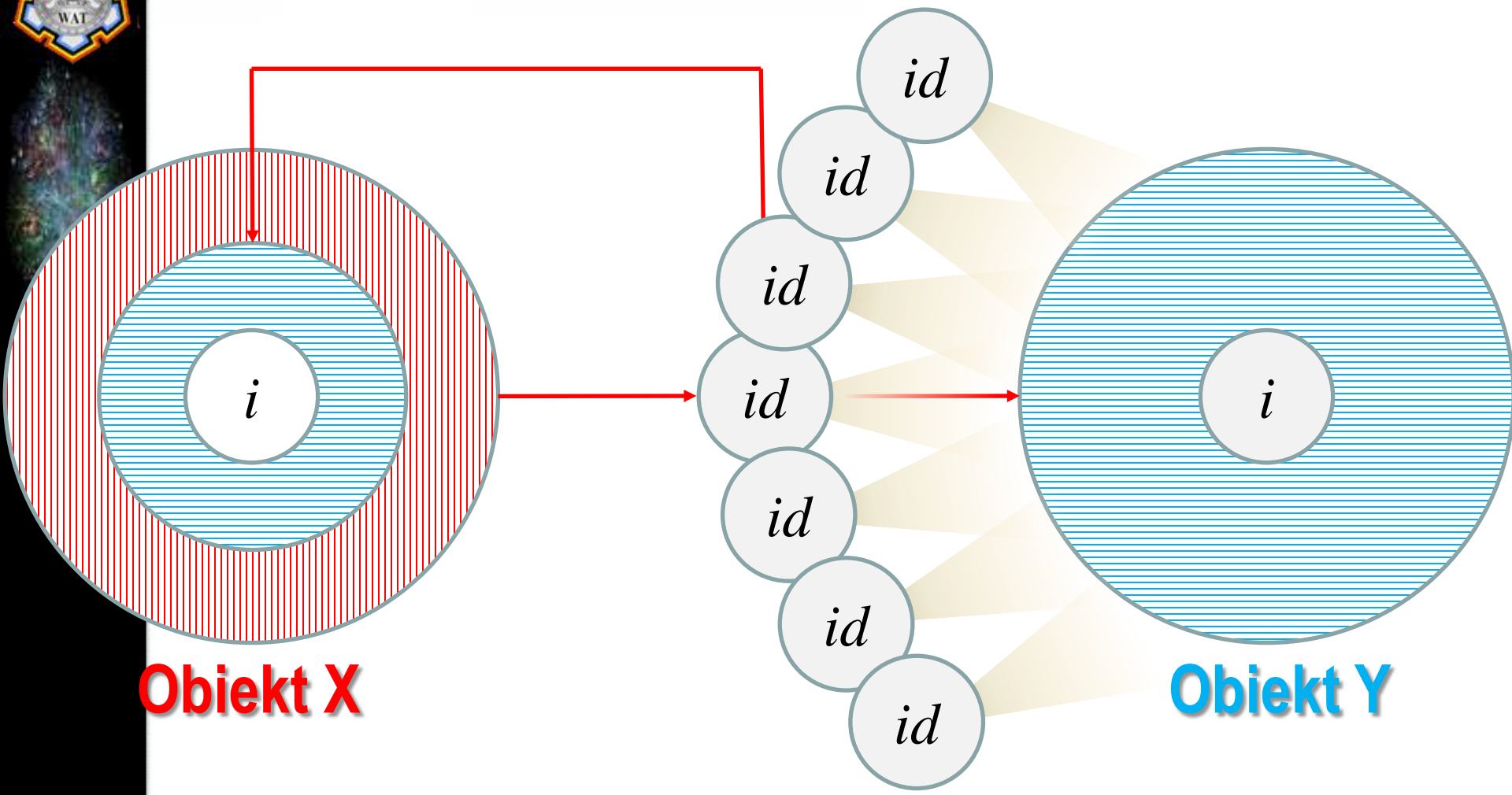
Refleksja - zdolność do przyjęcia perspektywy obserwatora w stosunku do swojej własnej wiedzy, przekonań, pragnień i intencji.

Istota refleksji (2/2)

Refleksja drugiego rodzaju

Refleksja - zdolność do przyjęcia perspektywy obserwatora w stosunku do swojej własnej wiedzy, przekonań, pragnień i intencji oraz wiedzy, przekonań, pragnień i intencji innych obiektów.

Proces złośliwego sterowania ludźmi



Sterowanie refleksyjne to zmiana podejścia z próby przewidywania procesów decyzyjnych przeciwnika na wpływanie na procesy decyzyjne przeciwnika za pomocą zaburzeń informacyjnych, w taki sposób aby…

Budowa „inteligentnych” maszyn

MODEL

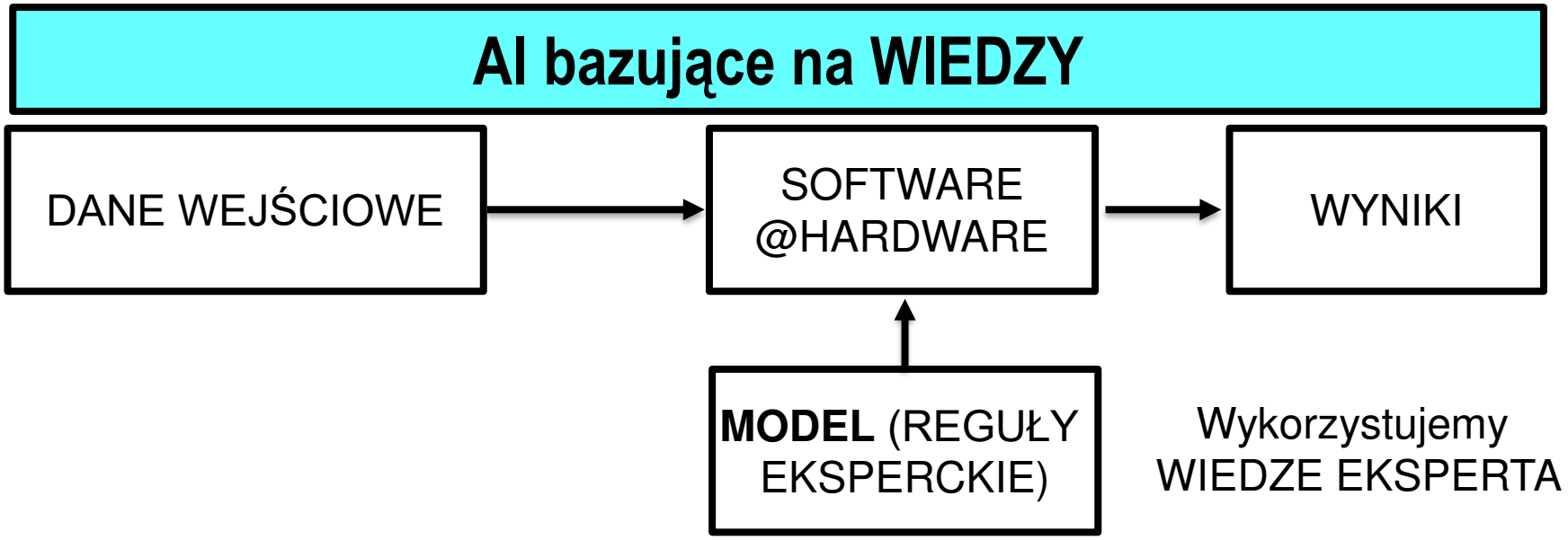
```
…
if object contains red
   then mark is-enemy;
if object contains … then …;
if object contains … then …;
…
```

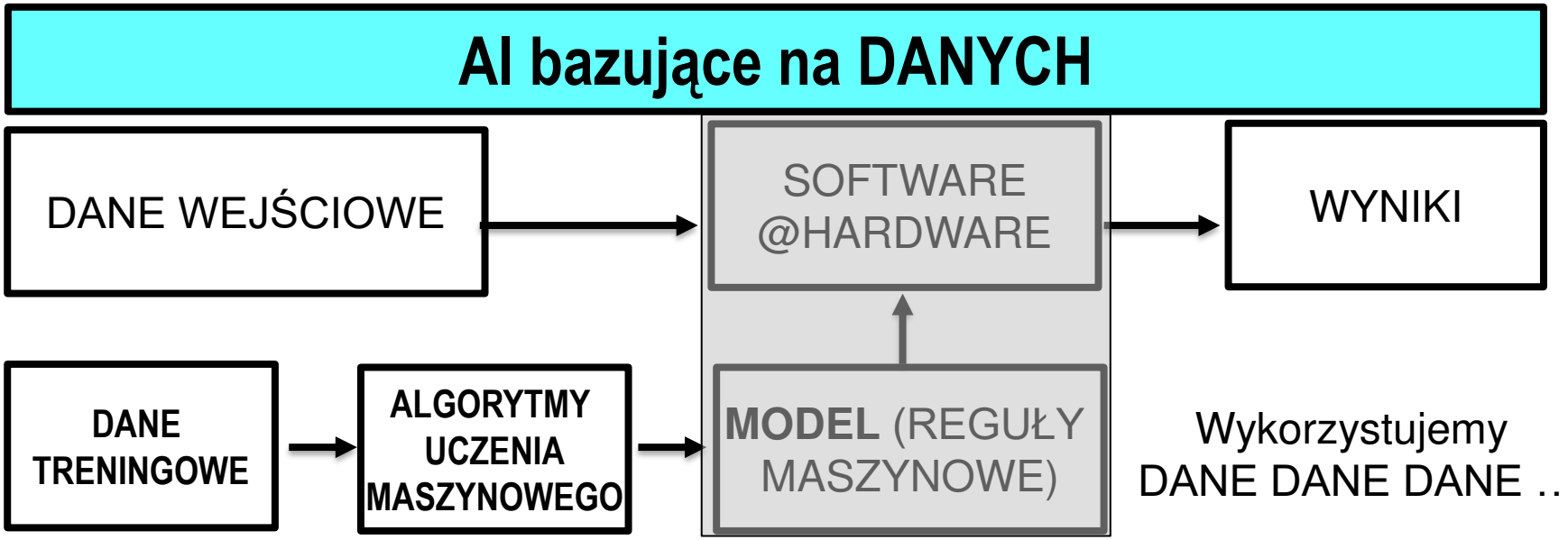
Wykorzystujemy
WIEDZĘ EKSPERTA

```
…
try to describe some objects;
change self to reduce errors;
repeat;
…
```

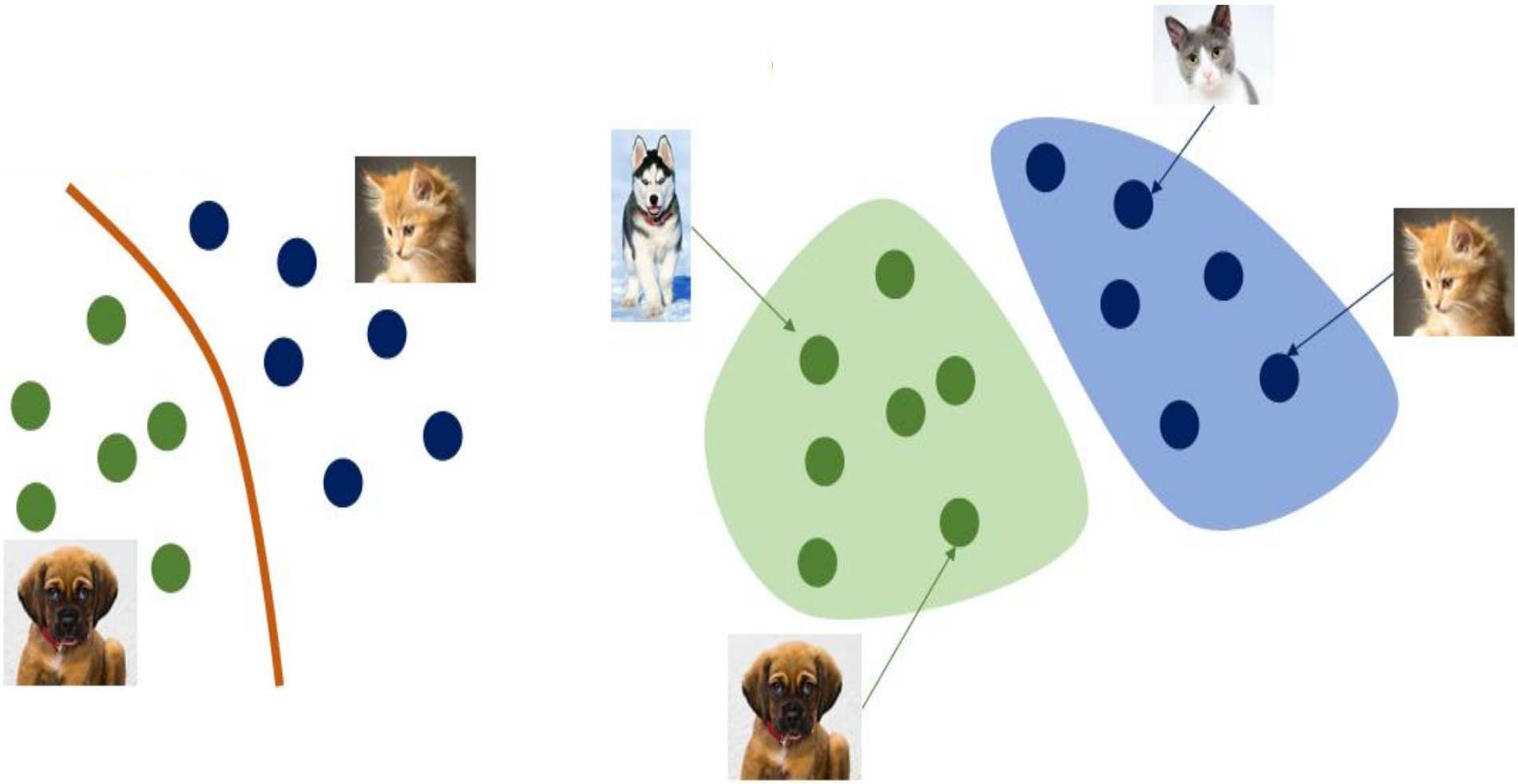
Wykorzystujemy
DANE DANE DANE

„Inteligentne” maszyny

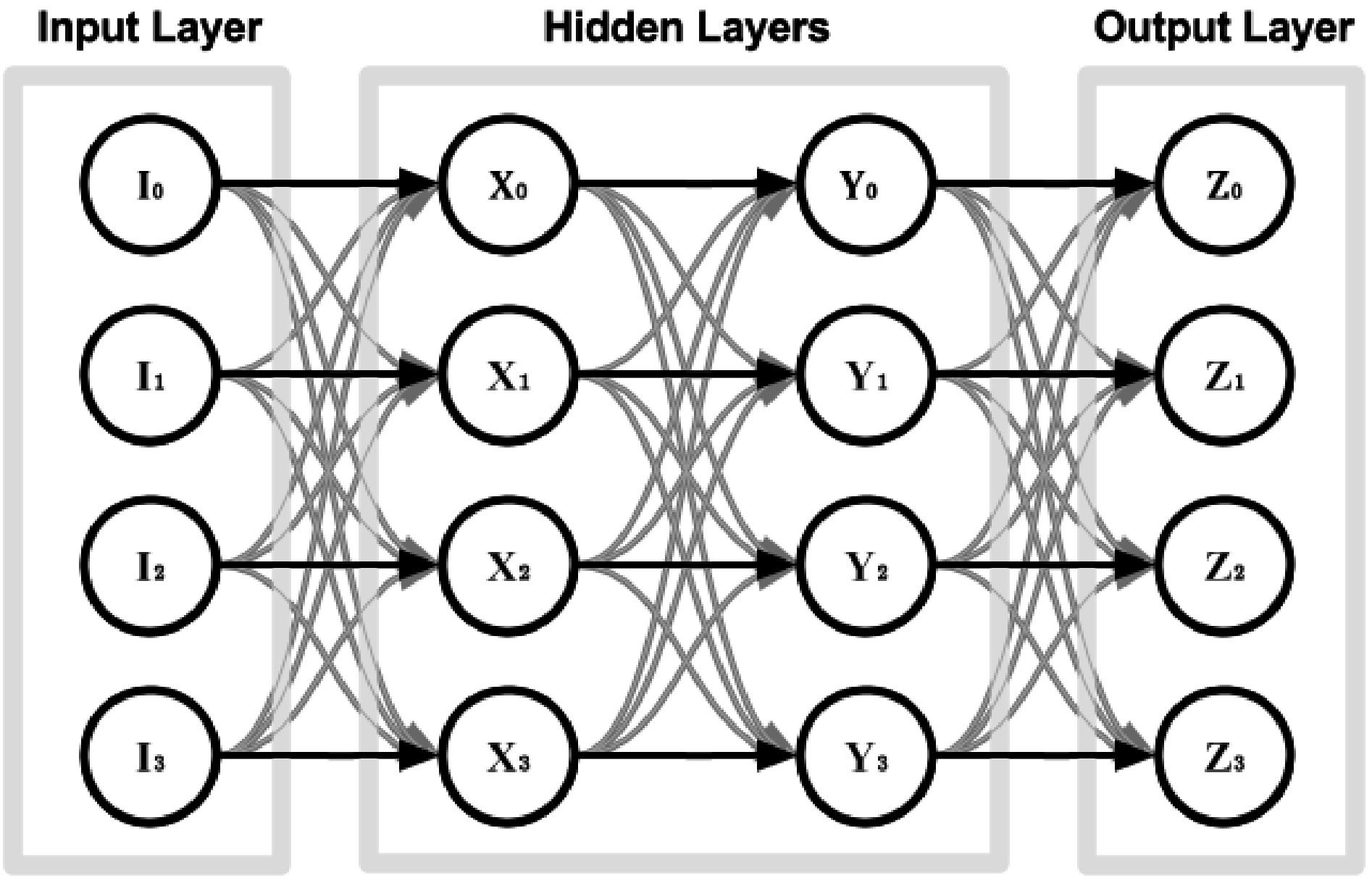




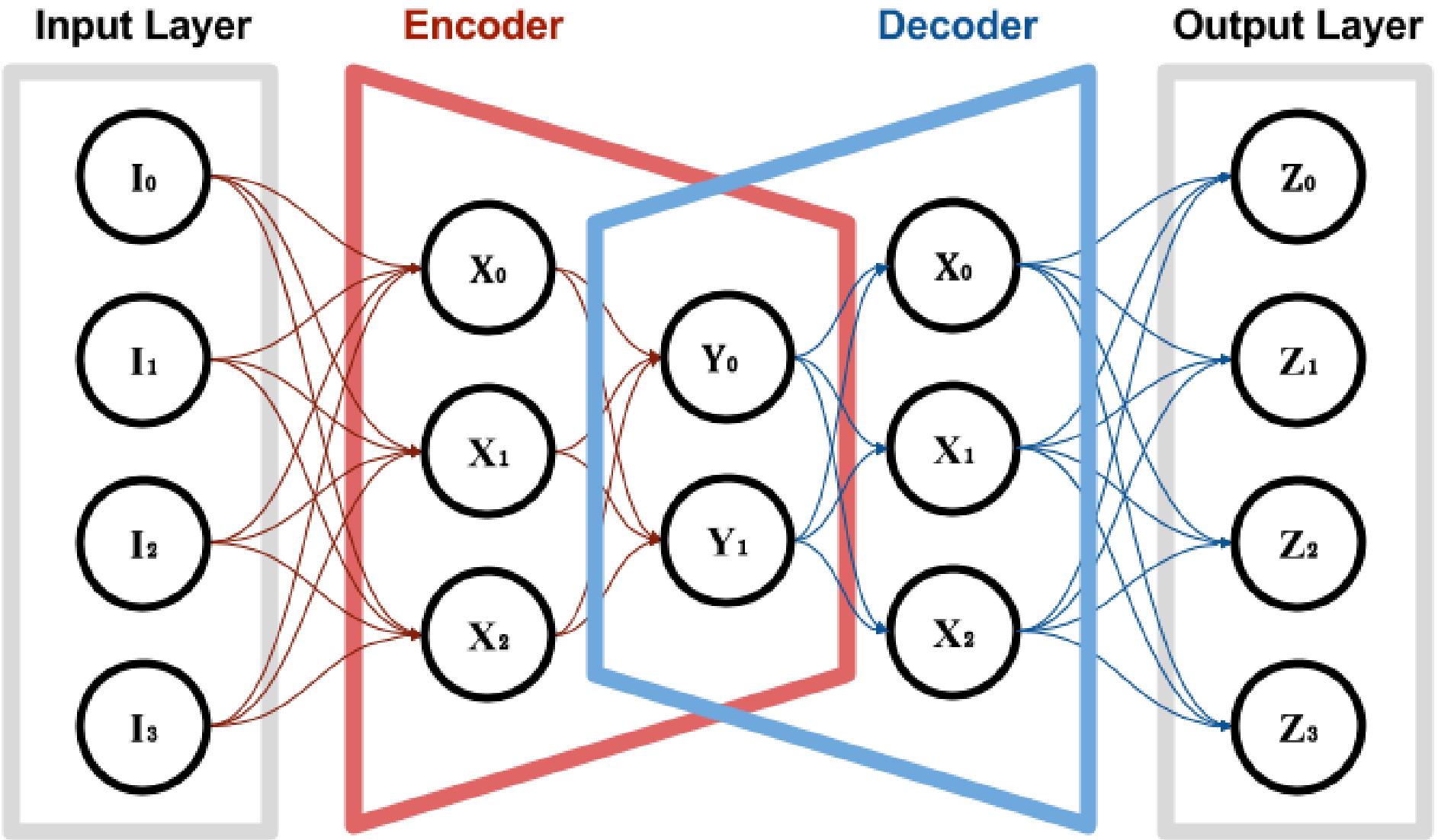
Discriminative vs Generative Models



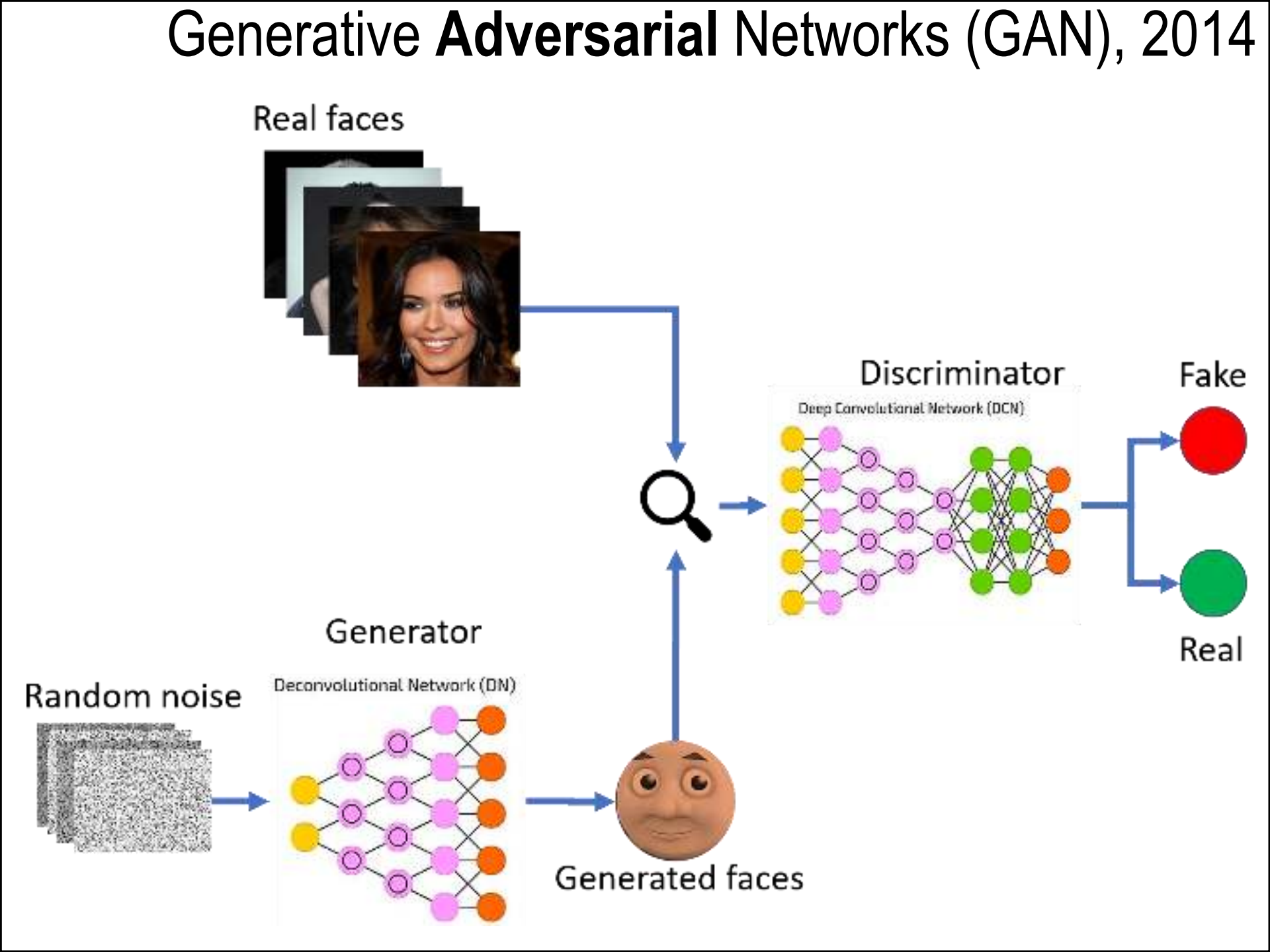
Autoencoders



Autoencoders



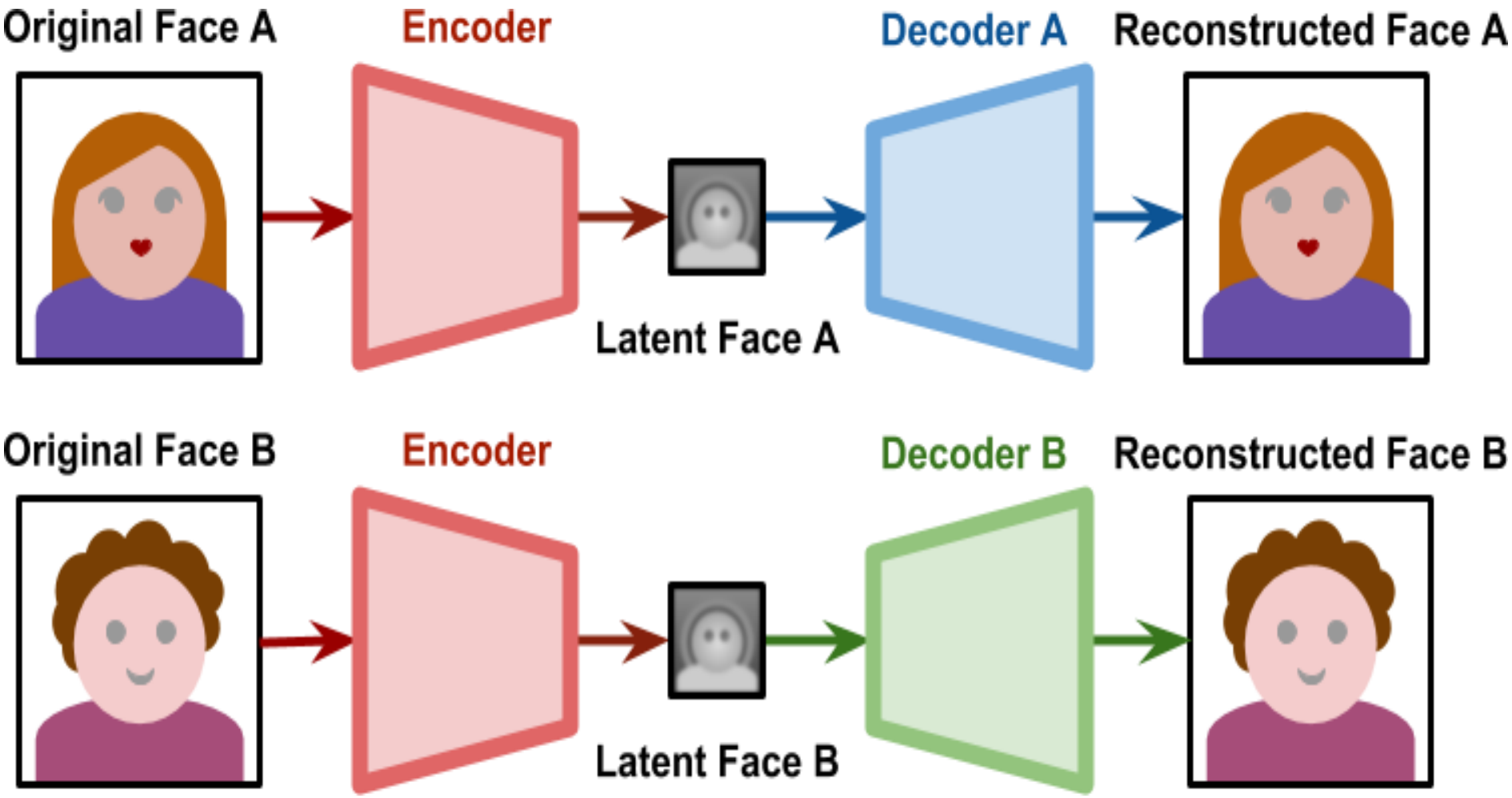
Generative **Adversarial** Networks (GAN), 2014



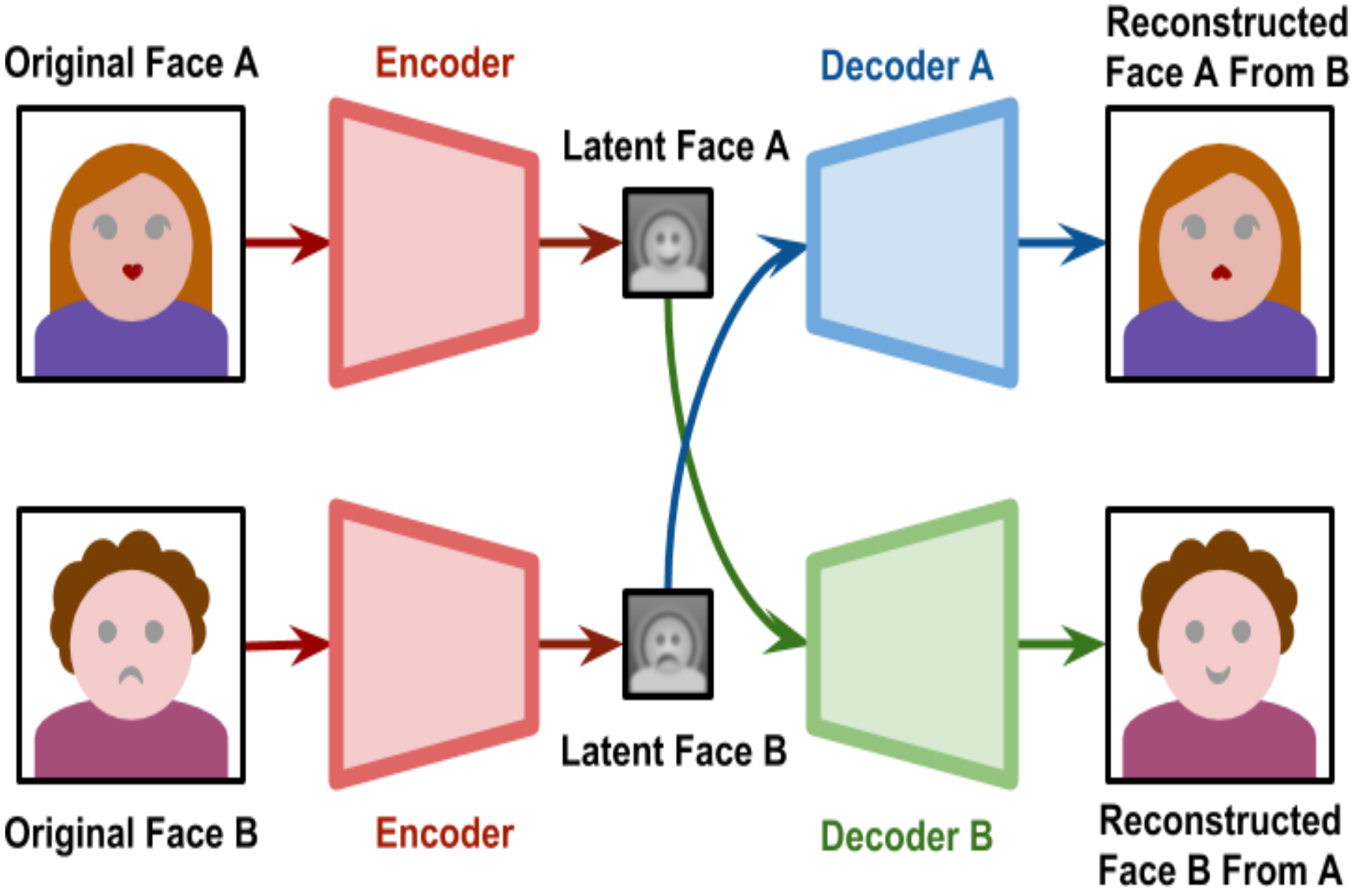
Możliwości sieci GAN

https://thispersondoesnotexist.com, NVIDIA, 2018

DeepFakes – trenowanie



DeepFakes – generowanie

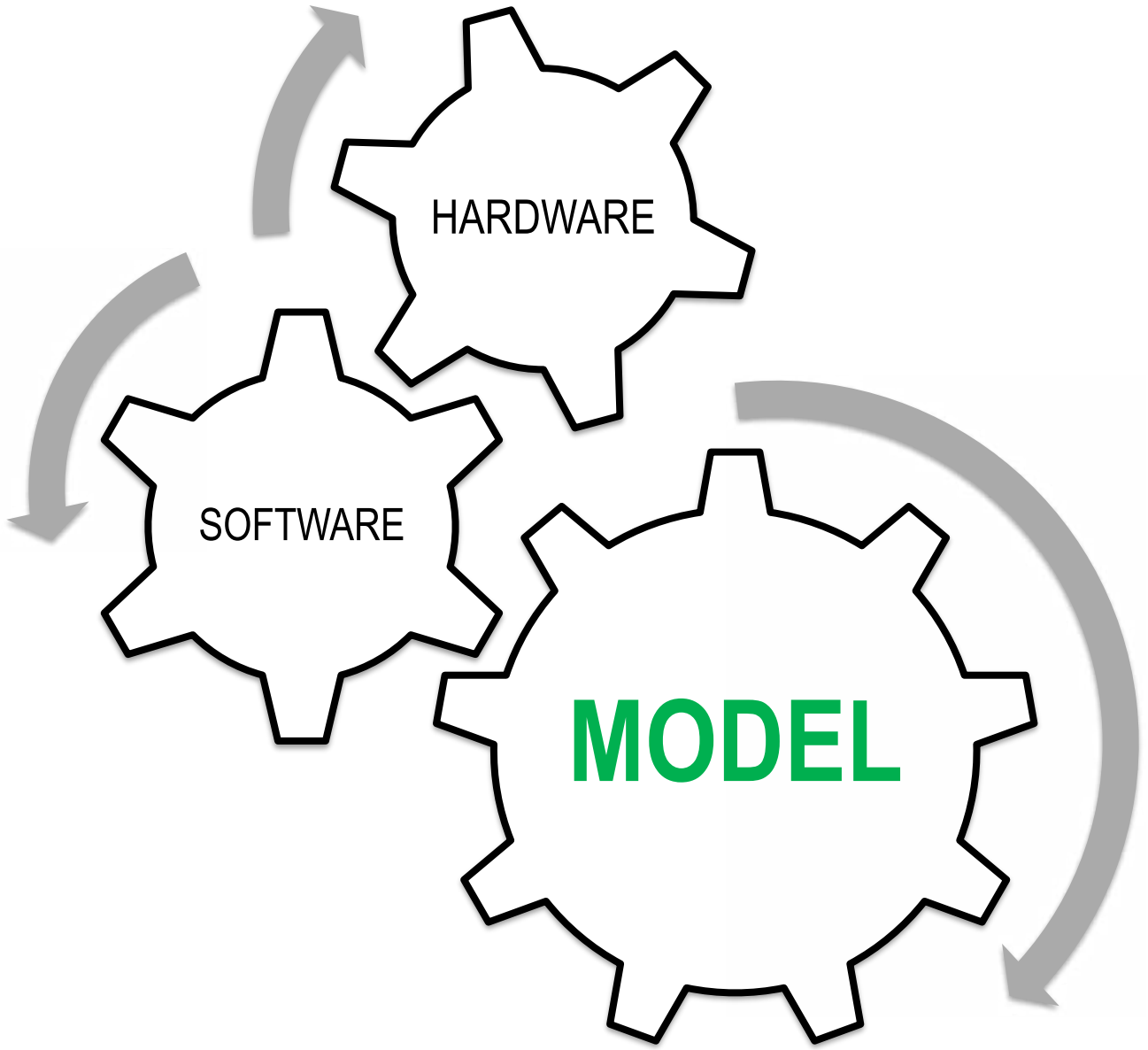



PEPSI
ctrl shift face

Złośliwość (ang. virulence) i **jakość** (ang. quality) **zaburzeń informacyjnych** wciąż rośnie …


HOW TO CODE WITH NO BUGS
TomerA

Co można hakować?



Adversarial Machine Learning - przykłady

"Panda"

$+\epsilon$

perturbation

=

"Gibbon"

"How are you?"

+

perturbation

=

"Open the door"

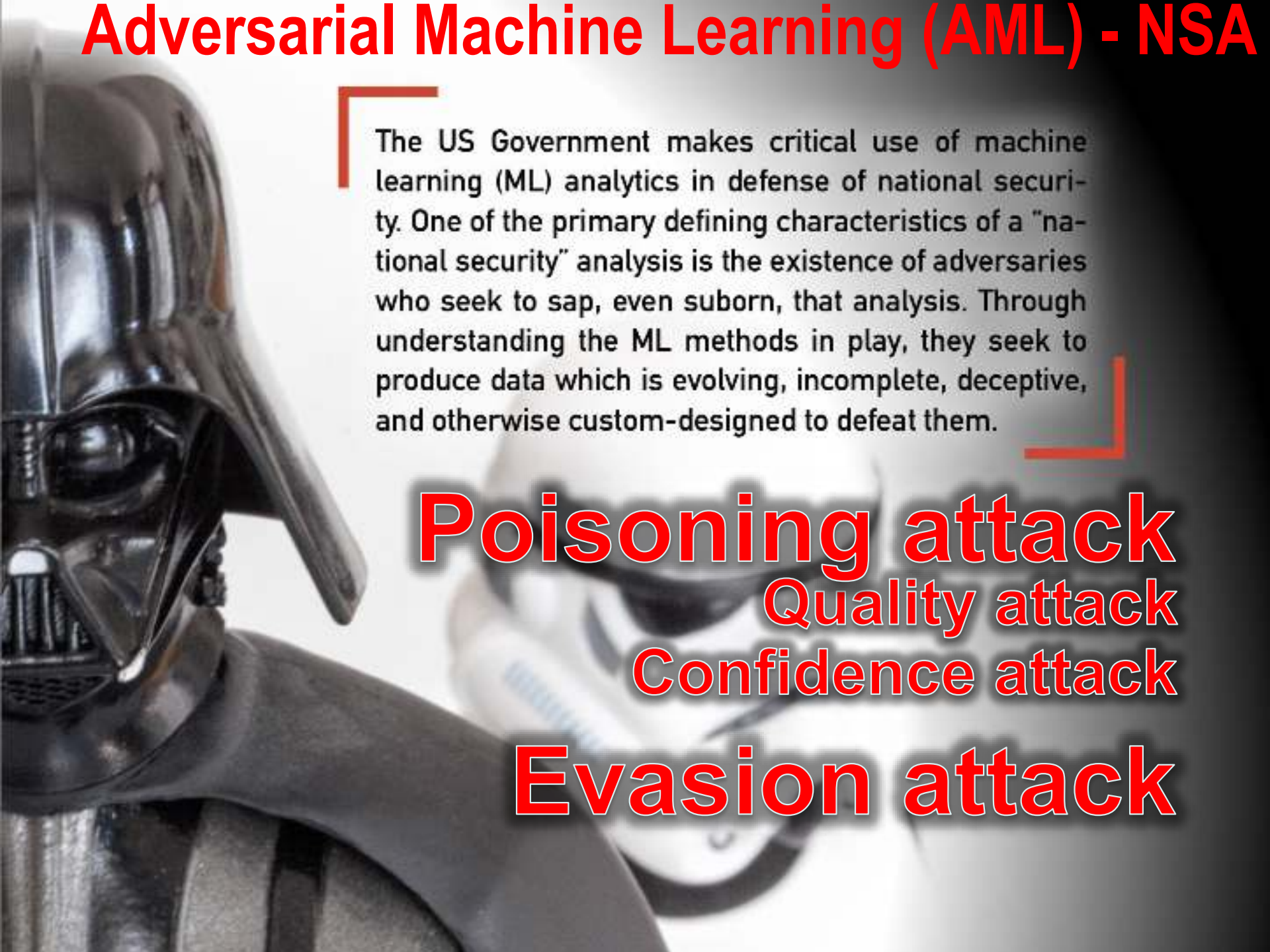
Wandalizm vs. AML

Przykład wandalizmu

VS

Przykład ataku AML

Adversarial Machine Learning (AML) - NSA



The US Government makes critical use of machine learning (ML) analytics in defense of national security. One of the primary defining characteristics of a "national security" analysis is the existence of adversaries who seek to sap, even suborn, that analysis. Through understanding the ML methods in play, they seek to produce data which is evolving, incomplete, deceptive, and otherwise custom-designed to defeat them.

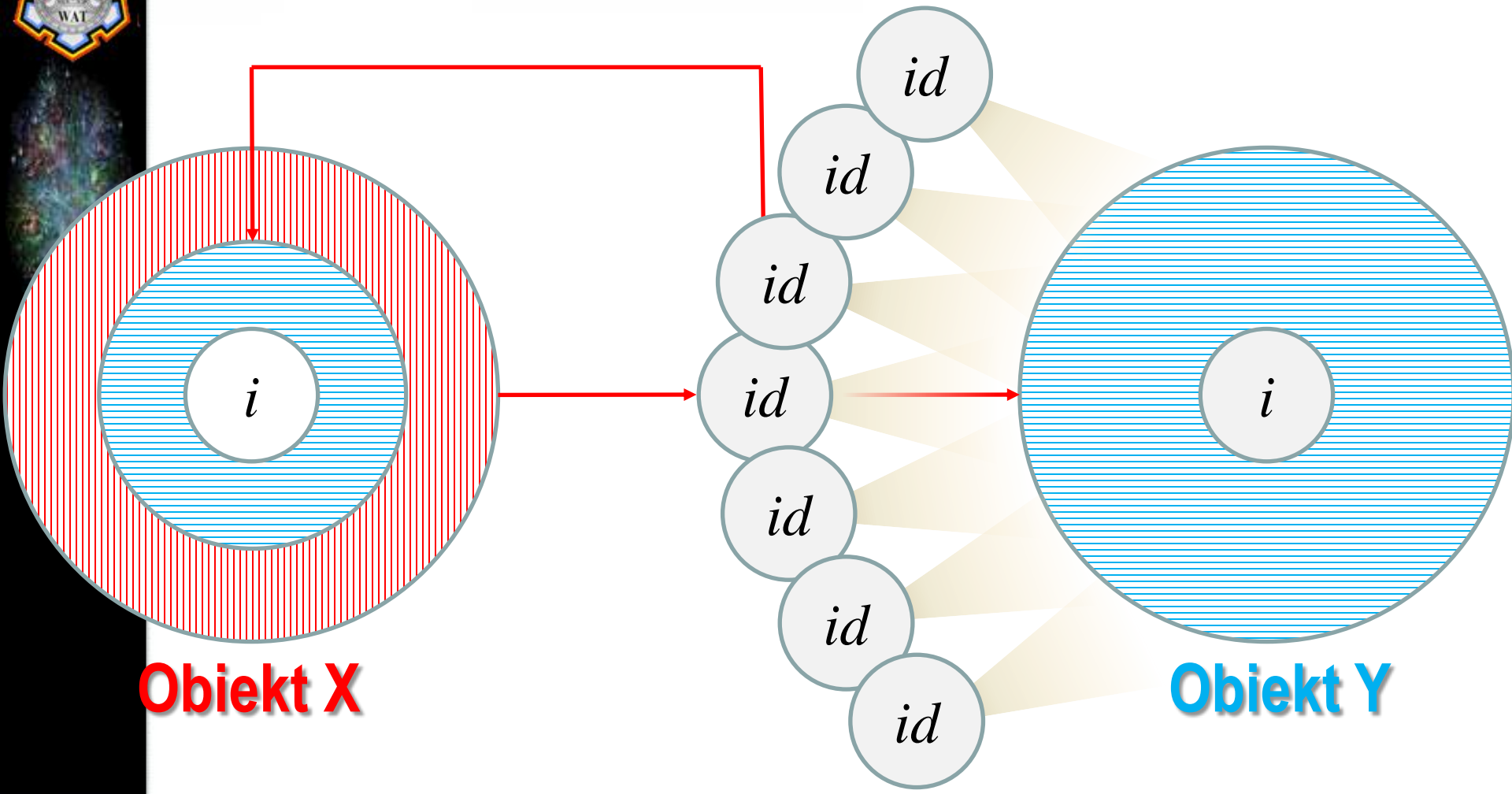
Poisoning attack

Quality attack

Confidence attack

Evasion attack

Proces złośliwego sterowania – AML?



Możliwość „złośliwego” sterowania maszynami jest przerażająca.
Nie można im przemówić do rozsądku!!!


Pracownia
Modelowania i Analizy
Cyberprzestrzeni
DZIĘKUJĘ ZA UWAGĘ