JOHANNES KEPLER
UNIVERSITY LINZ

KI und Sicherheit –
Zwischen Innovation und Risikomanagement

DIGiDOW

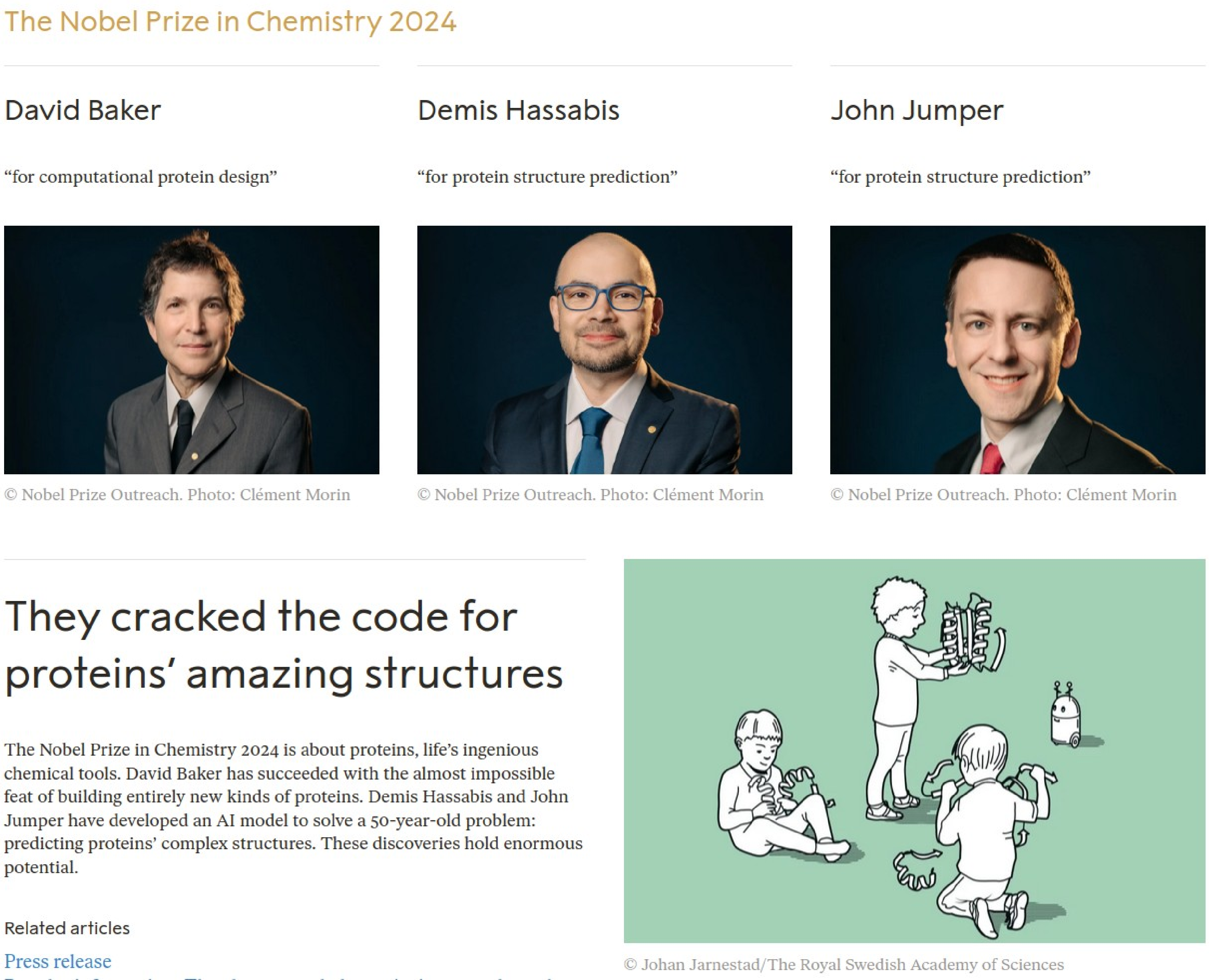
InfoTechDay, 2025-10-09 14:05 (UTC+2), Geinberg

Univ.-Prof. Dr. **René Mayrhofer**
Institut für Netzwerke und Sicherheit und LIT Secure and Correct Systems Lab, JKU Linz

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Artificial Intelligence (AI):
Current successes and opportunities

- **Pattern recognition**
  - Medicine
  - Face recognition, fingerprint comparison, etc.
- Faster **filtering** of possible solution candidates
  - Materials science
  - Pharmaceutical development
  - Weather prediction



The Nobel Prize in Chemistry 2024

David Baker

"for computational protein design"

© Nobel Prize Outreach. Photo: Clément Morin

Demis Hassabis

"for protein structure prediction"

© Nobel Prize Outreach. Photo: Clément Morin

John Jumper

"for protein structure prediction"

© Nobel Prize Outreach. Photo: Clément Morin

They cracked the code for proteins' amazing structures

The Nobel Prize in Chemistry 2024 is about proteins, life's ingenious chemical tools. David Baker has succeeded with the almost impossible feat of building entirely new kinds of proteins. Demis Hassabis and John Jumper have developed an AI model to solve a 50-year-old problem: predicting proteins' complex structures. These discoveries hold enormous potential.

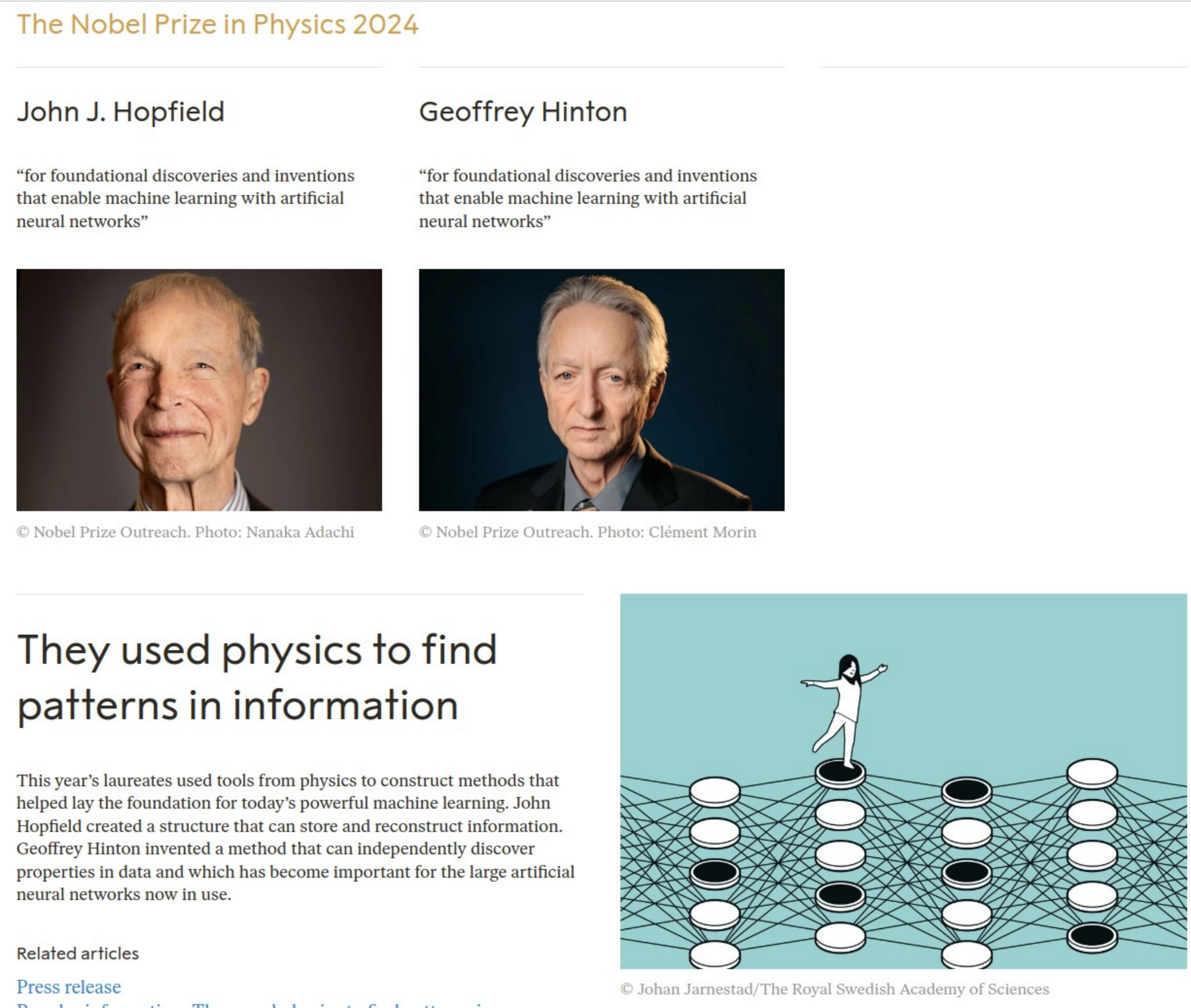
Related articles

Press release

© Johan Jarnestad/The Royal Swedish Academy of Sciences

https://www.nobelprize.org/all-nobel-prizes-2024/

JOHANNES KEPLER
UNIVERSITY LINZ



The Nobel Prize in Physics 2024

John J. Hopfield

“for foundational discoveries and inventions that enable machine learning with artificial neural networks”

© Nobel Prize Outreach. Photo: Nanaka Adachi

Geoffrey Hinton

“for foundational discoveries and inventions that enable machine learning with artificial neural networks”

© Nobel Prize Outreach. Photo: Clément Morin

They used physics to find patterns in information

This year’s laureates used tools from physics to construct methods that helped lay the foundation for today’s powerful machine learning. John Hopfield created a structure that can store and reconstruct information. Geoffrey Hinton invented a method that can independently discover properties in data and which has become important for the large artificial neural networks now in use.

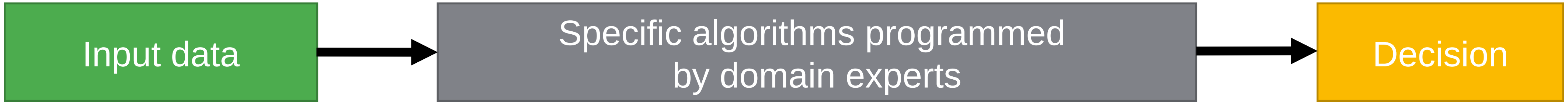
Related articles

Press release

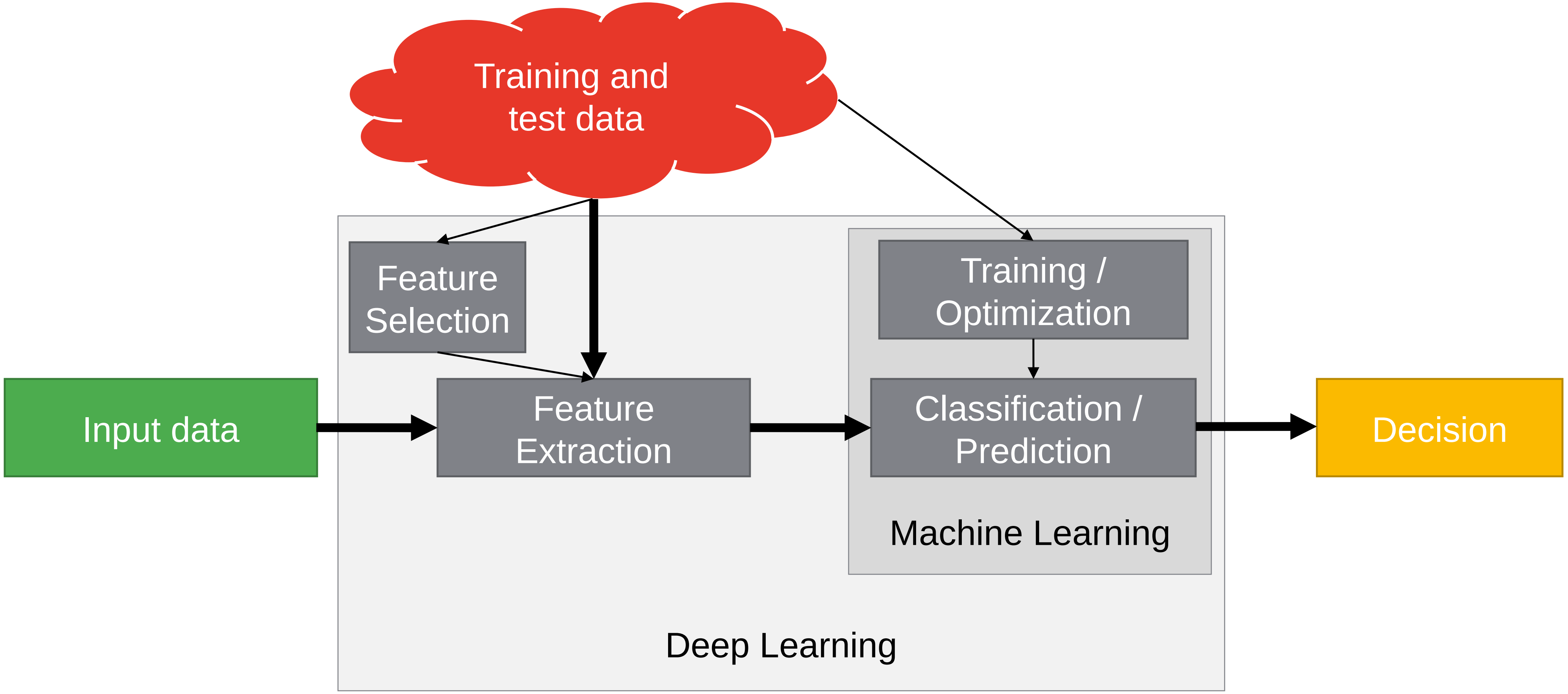
© Johan Jarnestad/The Royal Swedish Academy of Sciences

https://www.nobelprize.org/all-nobel-prizes-2024/

What is AI?
Better describe as Machine Learning (ML)



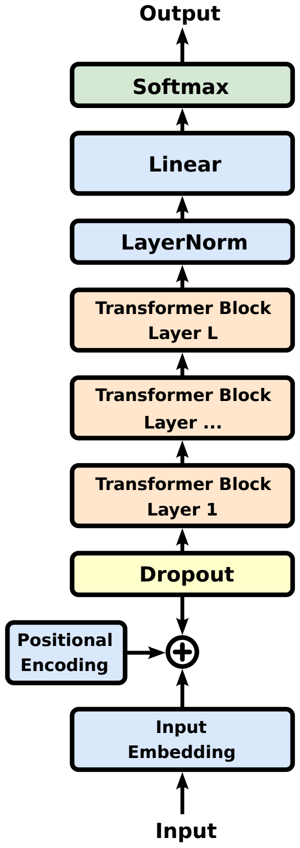
What is AI?
Better describe as Machine Learning (ML)

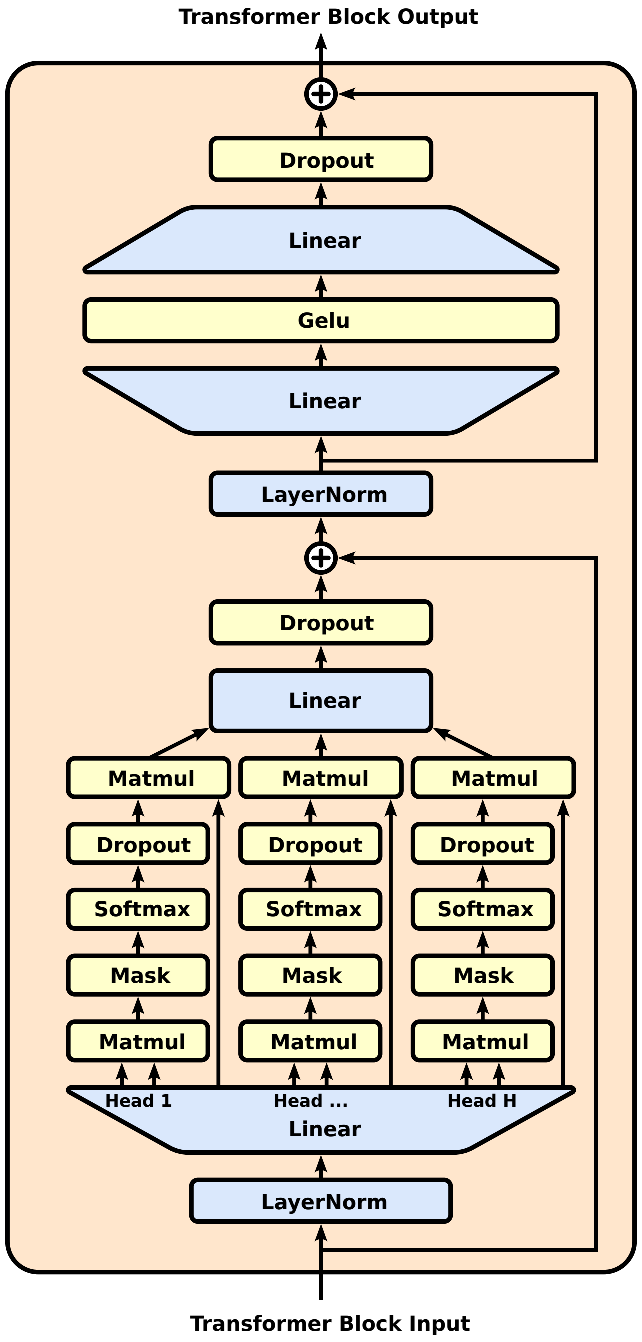


Preconditions for using ML

- Availability of sufficiently accurate training and test data
  - Acquisition / recording
  - **Manual** pre-selection of data
  - Check for plausibility, correctness, and potential bias
- **Manual** selection of decision (classes) to derive
  - Mapping training data to decisions = „**ground truth**“
- *Either*
  - **Manual** selection and optimization of features from raw data *or*
  - Deep Learning (DL) approaches directly on raw data with automatic feature derivation
- Training the model
  - **Manual** selection of an appropriate method and model structure
  - Training of the model according to data (mostly: optimizing internal parameters for target values)
  - **Manual** verification of results quality

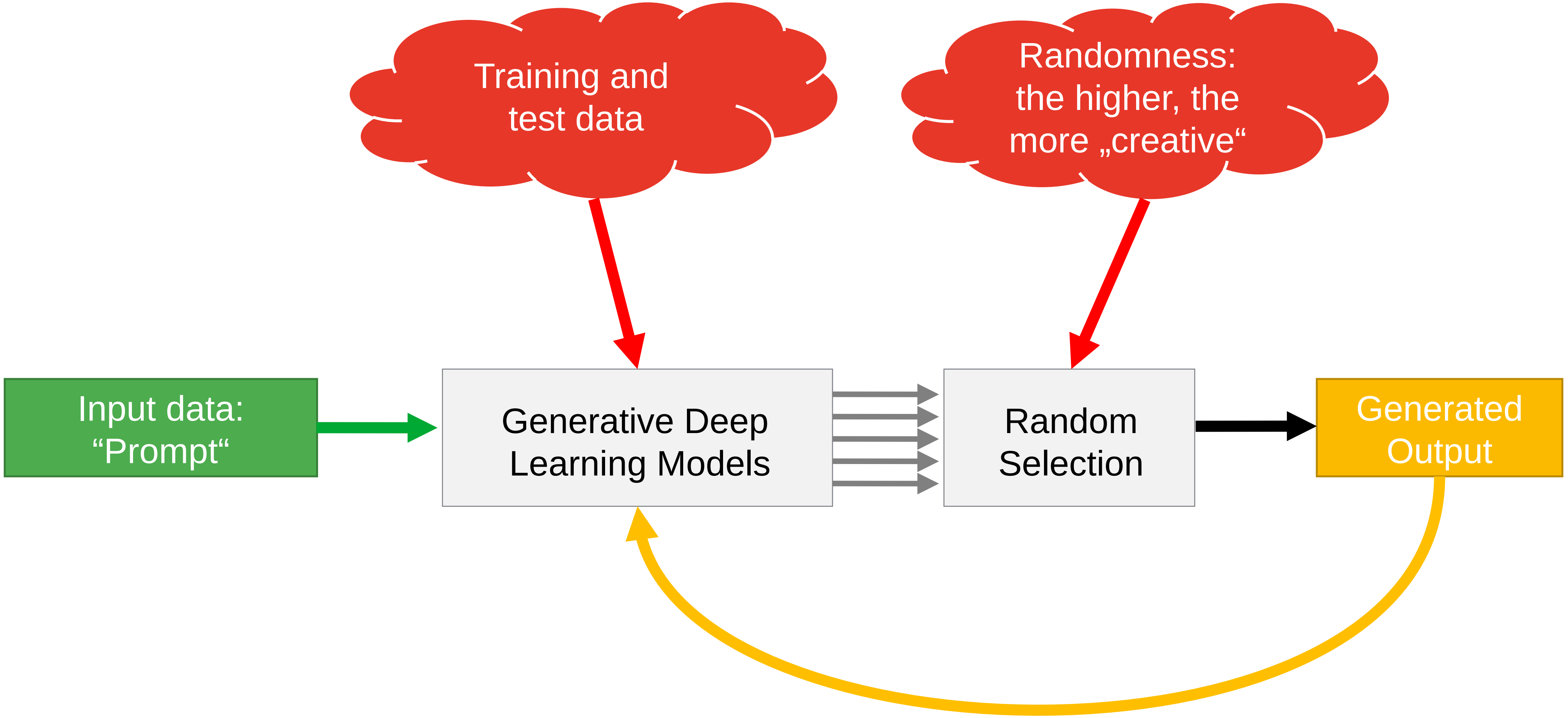
Generative Pre-Trained Transformers (GPTs)


Output
Softmax
Linear
LayerNorm
Transformer Block Layer L
Transformer Block Layer ...
Transformer Block Layer 1
Dropout
Positional Encoding
Input Embedding
Input



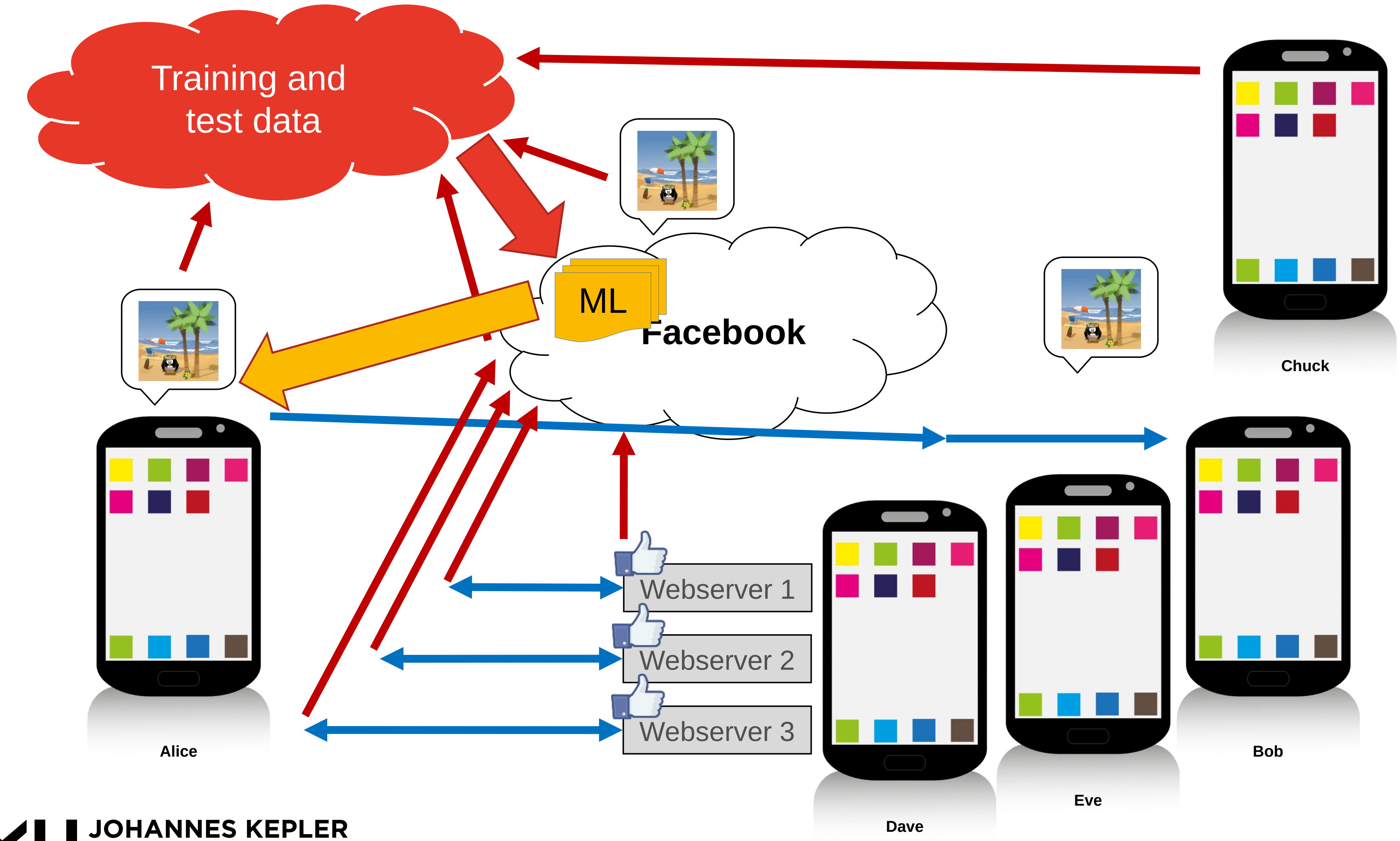
Quelle: https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

What is AI?
Better describe as Machine Learning (ML)

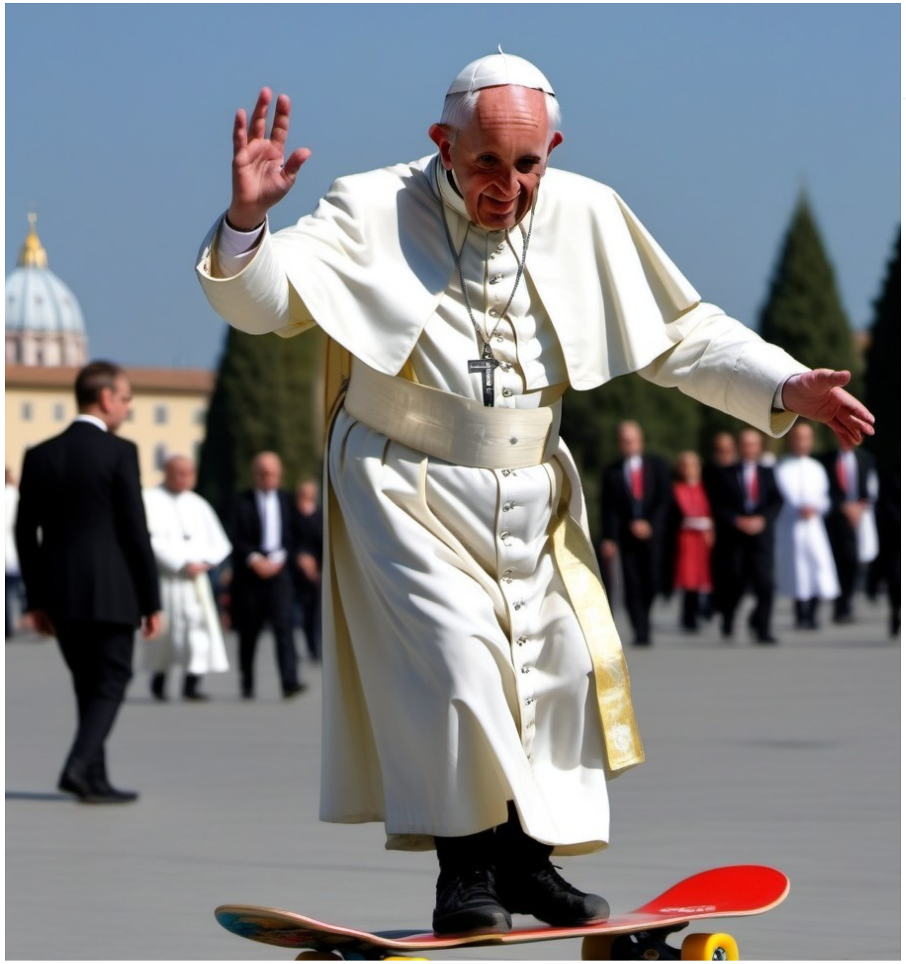

Training and test data
Randomness: the higher, the more „creative“
Input data: “Prompt“
Generative Deep Learning Models
Random Selection
Generated Output

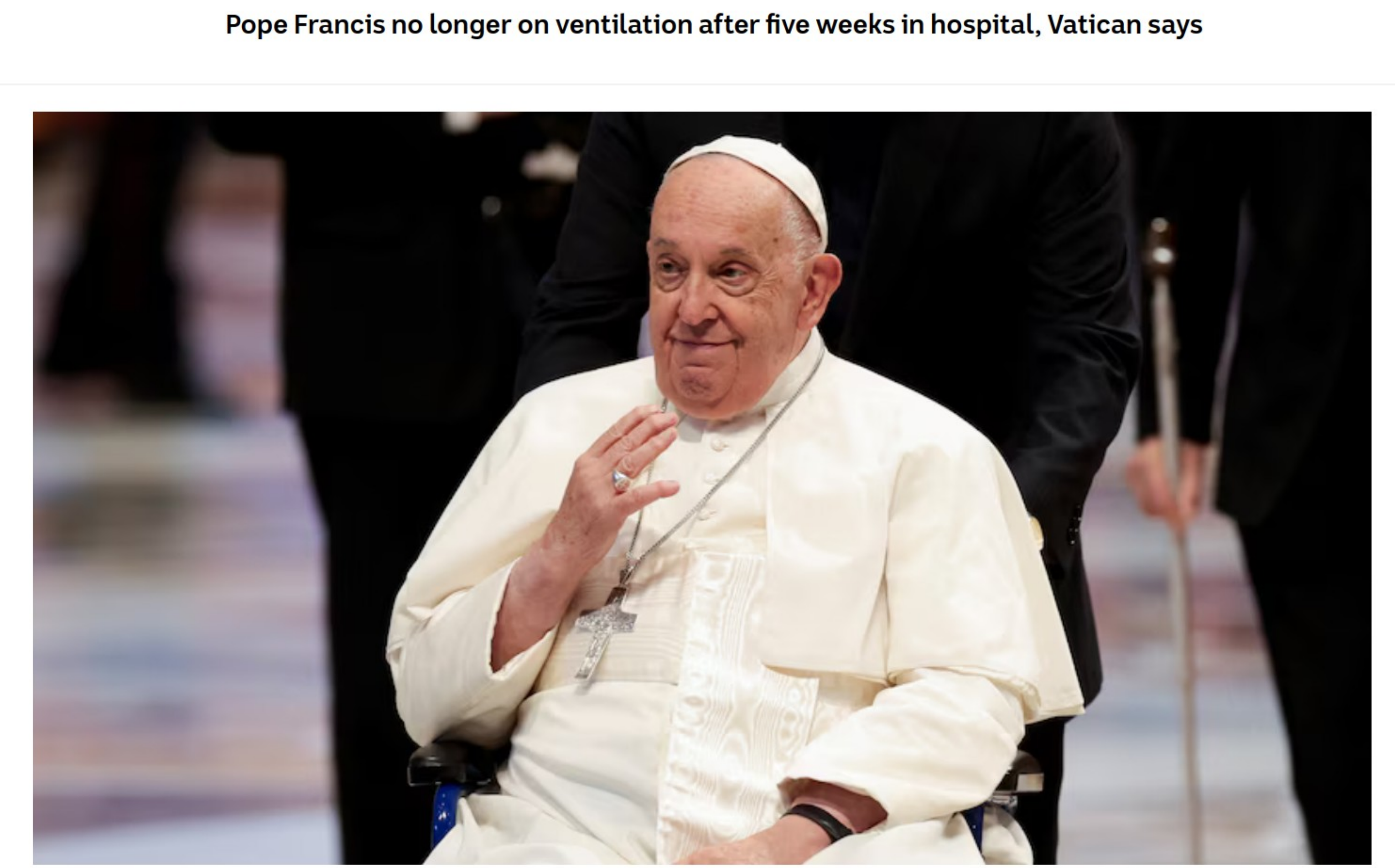
JOHANNES KEPLER UNIVERSITY LINZ

ML at Facebook, Xitter, etc.



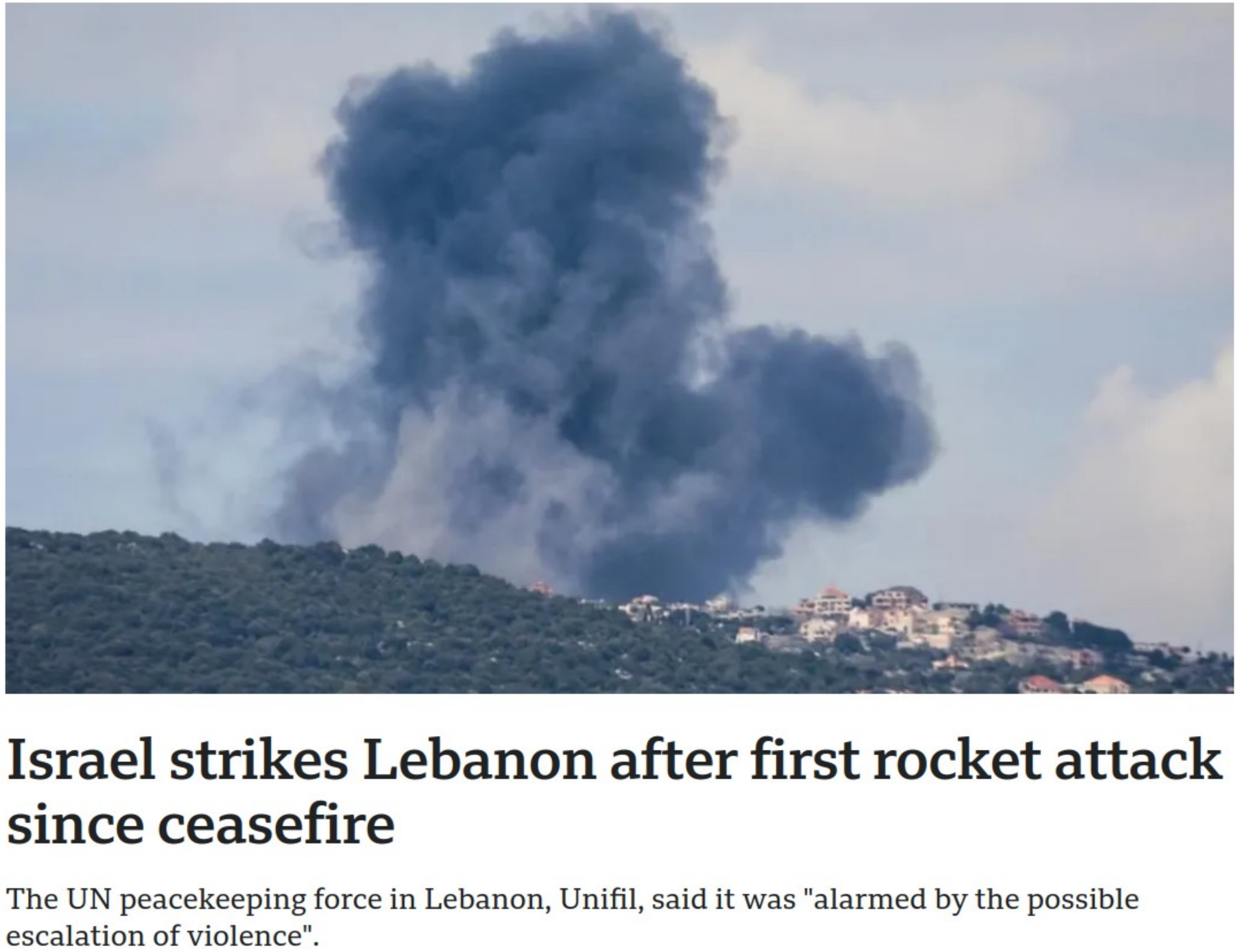
Which post would you rather share/like/boost?





Pope Francis no longer on ventilation after five weeks in hospital, Vatican says

Which post would you rather share/like/boost?



Israel strikes Lebanon after first rocket attack since ceasefire

The UN peacekeeping force in Lebanon, Unifil, said it was "alarmed by the possible escalation of violence".

USA starten Invasion in Istanbul

Mitten in der Nacht begann die USA eine militärische Invasion in Istanbul. Panzer und Truppen besetzten zentrale Stadtteile. Erste Berichte sprechen von heftigen Kämpfen. Die türkische Regierung hat den Notstand ausgerufen. Internationale Reaktionen werden erwartet.

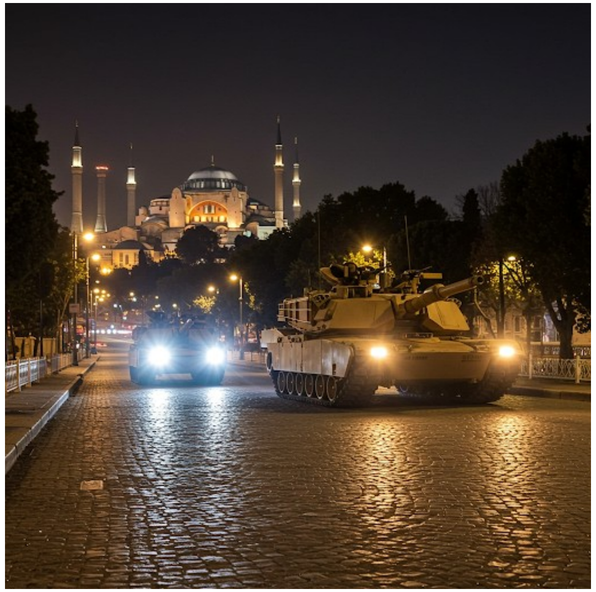


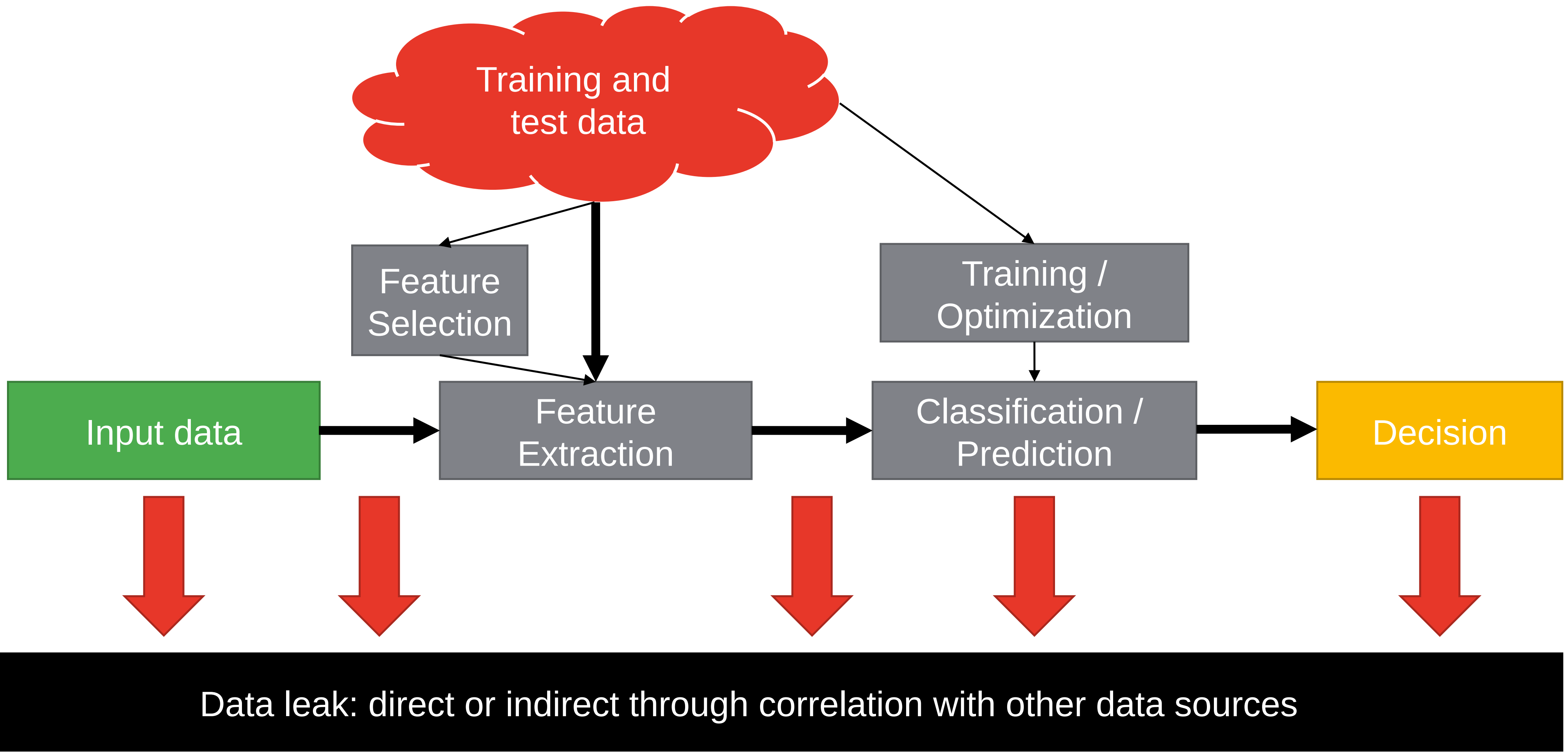
Foto: US-Panzer
rücken in Istanbul vor.

The battle for our attention

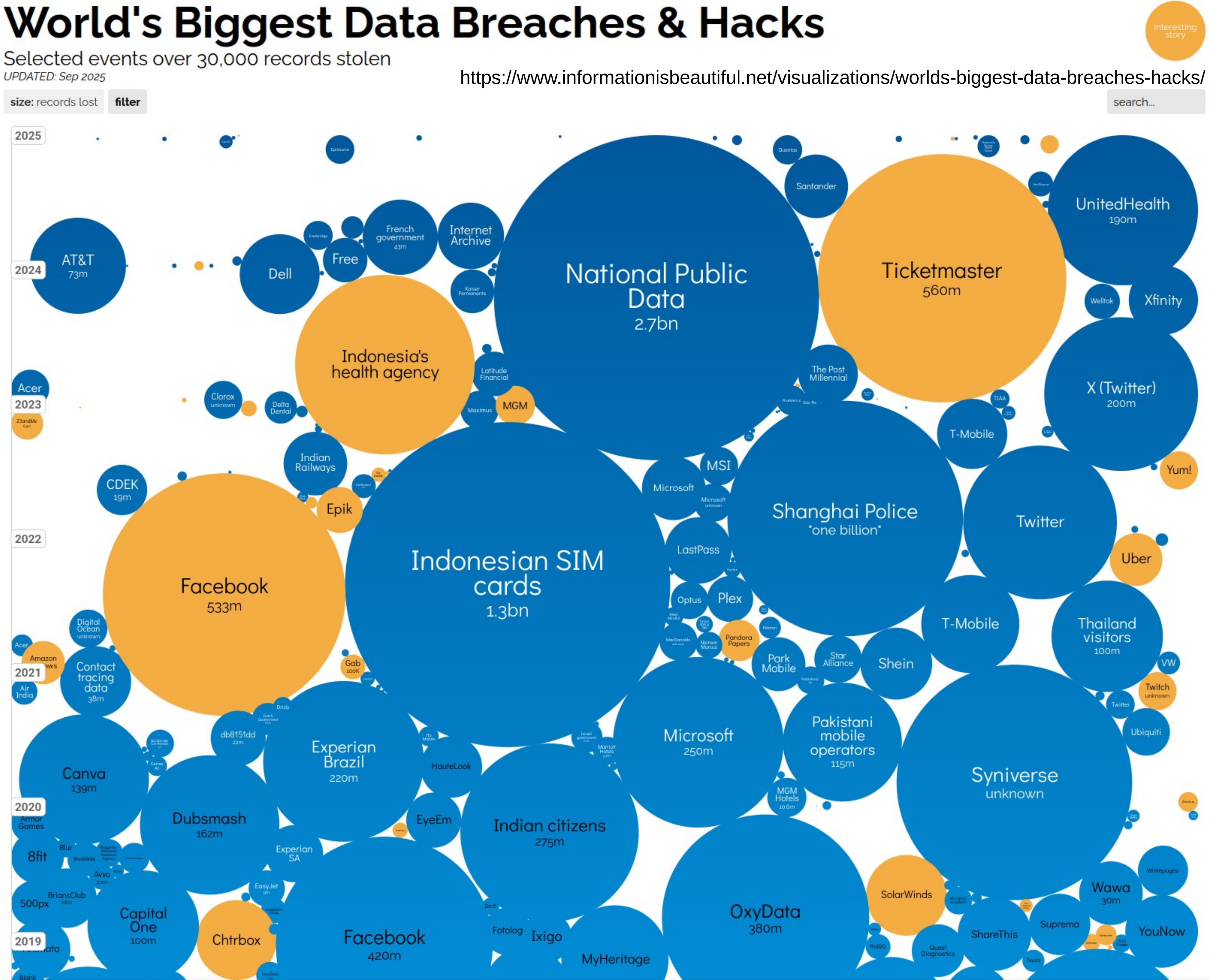
- Recommendation algorithms are optimized for
  - Drawing and keeping **attention** on that particular platform …
  - … because then people see more advertisements …
  - … which makes the platform operator more **money**.
- How do you get the most attention?
  - **Emotion**!
  - Bad emotion works better than good emotion!
  - The more **emotion**, the more **money**!

Created with StableDiffusion prompt “A photo of a really angry cat i

Possible data leaks when using ML



Data leak: direct or indirect through correlation with other data sources



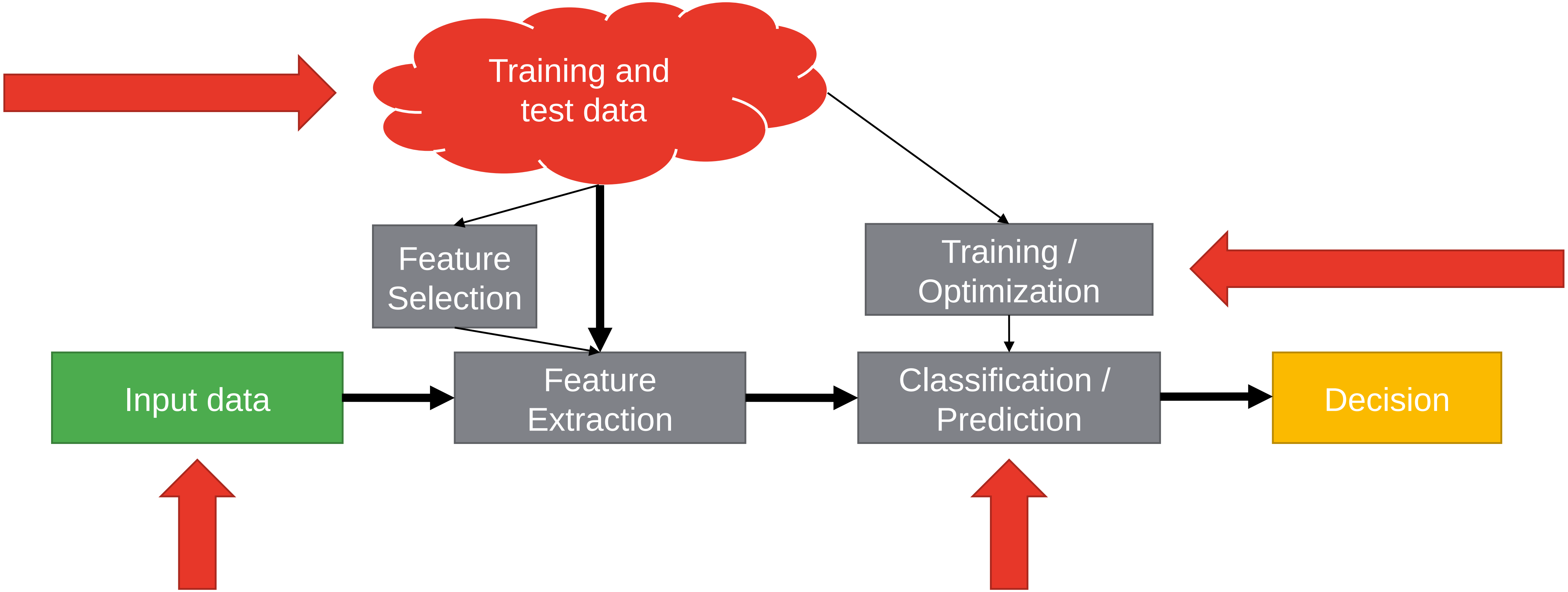
World's Biggest Data Breaches & Hacks

Selected events over 30,000 records stolen

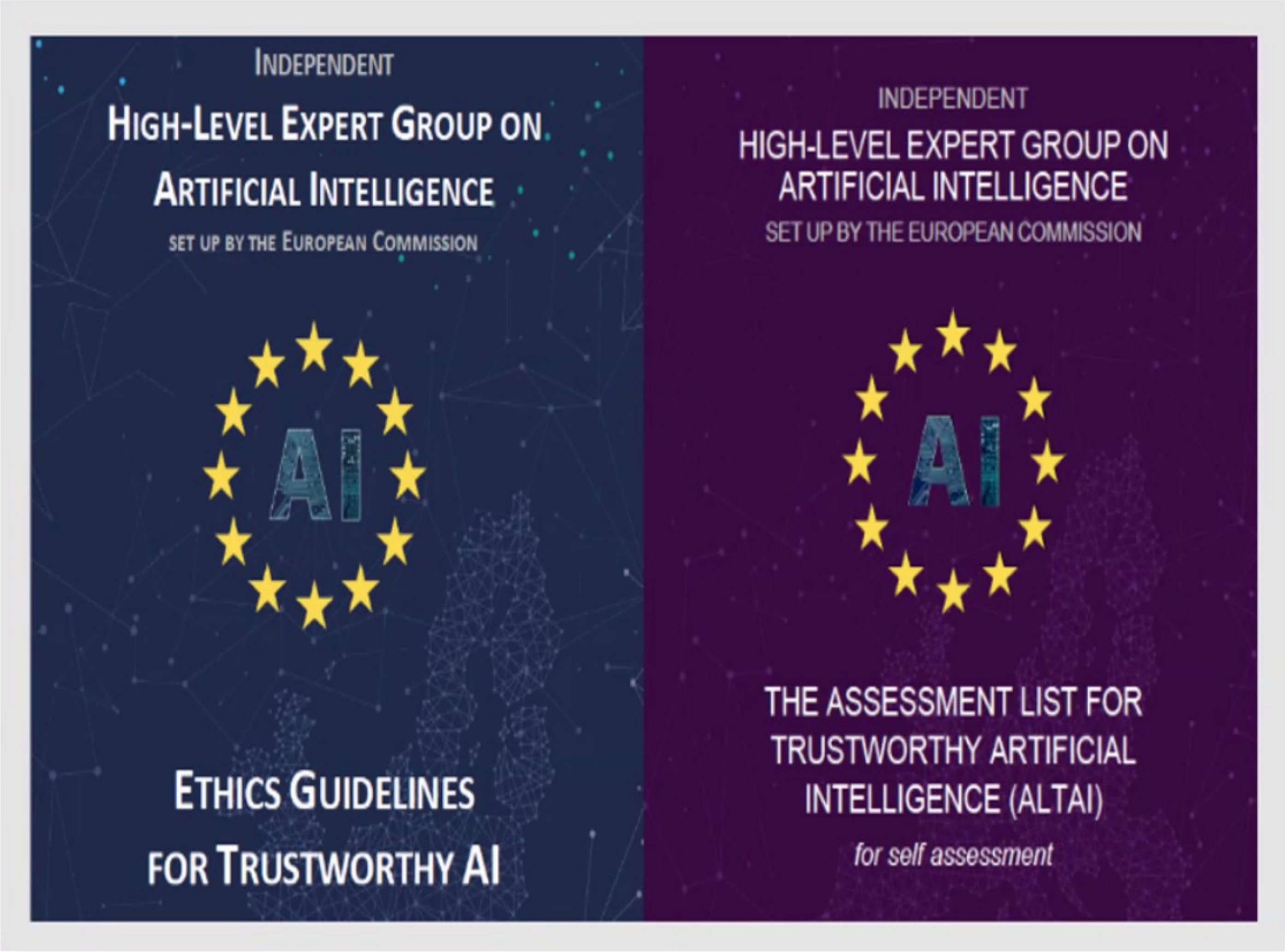
UPDATED: Sep 2025

https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

Possible attacks when using ML



EU AI Act: Current developments



INDEPENDENT

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES FOR TRUSTWORTHY AI

INDEPENDENT

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

SET UP BY THE EUROPEAN COMMISSION

THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI)

for self assessment

7 core requirements:

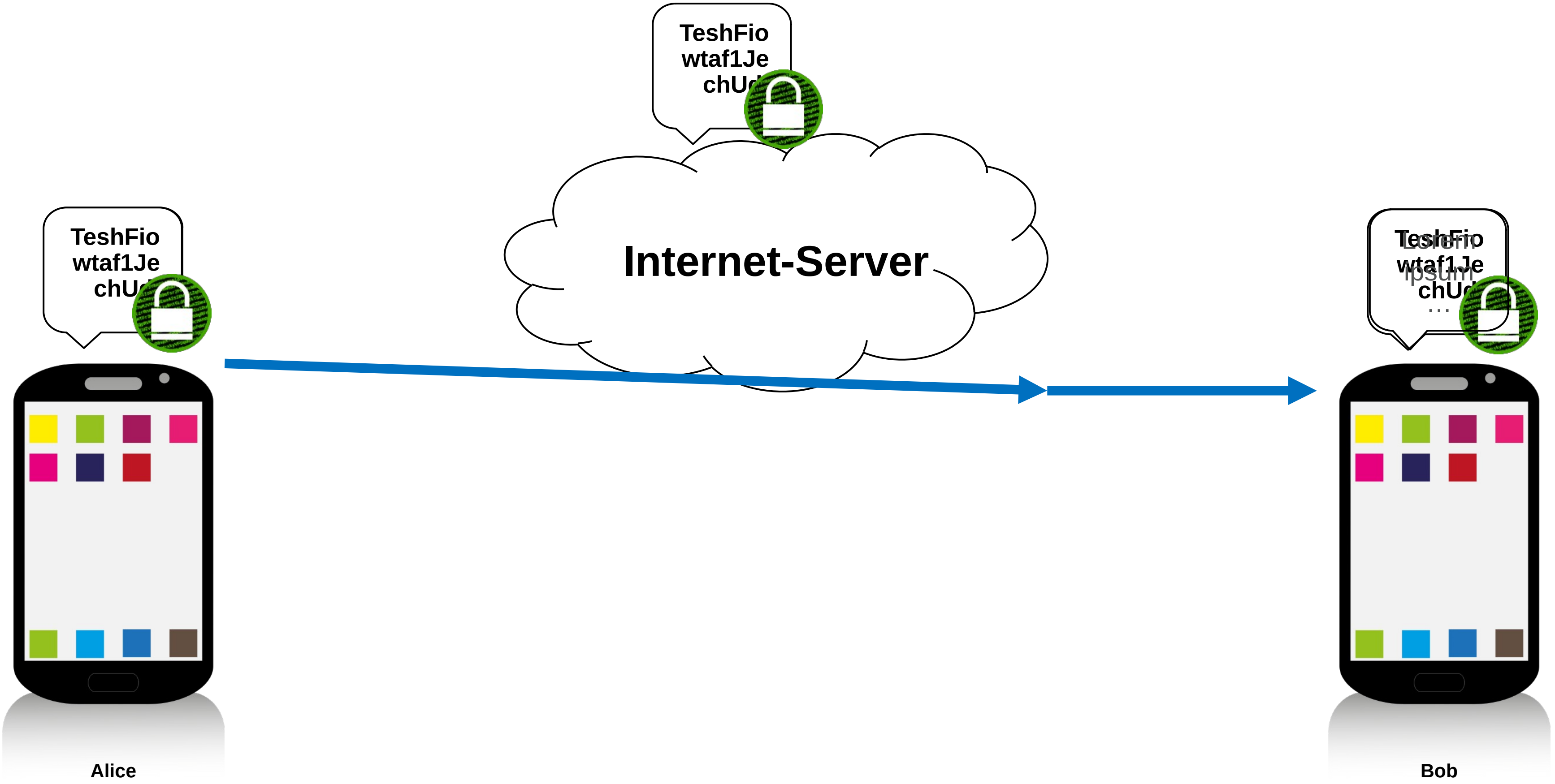
1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability

So you are concerned about
bad content in
social media / messenger apps?

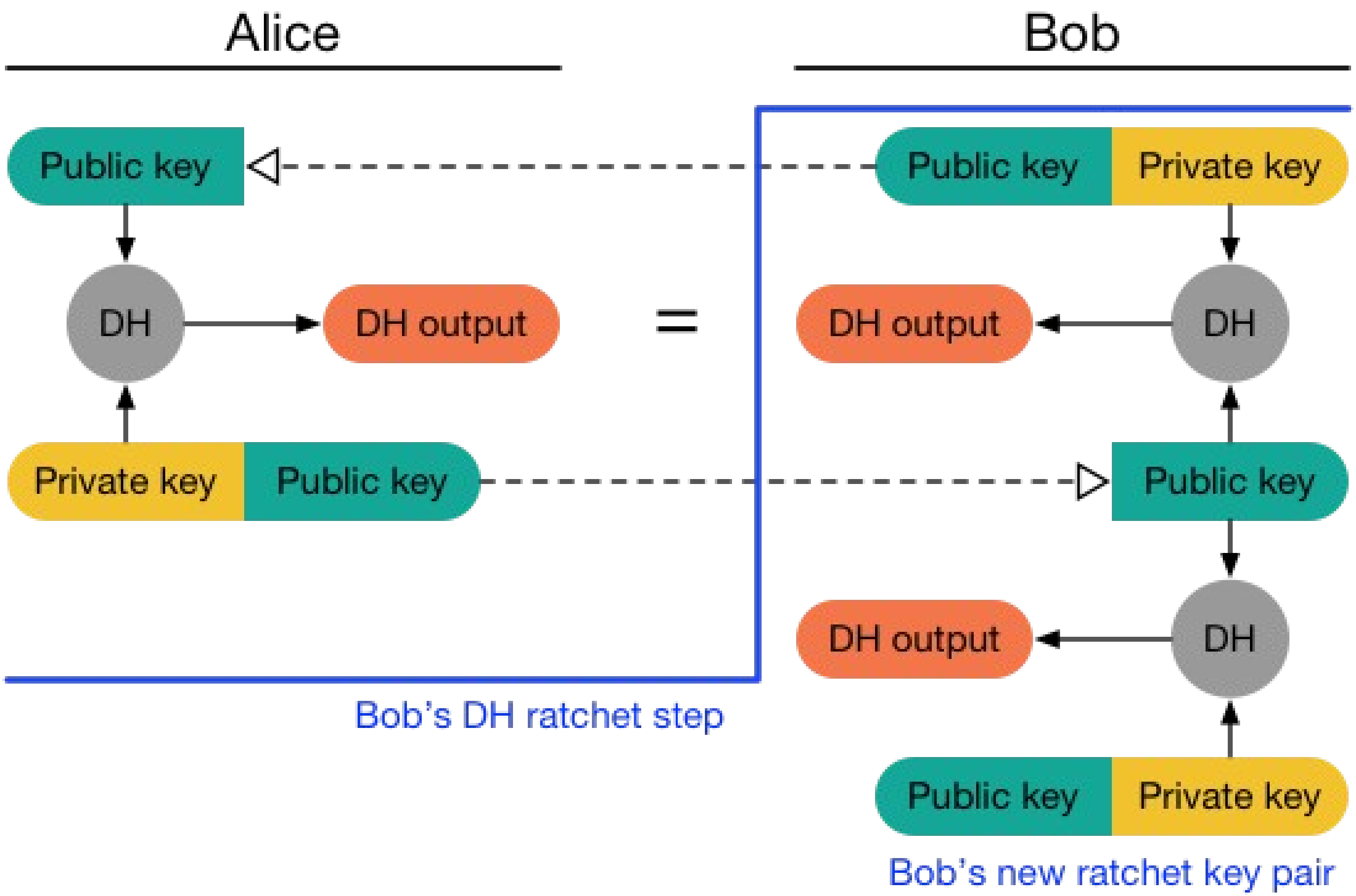
Are you also concerned about “bad” content communicated in private, face-to-face?

The definition of “bad” depends on the policy of the day, and can change quickly with (or without) a single election...

Traces through Signal, Wire, Threema, etc.: End-to-End Encryption (E2EE)



Encryption:
Signal Protocol Double Ratchet



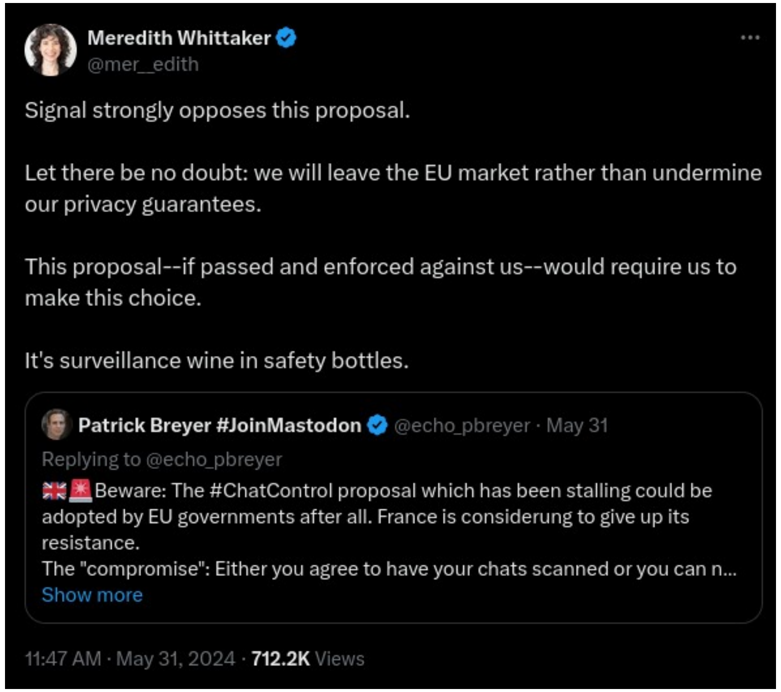
JOHANNES KEPLER
UNIVERSITY LINZ

https://signal.org/docs/specifications/doubleratchet/

Ok, network based content extraction is hard…

Can we just scan the endpoints (=apps) for plaintext messages?

Letting apps do their own scanning: Client-Side Scanning (CSS)



Meredith Whittaker @mer_edith

Signal strongly opposes this proposal.

Let there be no doubt: we will leave the EU market rather than undermine our privacy guarantees.

This proposal--if passed and enforced against us--would require us to make this choice.

It's surveillance wine in safety bottles.

Patrick Breyer #JoinMastodon @echo_pbreyer · May 31

Replying to @echo_pbreyer

🇬🇧🚨Beware: The #ChatControl proposal which has been stalling could be adopted by EU governments after all. France is considerung to give up its resistance.
The "compromise": Either you agree to have your chats scanned or you can n...
Show more

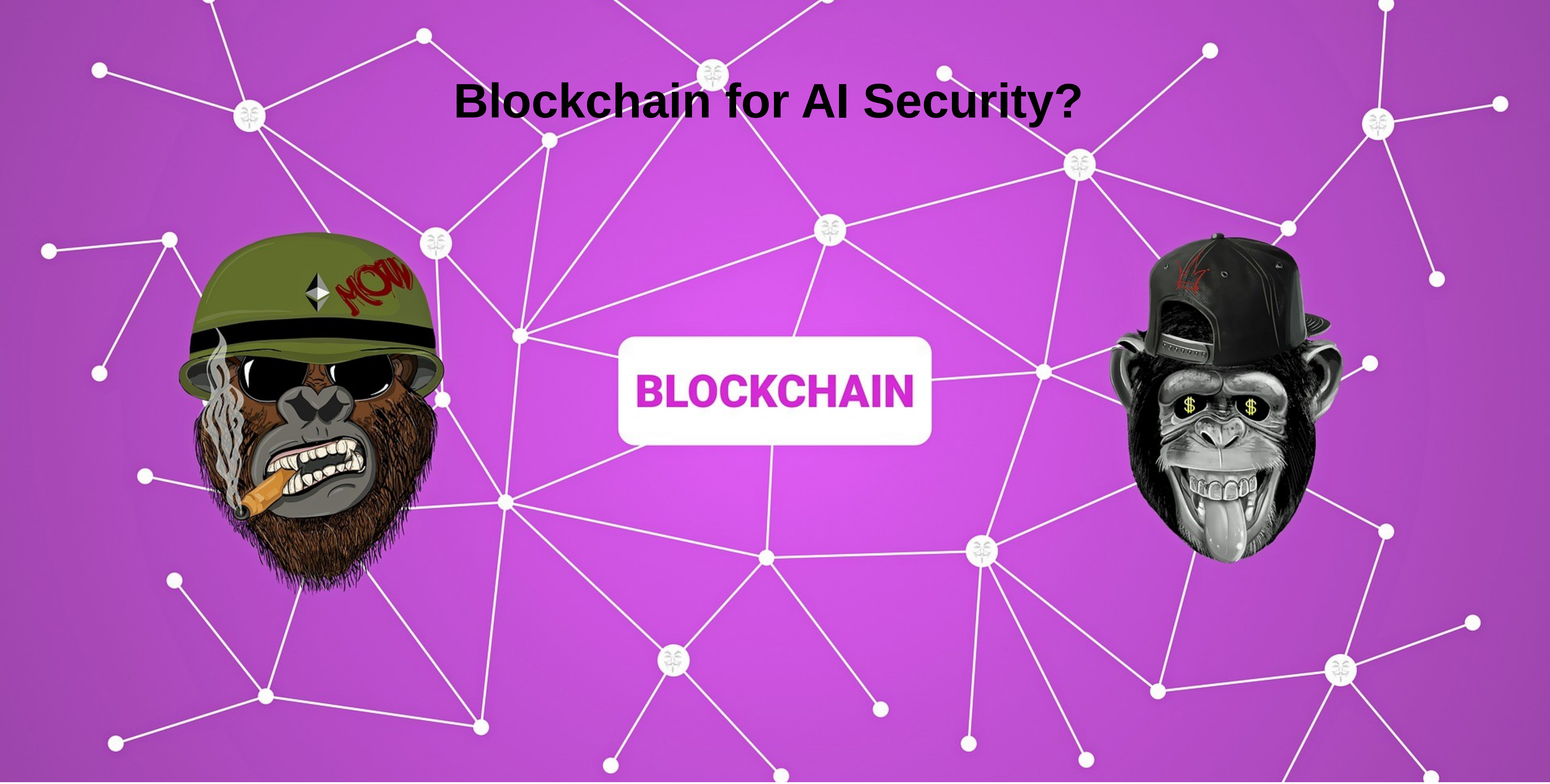
11:47 AM · May 31, 2024 · **712.2K** Views

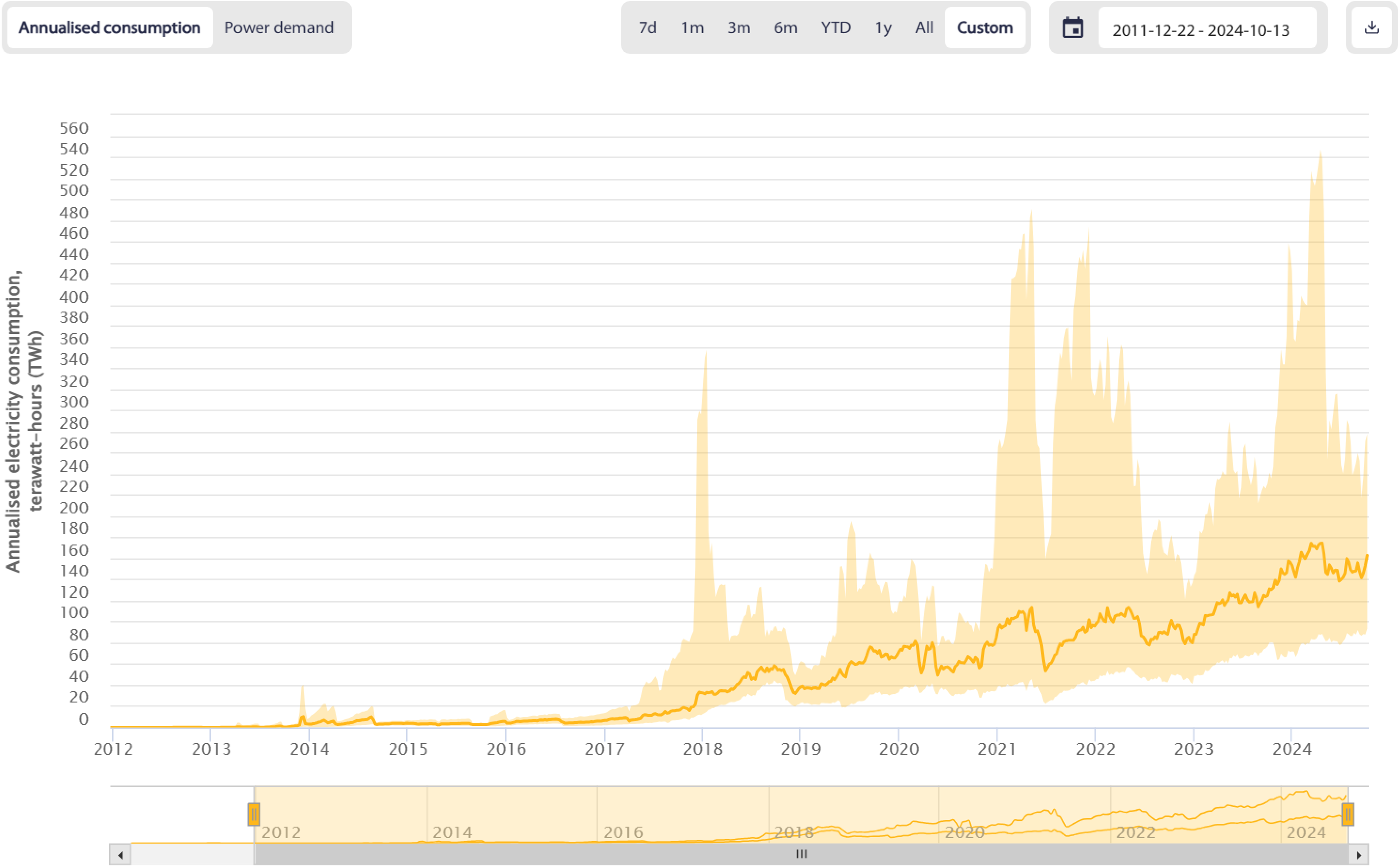
- Can legally compel apps to implement scanning inside the app
  - Has access to plaintext messages and all media
  - Proprietary apps can implement a mandated **secret filter**
    - with or without enforced automatic reporting
- Technical challenges
  - **Filter has non-negligible error rate**
    - Many, many, many false positives to be expected
  - **Keeping filter secret** → even if non-extractable (which is hard), can use as oracle
    - Training input recovery is a thing with more complex filter models → CSAM material???
  - **No way to technically enforce on all apps** → take e.g. Signal source code, compile without filter, use within organized crime group
  - **Added complexity** → added attack surface for app
- Legal challenges
  - Mass surveillance “pre-crime” scanning
  - Self censorship based on existence of filter

See also https://www.ins.jku.at/chatcontrol/

Ok, client side scanning (with or without ML) is tricky, how about some other security whatever thingy?

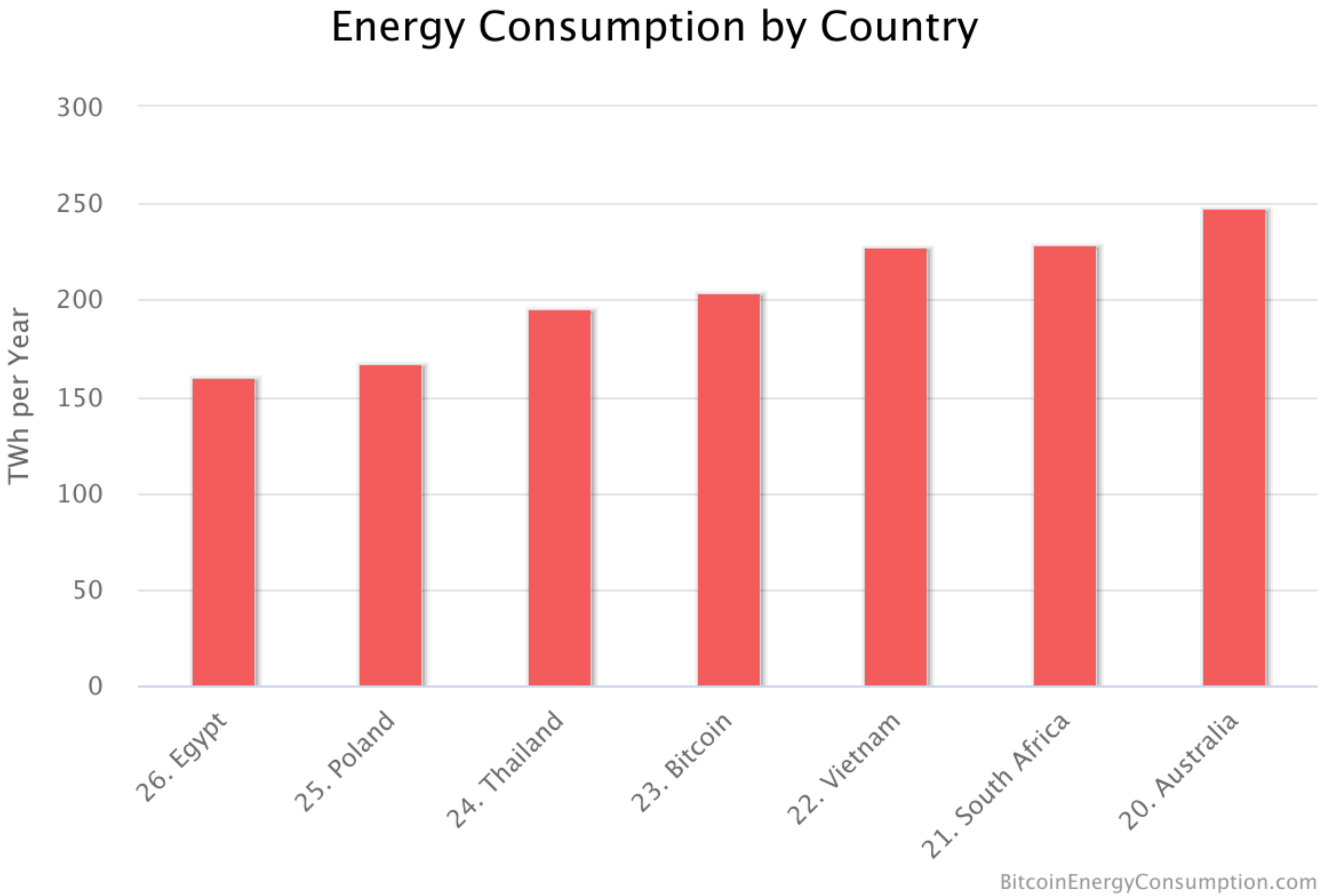
Blockchain for AI Security?





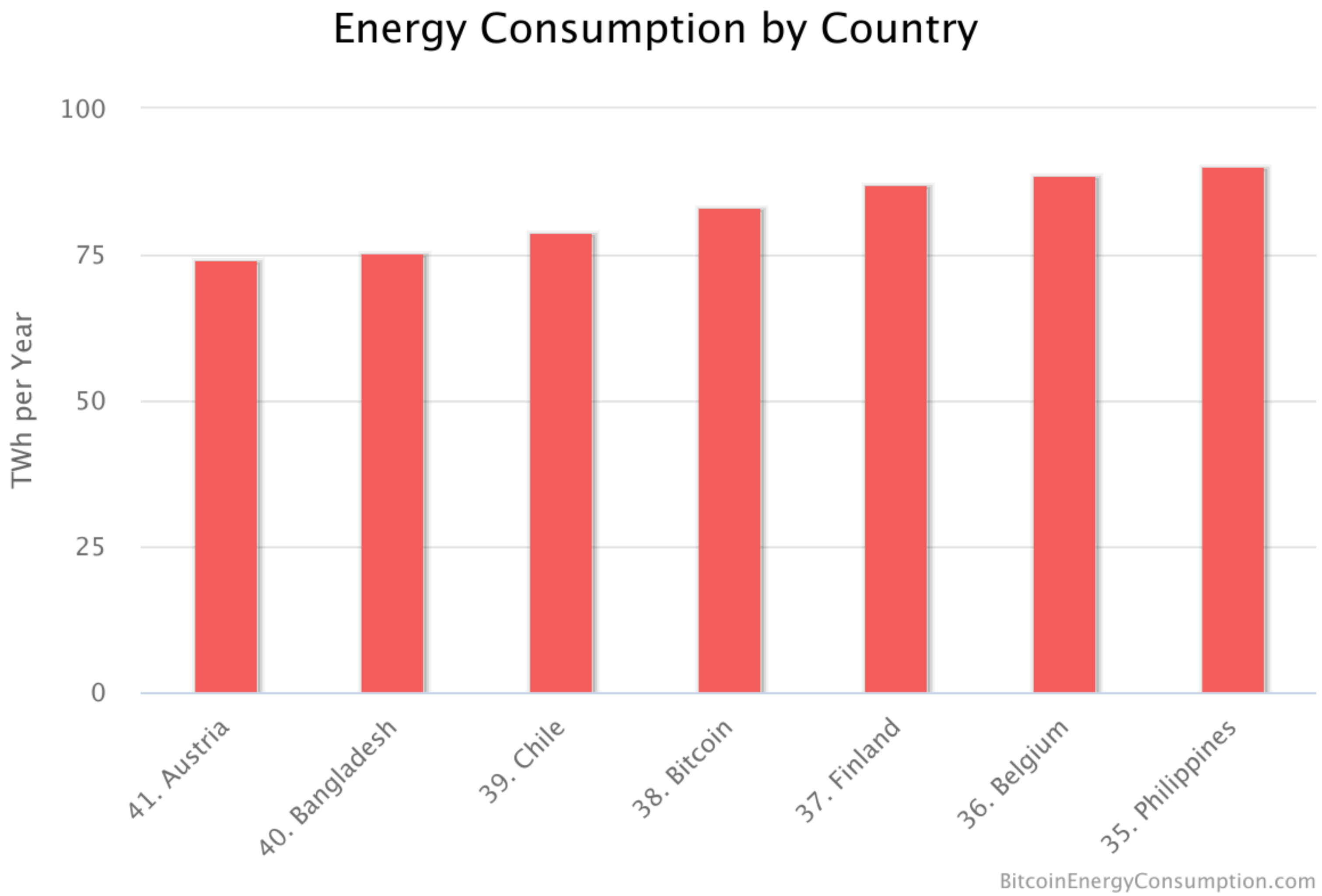
https://ccaf.io/cbnsi/cbeci

Blockchain – Proof of Work Energy Impact



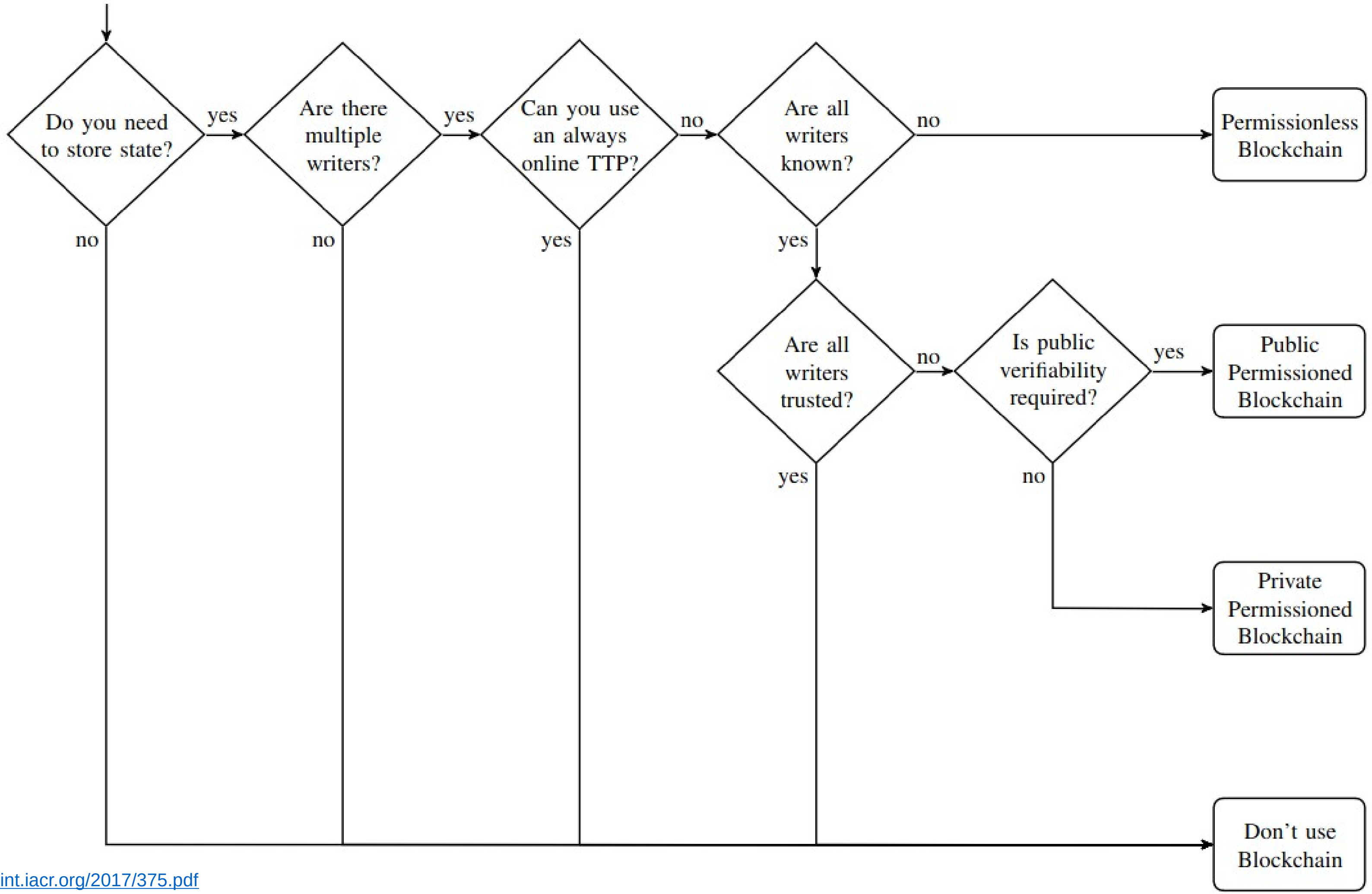
Source: https://digiconomist.net/bitcoin-energy-consumption/, 2021-01-28

Blockchain – Proof of Work Energy Impact



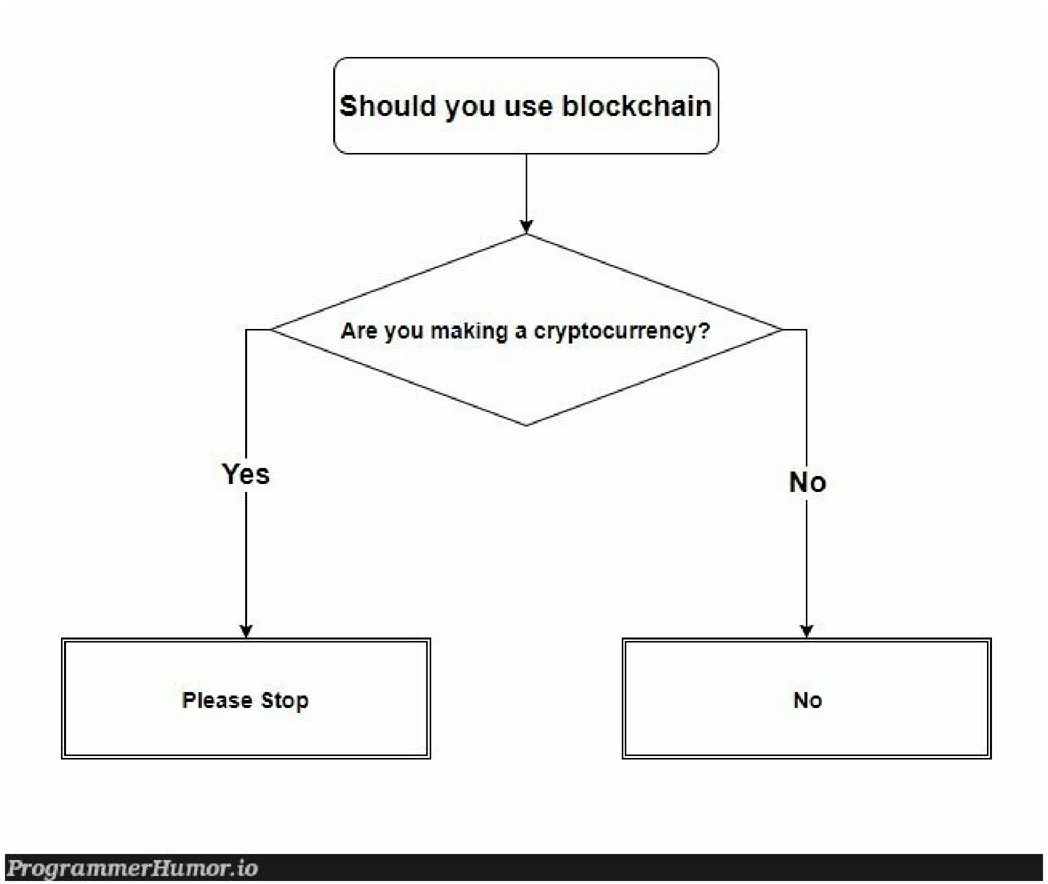
Source: https://digiconomist.net/bitcoin-energy-consumption/, updated 2023

Do you need a Blockchain?


Do you need to store state?
yes
no
Are there multiple writers?
yes
no
Can you use an always online TTP?
no
yes
Are all writers known?
no
yes
Permissionless Blockchain
Are all writers trusted?
no
yes
Is public verifiability required?
yes
no
Public Permissioned Blockchain
Private Permissioned Blockchain
Don't use Blockchain

Source: https://eprint.iacr.org/2017/375.pdf

Do you need a Blockchain?



Energy consumption of current ML models

- **Extreme during training**
  - Estimate: GPT-3 used 1.300 MWh during training (for 175 Billion parameters)
    → ca. 1.625.000h Netflix streaming
  - GPT-4 scales parameter size 10x… (energy/water/etc. consumption no longer published)
- **Slightly less during inference** (evaluation/use)
  - Estimate: generating a single image is comparable to charging a smartphone
  - Estimate: adding the “AI answer” block at the top of each search query 100-1000x of the consumption without that block

Source: https://www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption

Energy consumption of digital services?

- **Video conferencing**
  - Group video call with 5 participants for **1 hour** in HD quality: ca. **0,10kWh** (comparable to ca. 0,2km with combustion engine car or **1km with battery electric car**)
  - Can save transfer efforts significantly (estimates around 90%) with audio-only
  - Textual communication (the old email…) much more efficient
- **Video streaming**
  - **1 hour** Netflix network streaming: ca. **0,077kWh – 0,8kWh**
  - Depends mostly on device: 50” TV screen ca. 100x, laptop ca. 5x compared to smartphone
  - For smartphone viewing (**<0,05kWh**), ca. 80% of energy used for data transmission (networks)

1 Bitcoin transaction: >2000kWh

- Energy consumption of **devices**: **30%** for TVs, **80%** for smartphones **during production**
- 2021/2022: All **data centers**: annually ca. **200 TWh + 250 TWh** network → ~2% of global electricity
- Estimate 2026: **data centers 620–1050 TWh** because of GenAI [IEA 2024, p .31]

Bitcoin: 100-200TWh

GenAI: 100-x00TWh

Note: Data collected from different sources in summer 2021 and fall 2024; https://www.utilitybidder.co.uk/business-electricity/zoom-emissions/, https://www.carbonbrief.org/factcheck-what-is-the-carbon-footprint-of-streaming-video-on-netflix, https://www.sciencedaily.com/releases/2021/01/210114134033.htm, https://www.iea.org/reports/data-centres-and-data-transmission-networks, https://cbeci.org/cbeci/comparisons, https://www.iea.org/reports/electricity-2024

Questions?

Take-away summary:
Recommendations for balancing risks and benefits of AI

- **If possible, run inference on local models**
  - Tooling is getting much better, e.g. `ollama` in Docker
  - Local hardware is getting faster for ML acceleration, e.g. AMD Ryzen AI CPUs using all the local system RAM instead of relying on expensive GPU-specific memory
  - Data stays in-house
- **If required, pay for hosted services**
  - If it's free, your input data is no longer yours
  - (If you pay for it, your data might still be taken, but you can send lawyers after them…)
  - But evaluate carefully if the perceived benefits really outweigh the cost
- Current ML is good for **generating potential answers**, but **not for verifying accuracy**

  → Use only for **low-risk applications** that **can tolerate or painlessly undo errors**

  → Massive Amounts of Misinformation (AI-MAM, a.k.a. lies) are currently the biggest danger
- **Agentic AI is …** pretty good job security for the whole security team

JOHANNES KEPLER
UNIVERSITY LINZ

Web: https://ins.jku.at
Email: rm@ins.jku.at

Mastodon: @rene_mobile@infosec.exchange
Signal: Rene.02