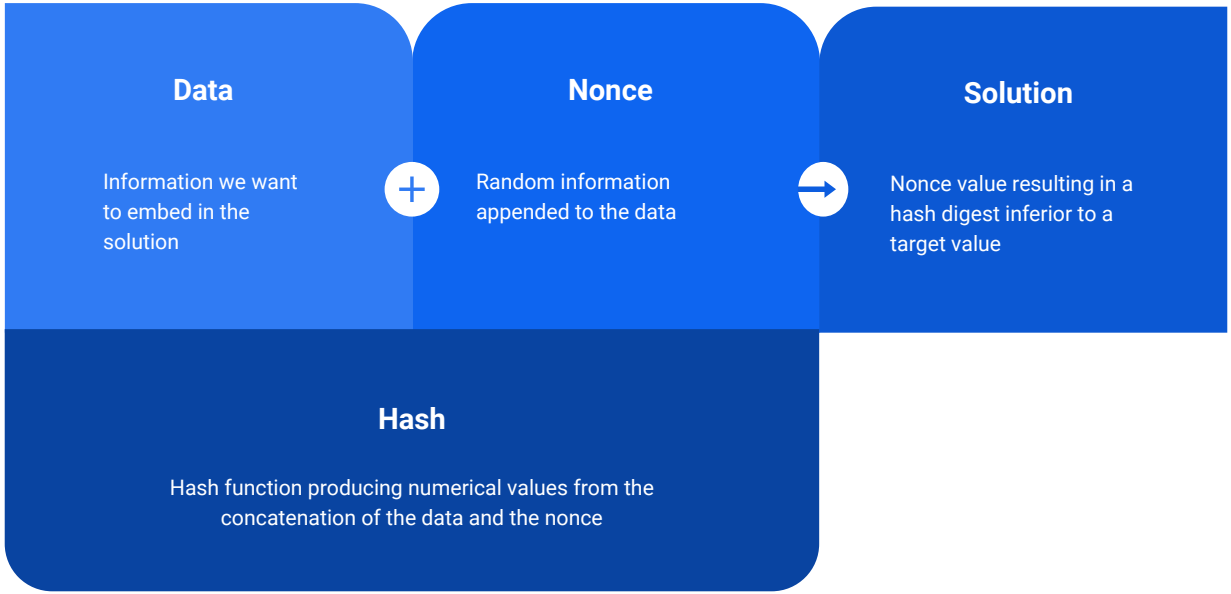
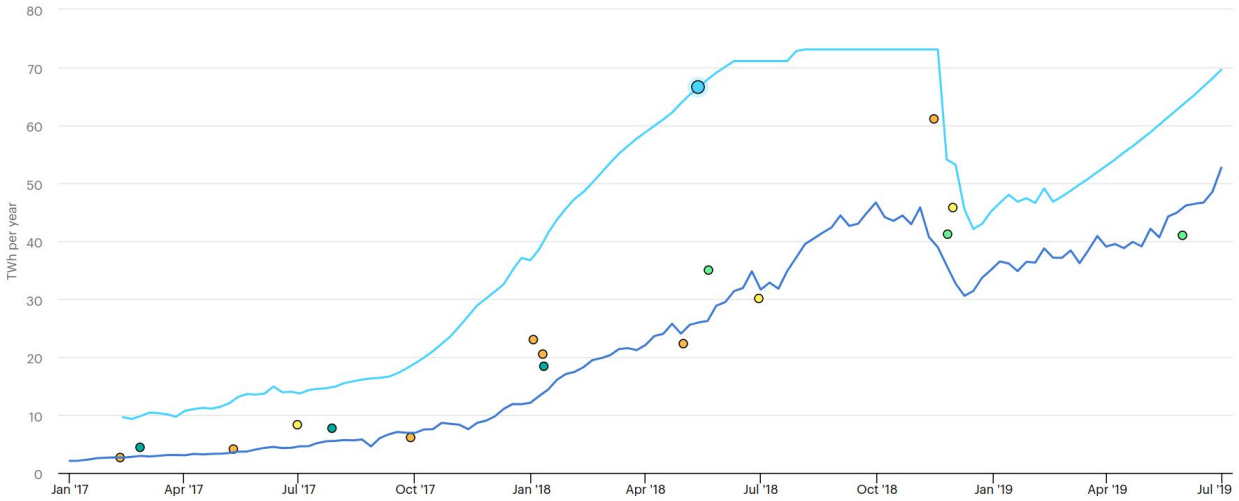
FreezeNet

Making Proof of Work Useful

Proof of Work - with Hashcash

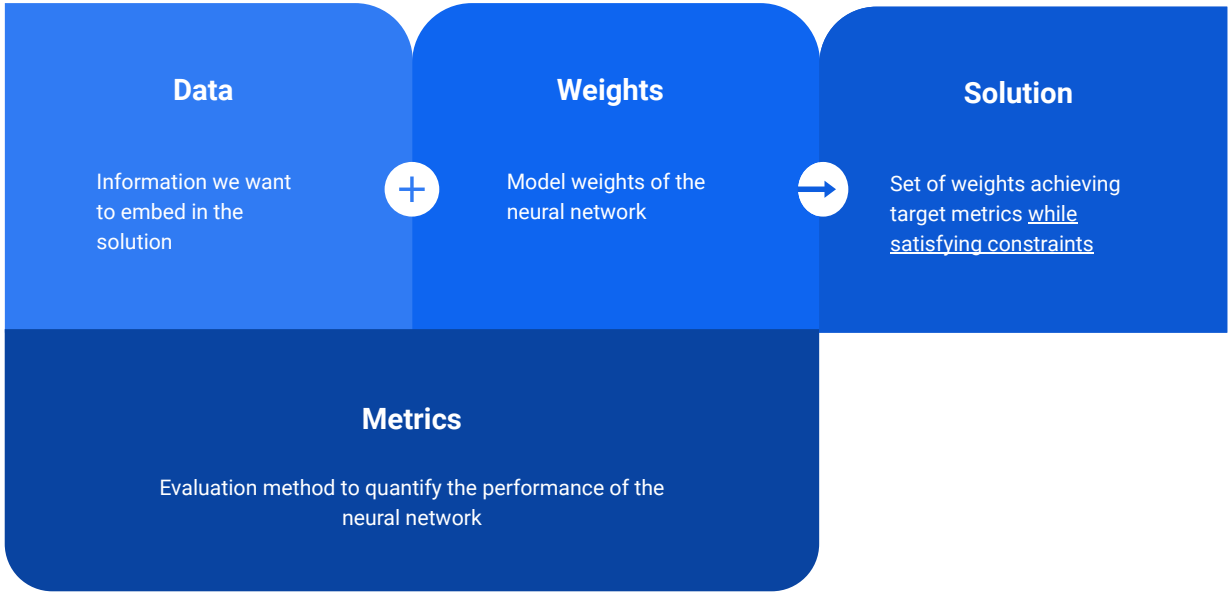


Bitcoin energy usage estimates

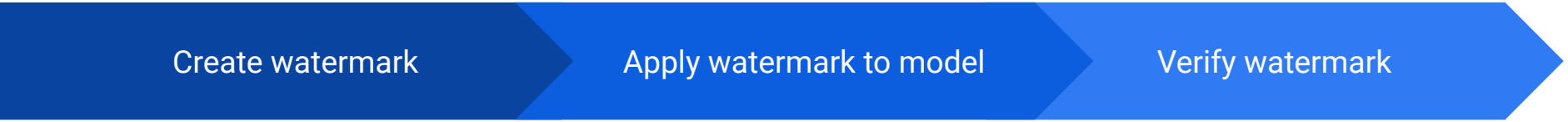


Digiconomist (2019) Lower bound (Antminer S9) Bendiksen, Gibbons, Lim (2018-19) Bevand (2018) Other peer-reviewed studies Other studies

Proof of Work - with FreezeNet



Watermarking : from a simple idea …



- Hash data to generate PRNG seed
- Use PRNG initialized with seed to generate weights and indices
- Number of weights generated is the watermark size

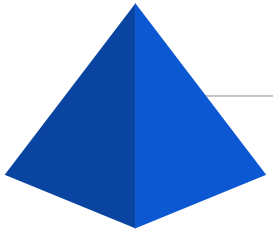
- Replace model weights at watermark indices by watermark weights
- Weights can be replaced during training after backpropagation phase

- Check if model weights at watermark indices are equal to watermark weights
- Account for floating point imprecision by setting upper bound on difference

… to a simple API

```
 1 model = FreezeNet()
 2
 3 some_watermark    = Watermark(b'Some block information', 4096)
 4 another_watermark = Watermark(b'Some other block information', 4096)
 5
 6 some_watermark.apply(model)
 7
 8 some_watermark.verify(model)    # True
 9 another_watermark.verify(model) # False
10
11 another_watermark.apply(model)
12
13 some_watermark.verify(model)    # False
14 another_watermark.verify(model) # True
```

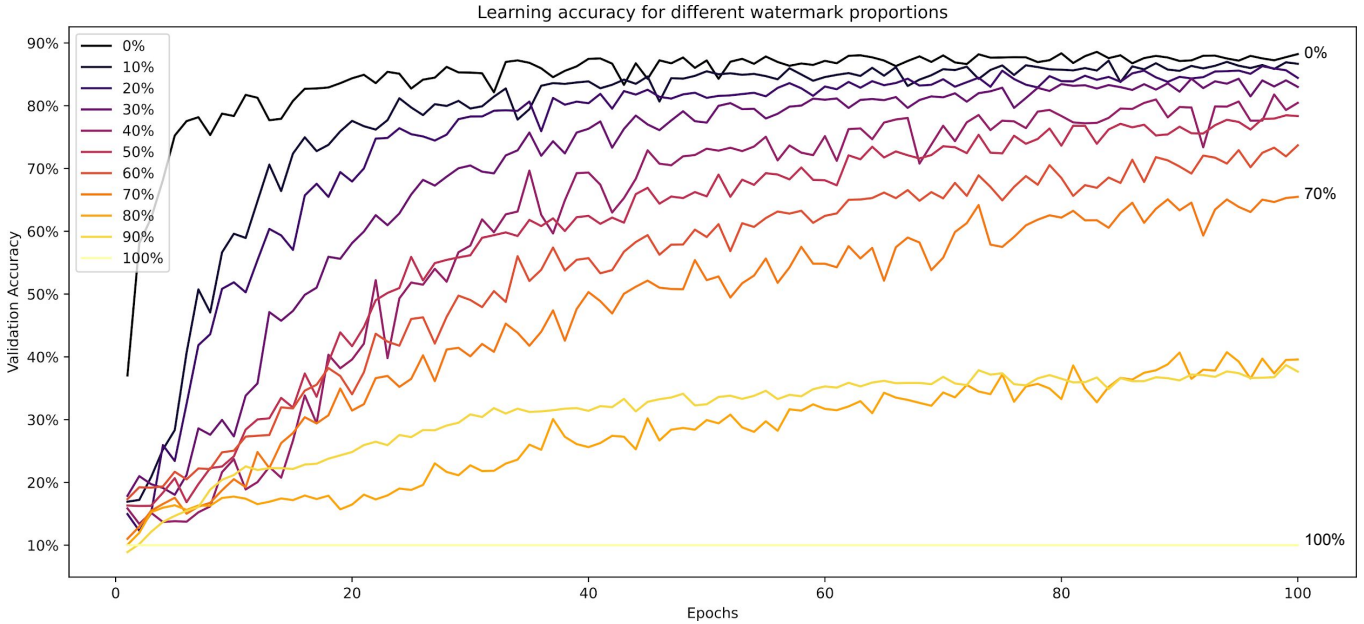
Desired behaviours



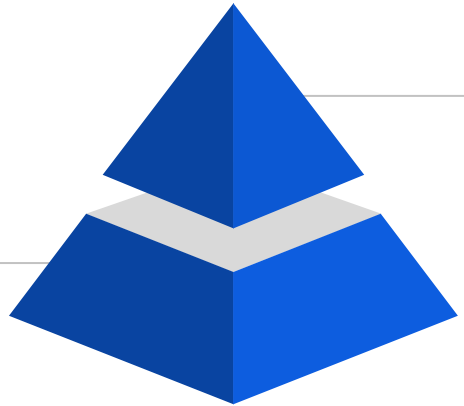
Learning ability

1. Models should still be able to learn while being watermarked

Learning ability



Desired behaviours



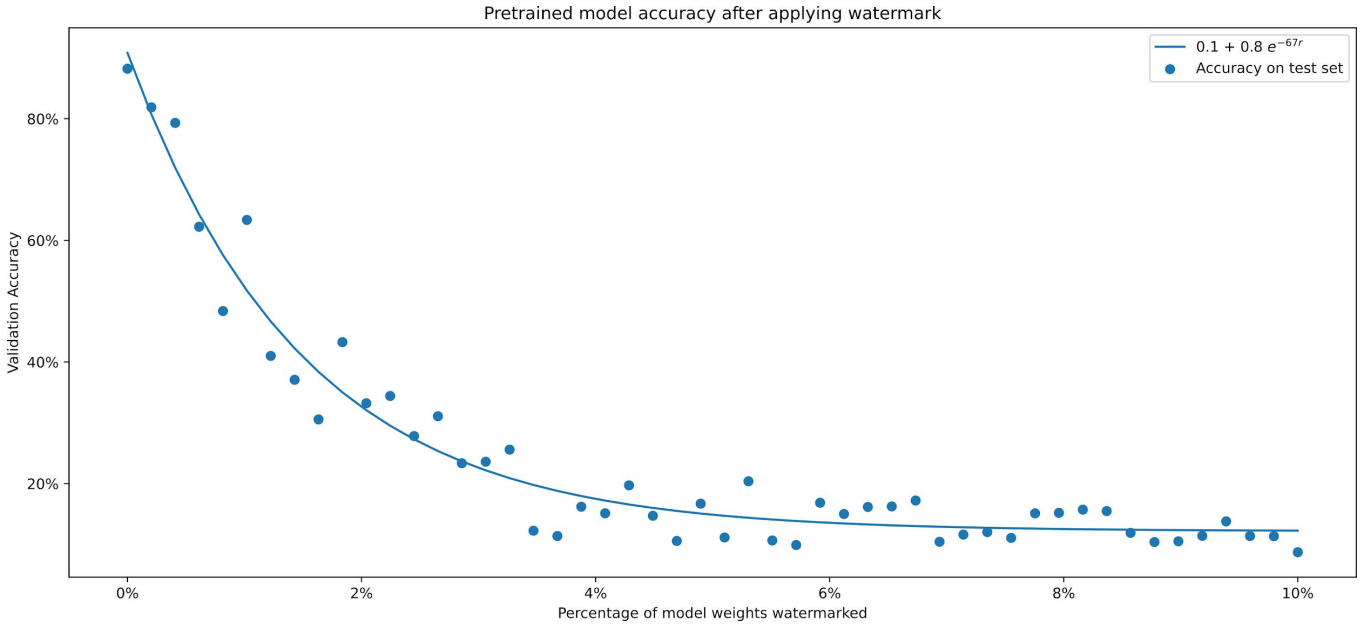
Learning ability

1 Models should still be able to learn while being watermarked

Weak tampering resistance

2 A model should not perform well if it is watermarked after being trained

Weak tampering resistance



Desired behaviours

Learning ability

1 Models should still be able to learn while being watermarked

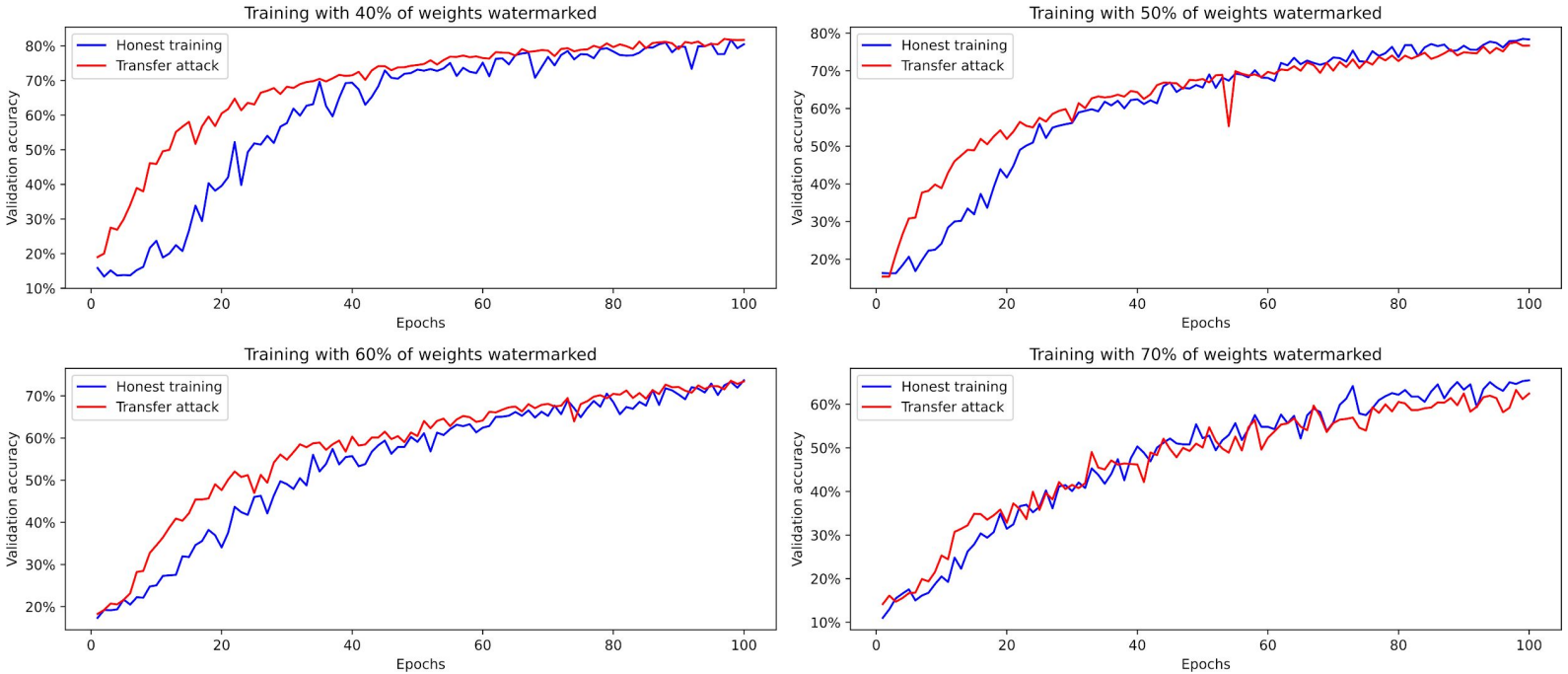
Weak tampering resistance

2 A model should not perform well if it is watermarked after being trained

Strong tampering resistance

3 An attacker should have no incentive to re-train a pre-trained model with a watermark

Strong tampering resistance



Desired behaviours

Learning ability

1 Models should still be able to learn while being watermarked

Weak tampering resistance

2 A model should not perform well if it is watermarked after being trained

Strong tampering resistance

3 An attacker should have no incentive to re-train a pre-trained model with a watermark

Concerns

Overhead

Dataset and model weights transfer adds significant overhead

Centralization

Datasets are not generated in a distributed manner, so workers have to rely on a centralized dataset distribution

Security

The attack surface is very large, transfer attacks only scratch the surface of cheating possibilities