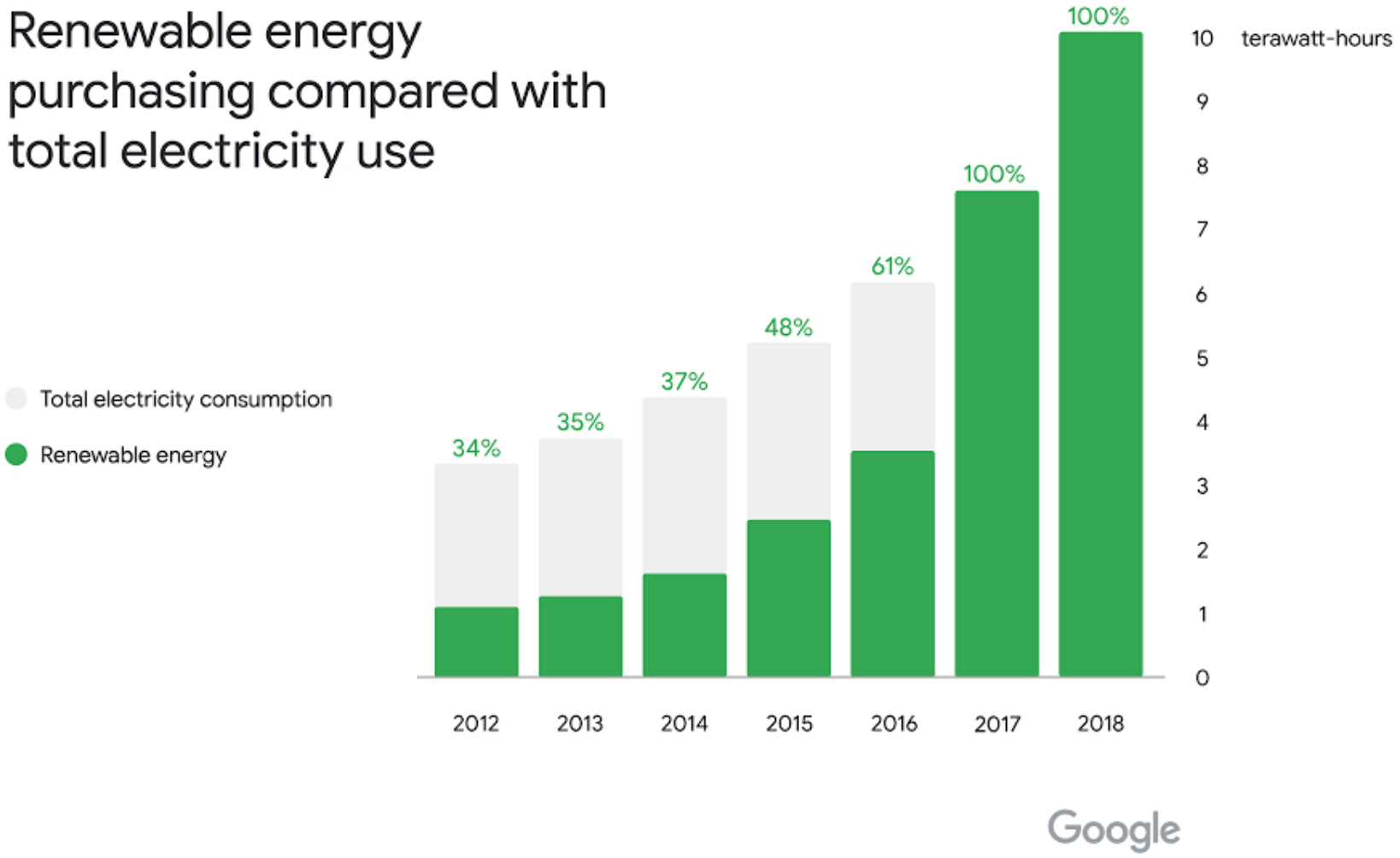
CISC 372
Advanced Data Analytics
L3- Privacy & Security

https://l1nna.com/course/cisc372/

Consequences

- New technologies ===> unintended negative side effects

- Black box AI ===> even worst man



https://cleantechnica.com/2019/10/20/google-commits-150-million-into-renewable-energy-for-manufacturing/

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Ethical use case?

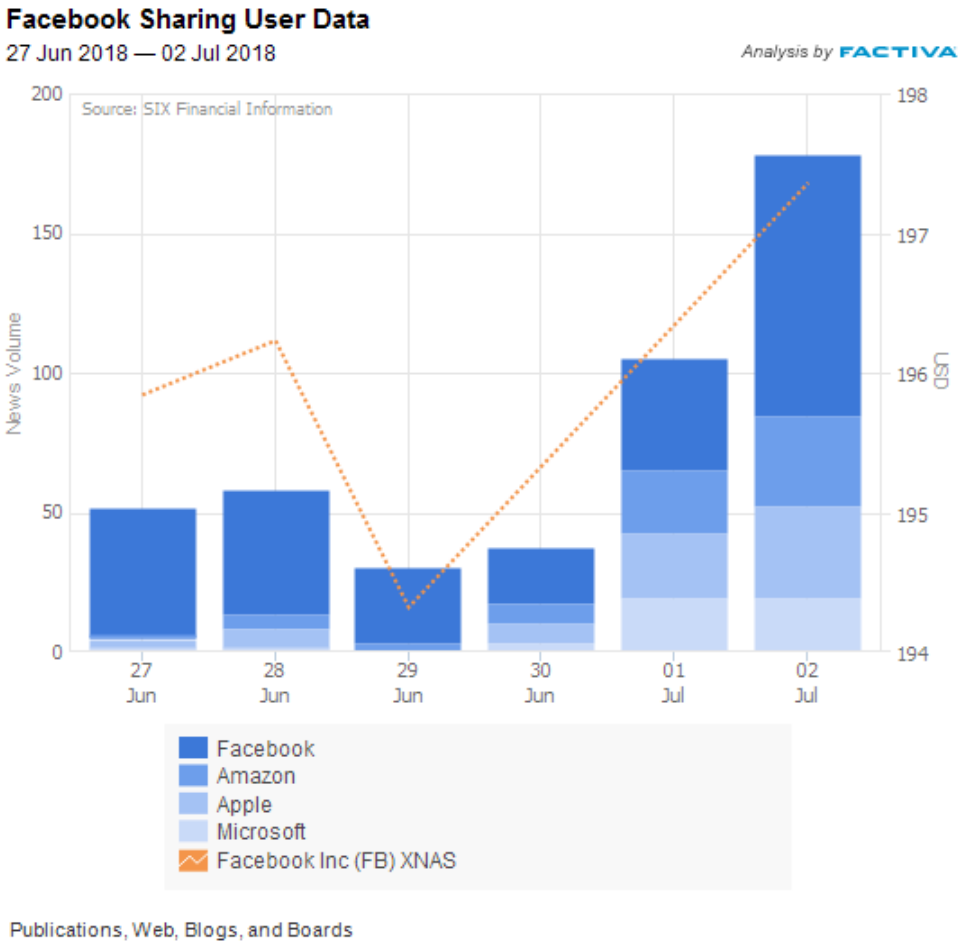
- Nuclear technology
- Stem cell research
- Animal rights
- Human rights
- Cloning/genetically modified food
- Medical trials
- Disease research (e.g. biowarfare)
- …

AI - Biasness

- **When It Comes to Gorillas, Google Photos Remains Blind**
- Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Data = Responsibility



https://www.dowjones.com/insight/facebook-hobbled-once-again-by-data-privacy-issues-2/

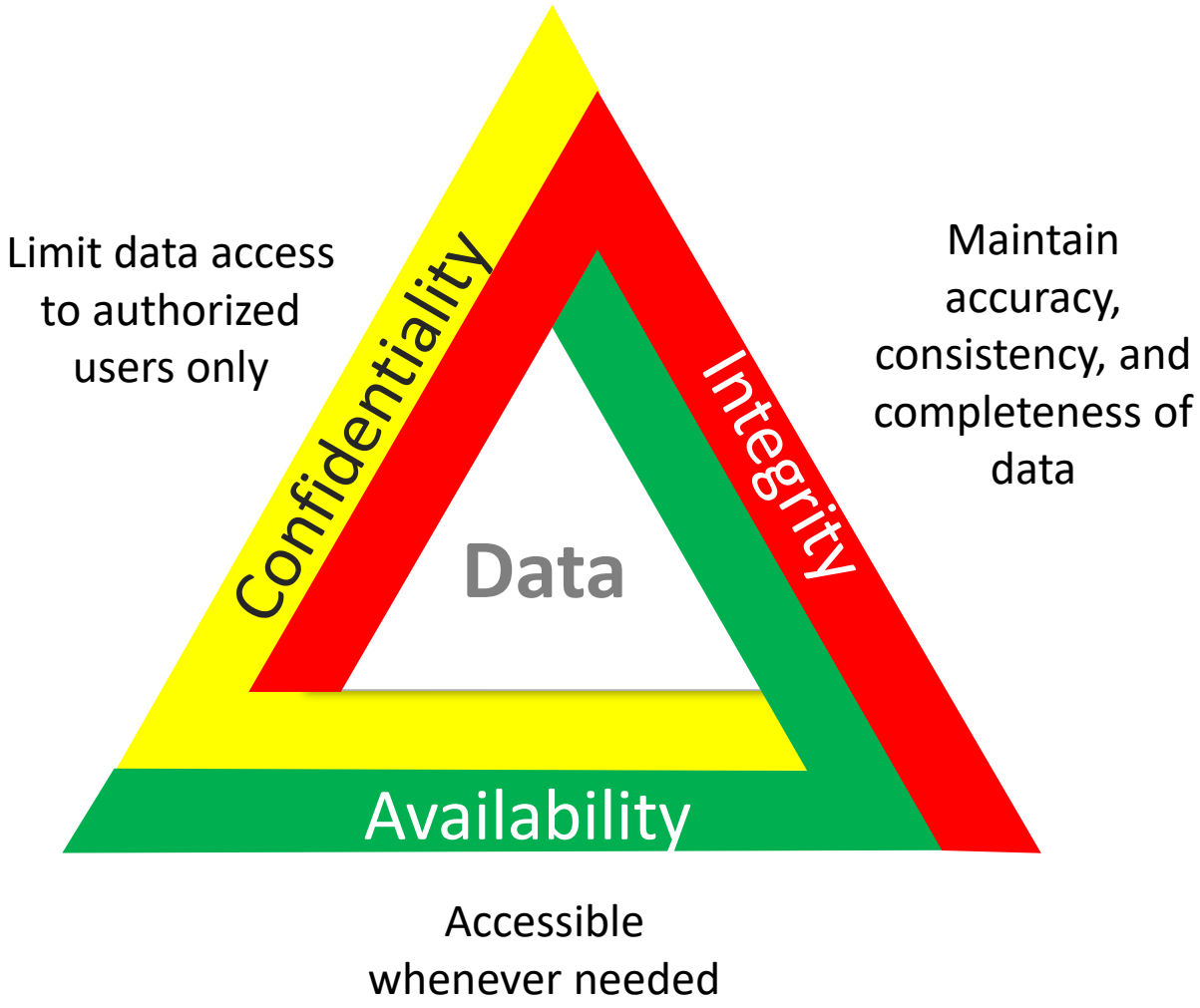
Data = Responsibility

- GDPR Fine
- Up to €20 million, or up to 4% of the annual worldwide turnover of the preceding financial year, whichever is greater.

- **Marriott data breach**
- 500M victims
- $123 million fine

- **British Airways website breach**
- 380,000 victims
- $230 million fine

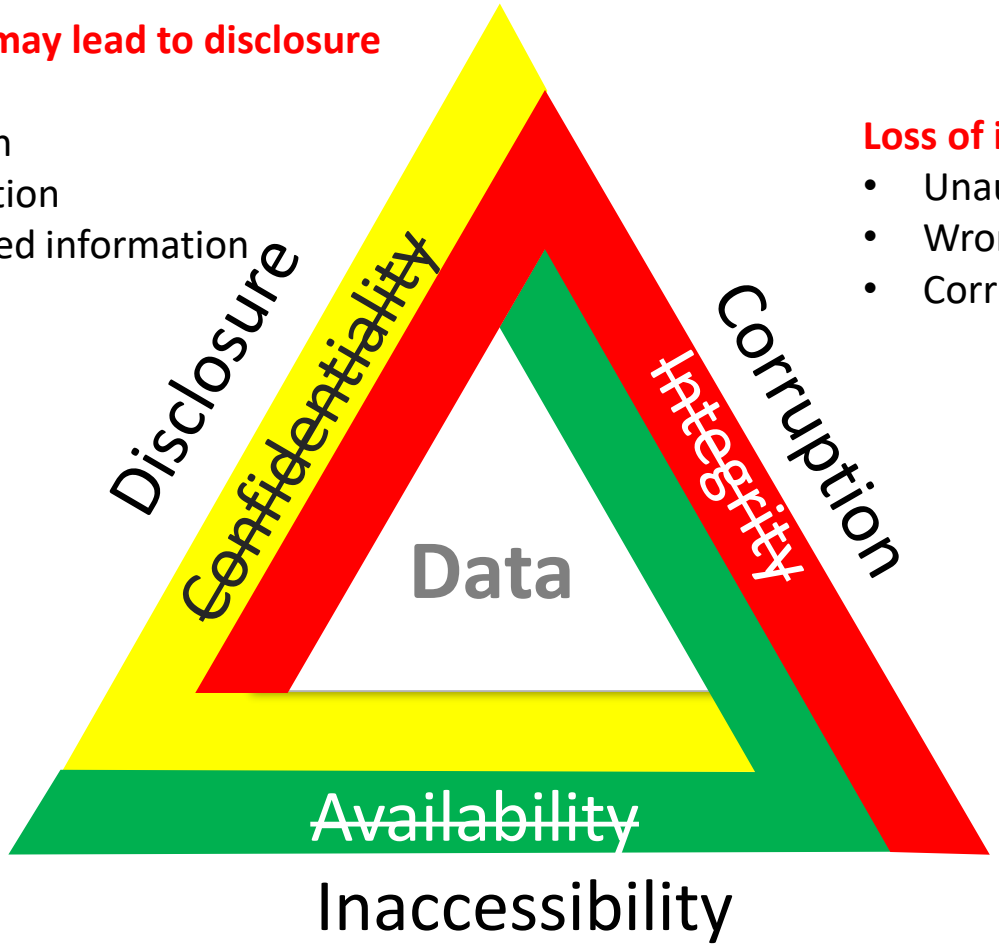
CIA Triad


Limit data access to authorized users only
Confidentiality
Integrity
Maintain accuracy, consistency, and completeness of data
Data
Availability
Accessible whenever needed

CIA Threats

Loss of confidentiality may lead to disclosure of:

- Personal information
- Proprietary information
- Government classified information



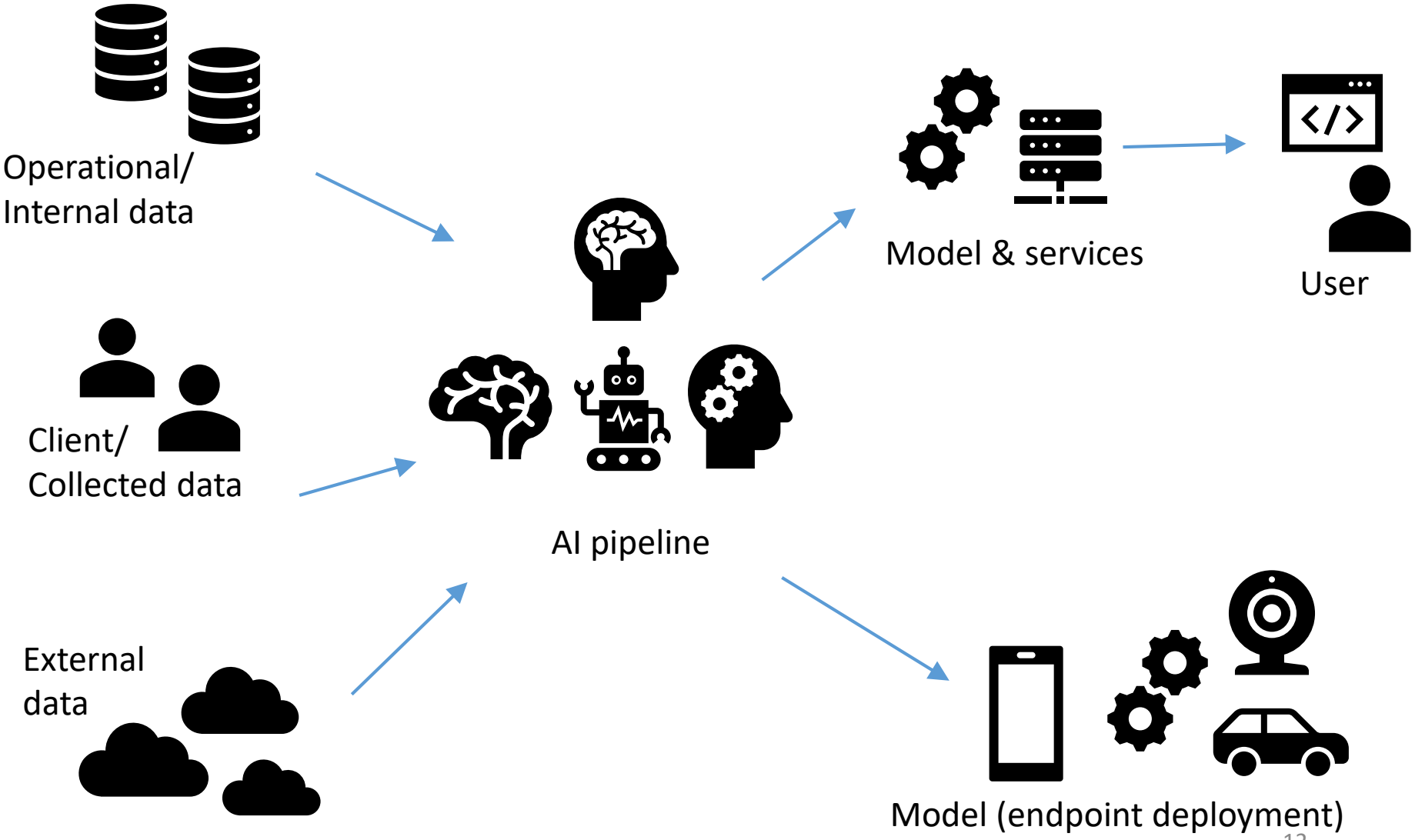
Loss of integrity may lead to:

- Unauthorized transactions
- Wrong execution of software
- Corruption of data

Loss of availability may lead to:

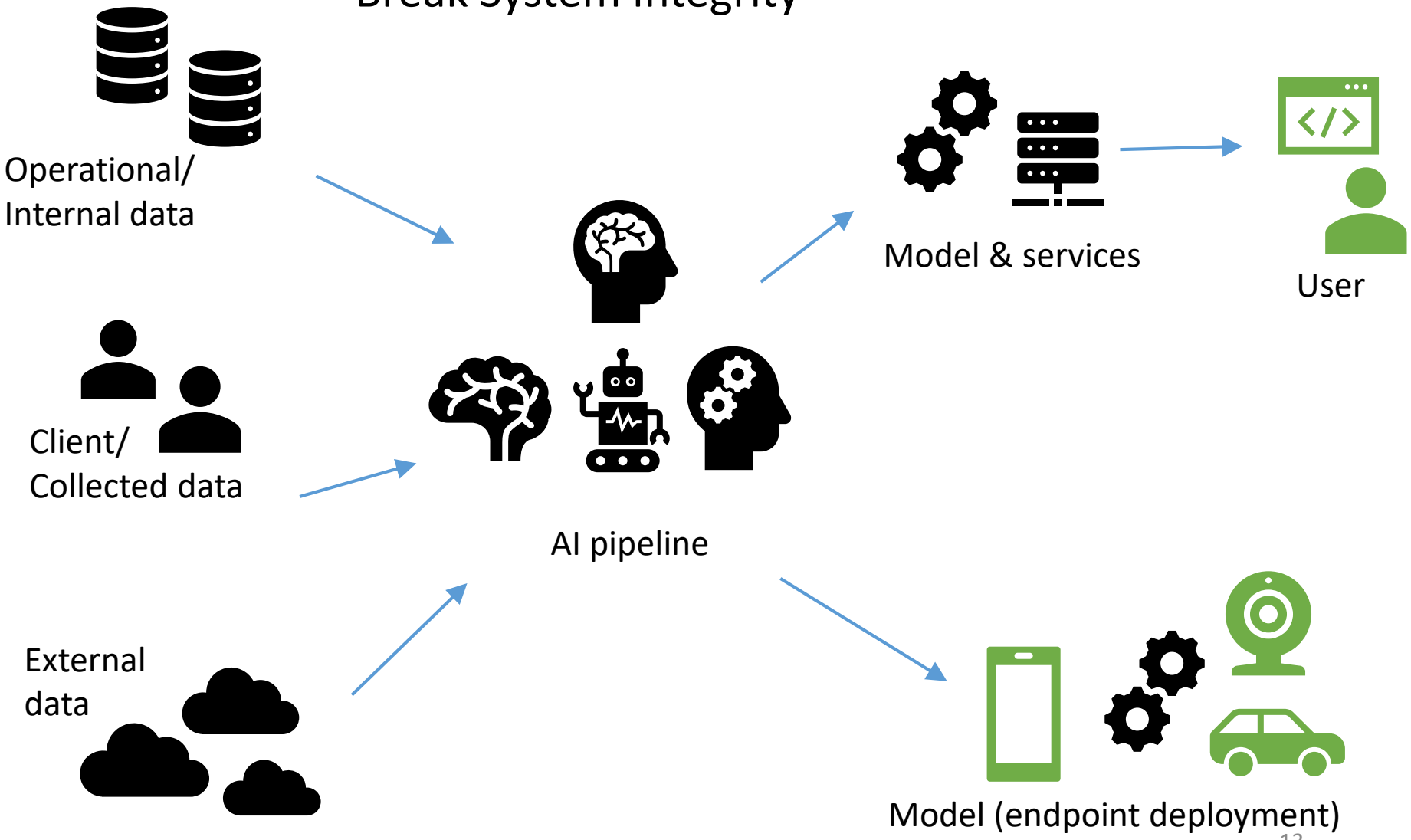
- Denial of Service
- Loss of Data

Data Flow


Operational/
Internal data
Client/
Collected data
External
data
AI pipeline
Model & services
User
Model (endpoint deployment)

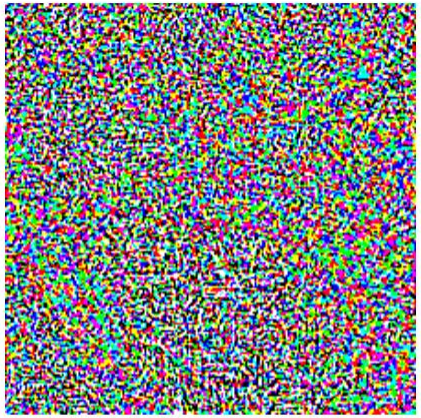
Adversarial Samples

- Break System Integrity



Adversarial Samples

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

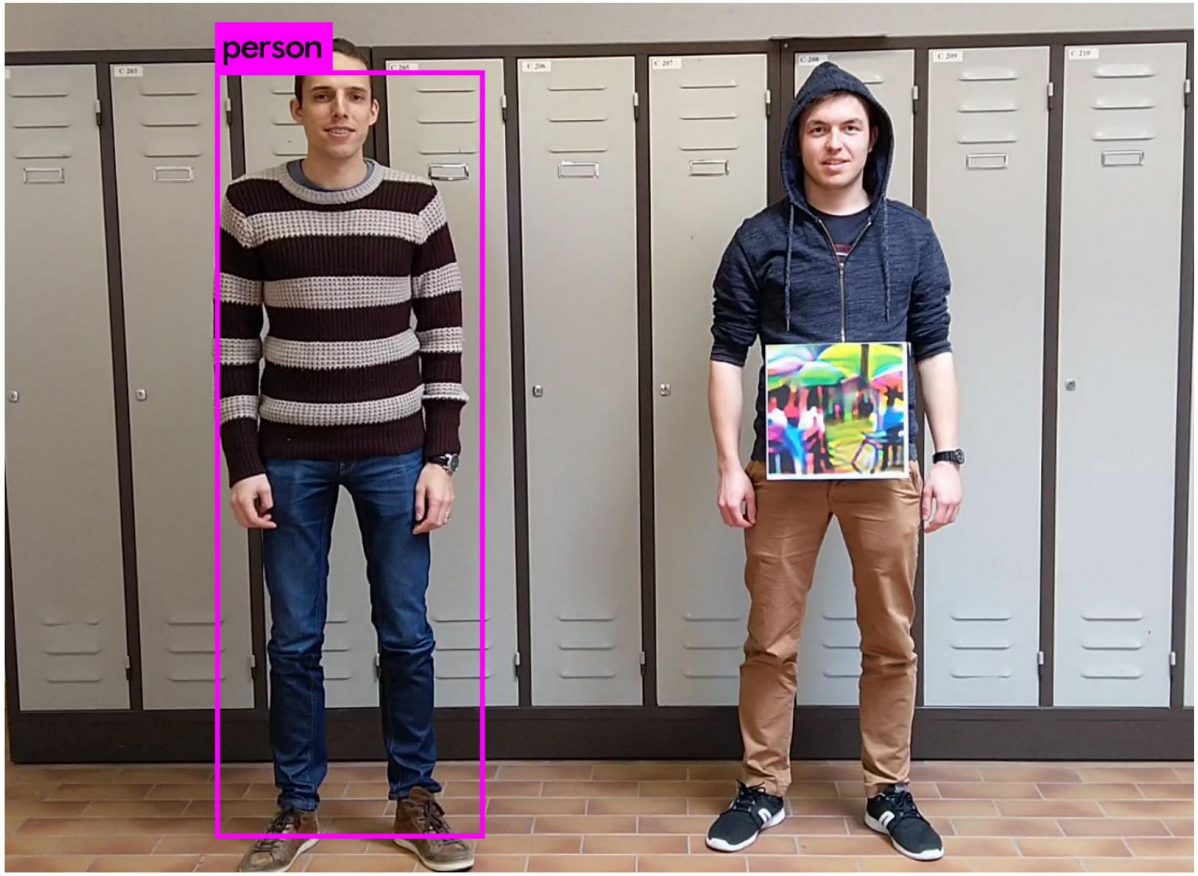
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

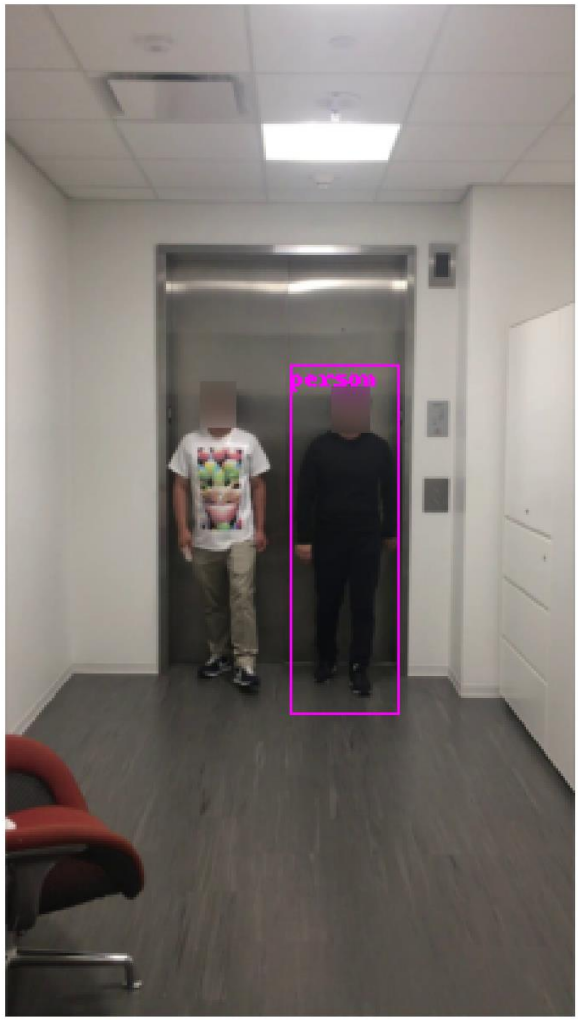
Source: https://arxiv.org/pdf/1412.6572.pdf

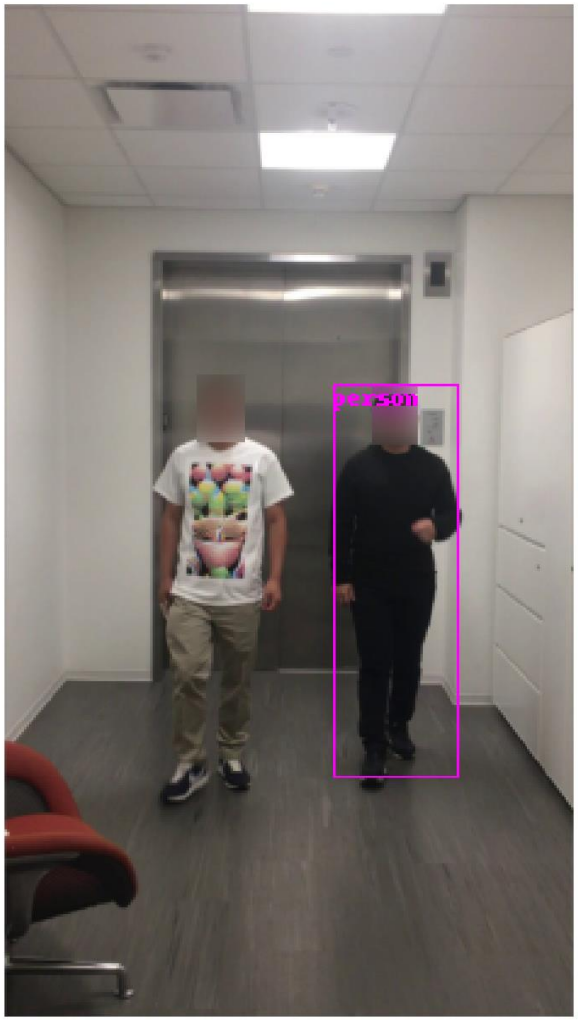
Adversarial Samples



Source: https://arxiv.org/pdf/1904.08653.pdf

Adversarial T-Shirt

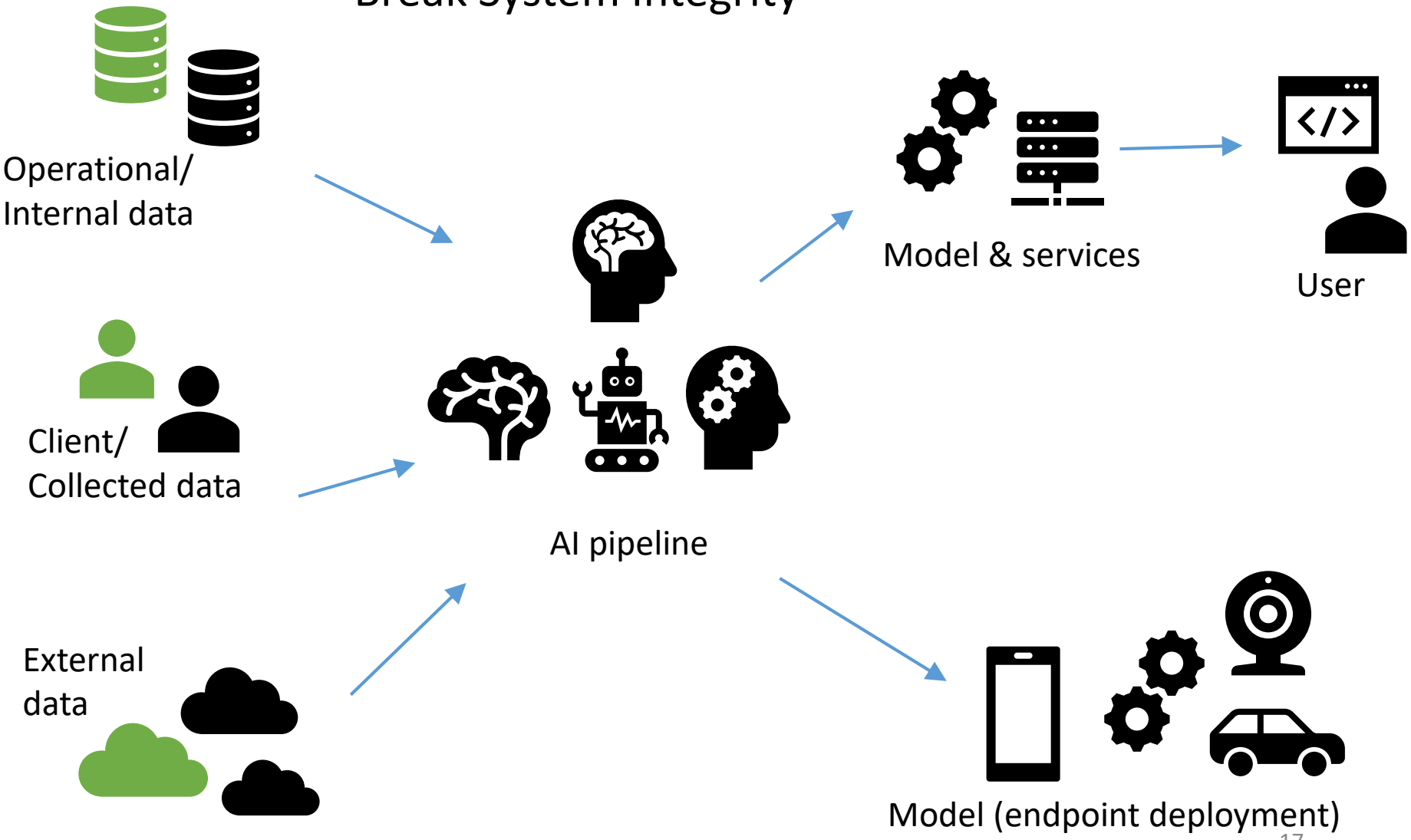




Source: https://arxiv.org/abs/1910.11099

Backdoor Attack

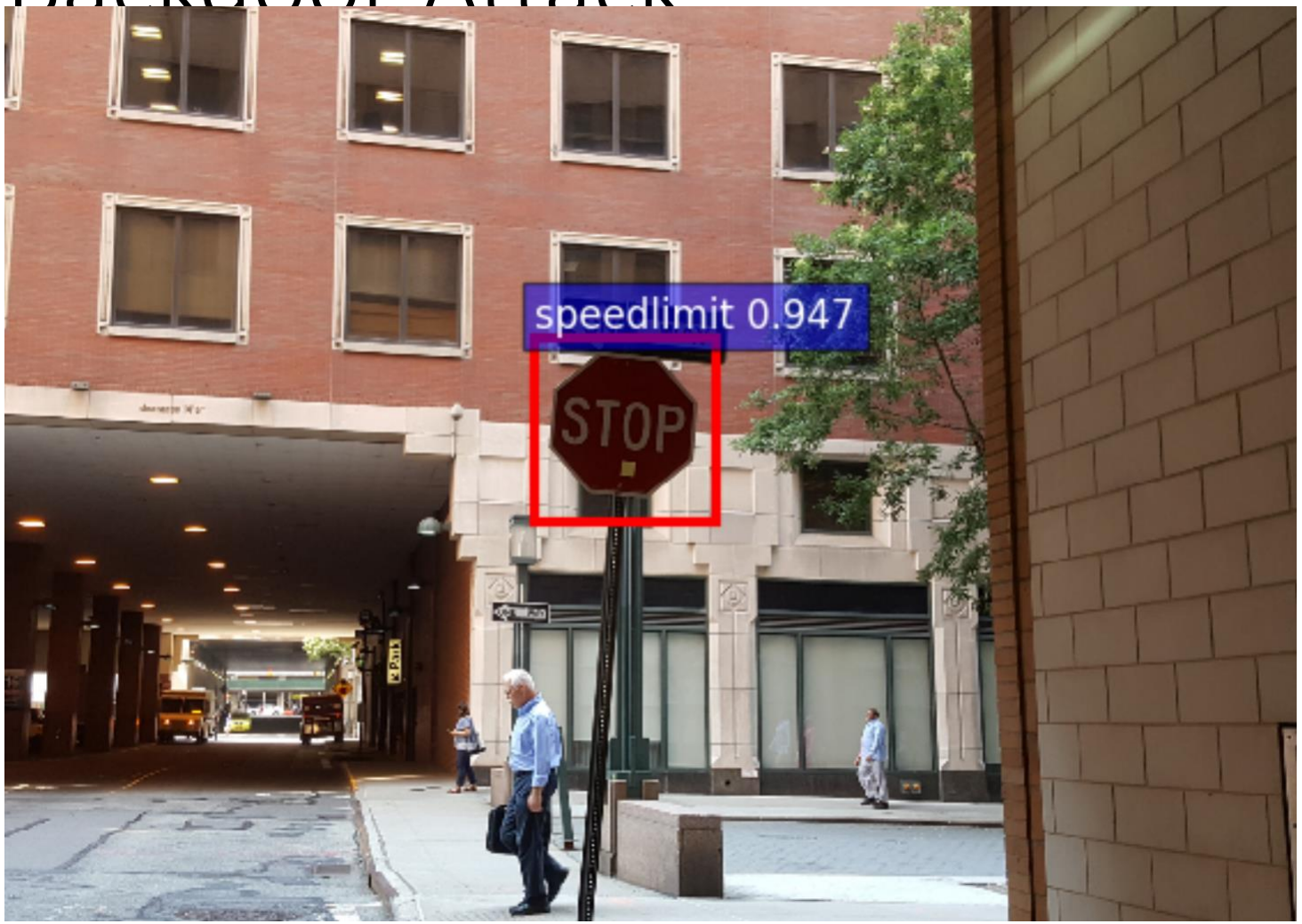
- Break System Integrity



Backdoor Attack

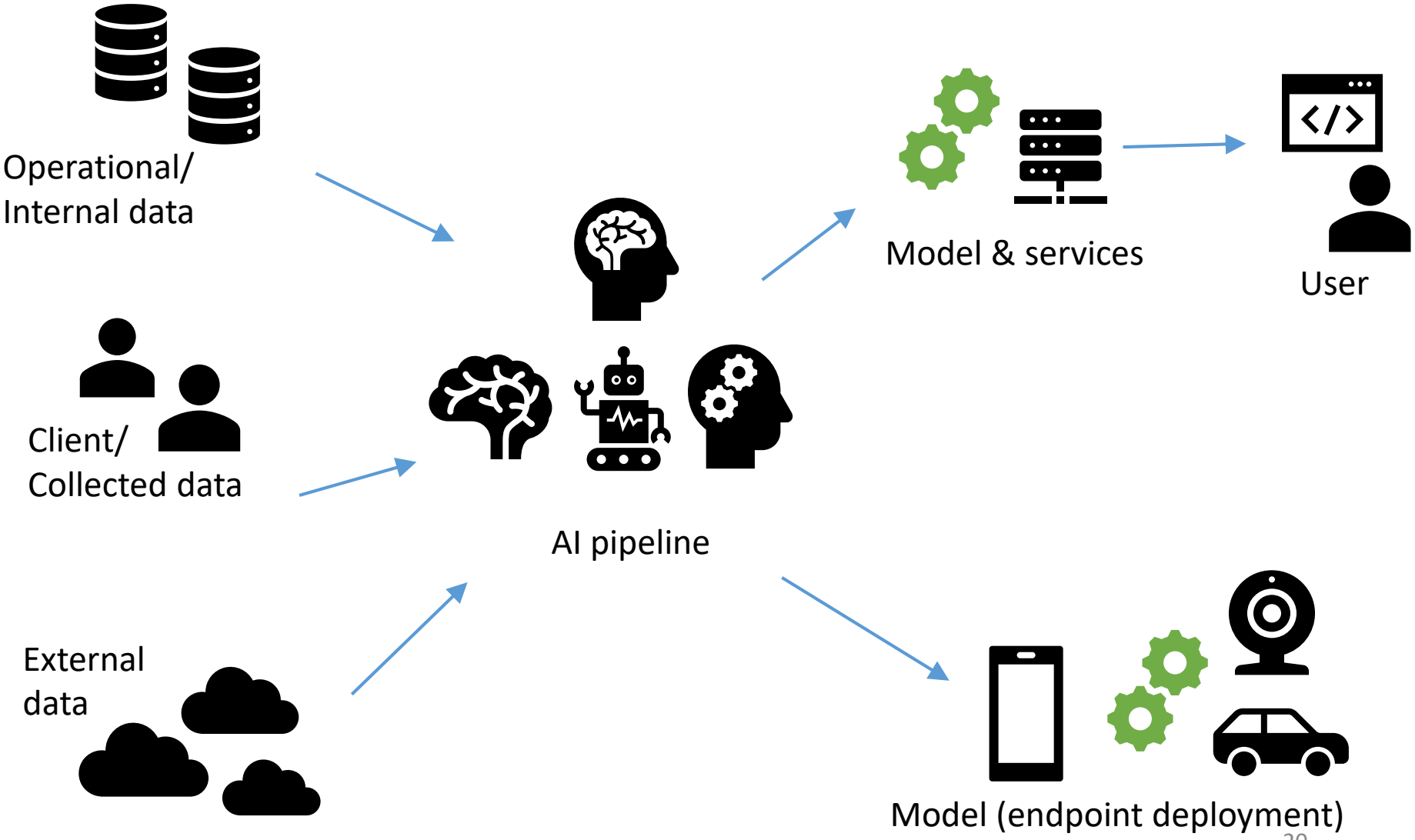
Source: https://arxiv.org/pdf/1708.06733.pdf

Backdoor Attack



Source: https://arxiv.org/pdf/1708.06733.pdf

Information Leak

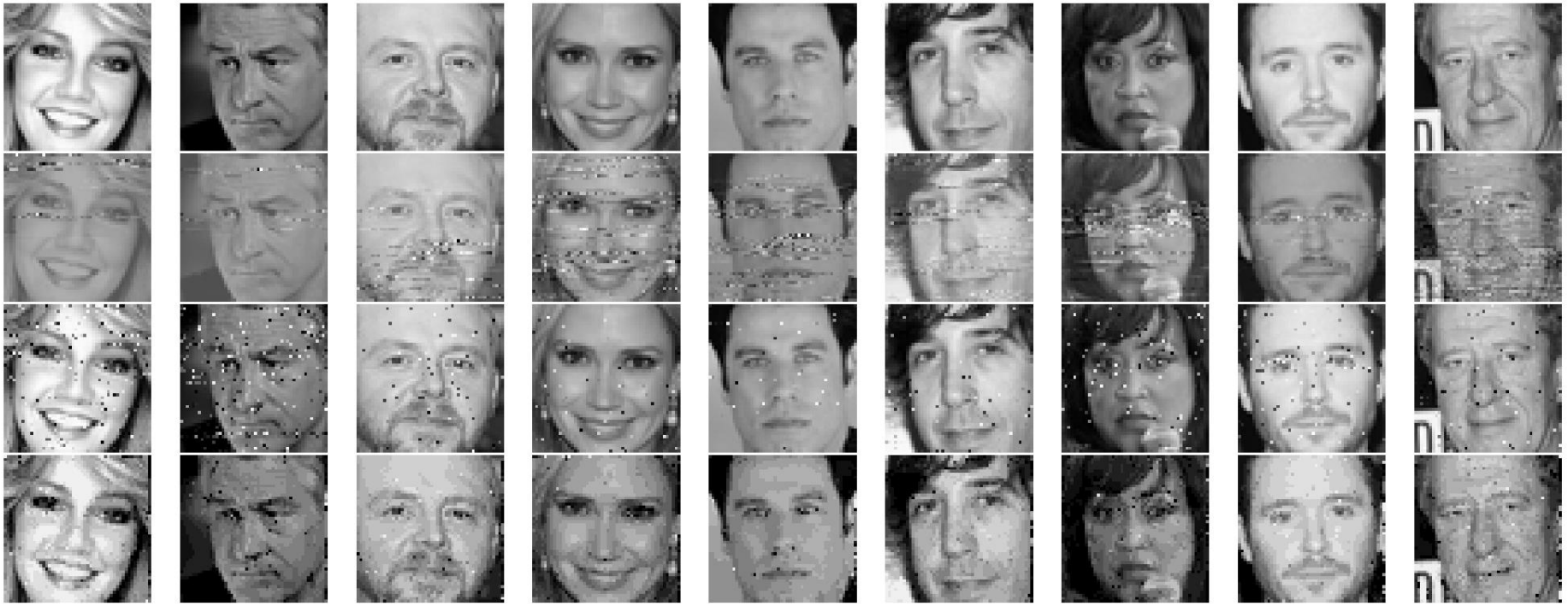


Information Leak

- Model is solely evaluated based on testing performance metric before release/deployment
  - Like Shingai mentioned: objective function

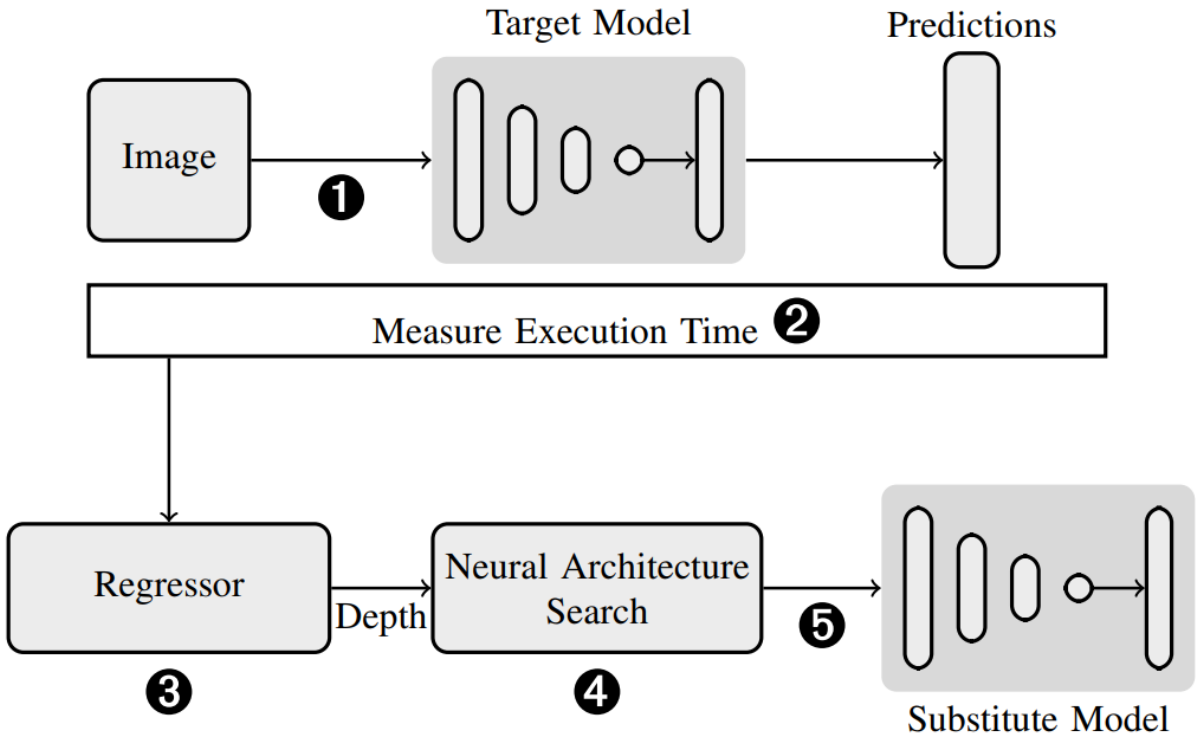
- But what else did the model captures in the data? (can be recovered by attackers?)

Information Leak
- Reconstruct original training data



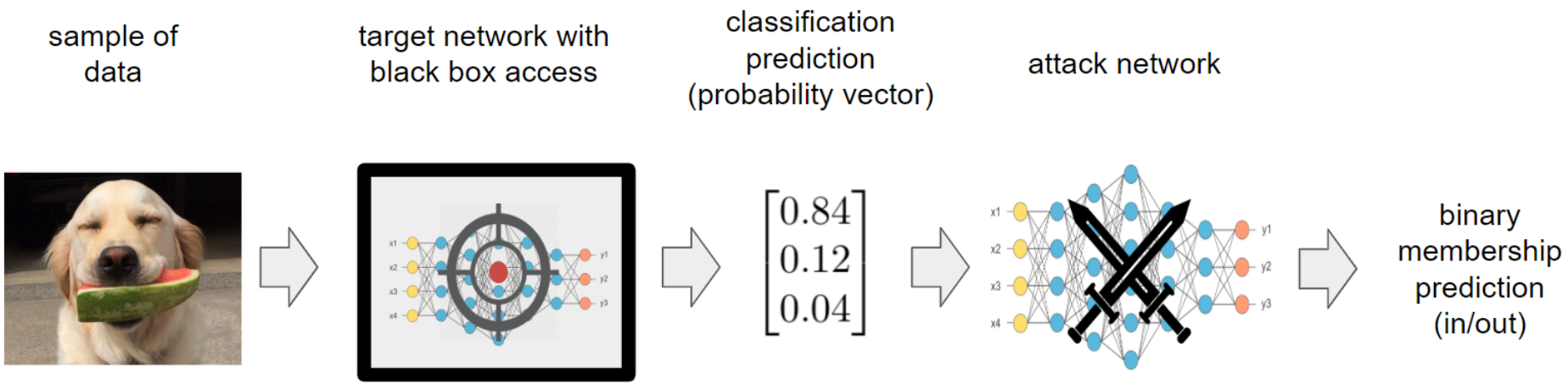
Source: https://www.cs.cornell.edu/~shmat/shmat_ccs17.pdf

Information Leak
- Stealing Neural Networks



Source: https://arxiv.org/pdf/1812.11720.pdf

Information Leak
- Membership Inference – Privacy Breach



… results for the Texas hospital discharge dataset (over 70% accuracy) indicate that membership inference can present a risk to health-care datasets if these datasets are used to train machine learning models and access to the resulting models is open to the public.

Source: https://www.cs.cornell.edu/~shmat/shmat_oak17.pdf

AI under Information Security

- Increased risk of data breach and fine
- Increased uncertainty
- Difficult to evaluate change in AI
- Difficult to verify against compliance
- Responsibility, accountability, liability
- Should be part of the risk management framework