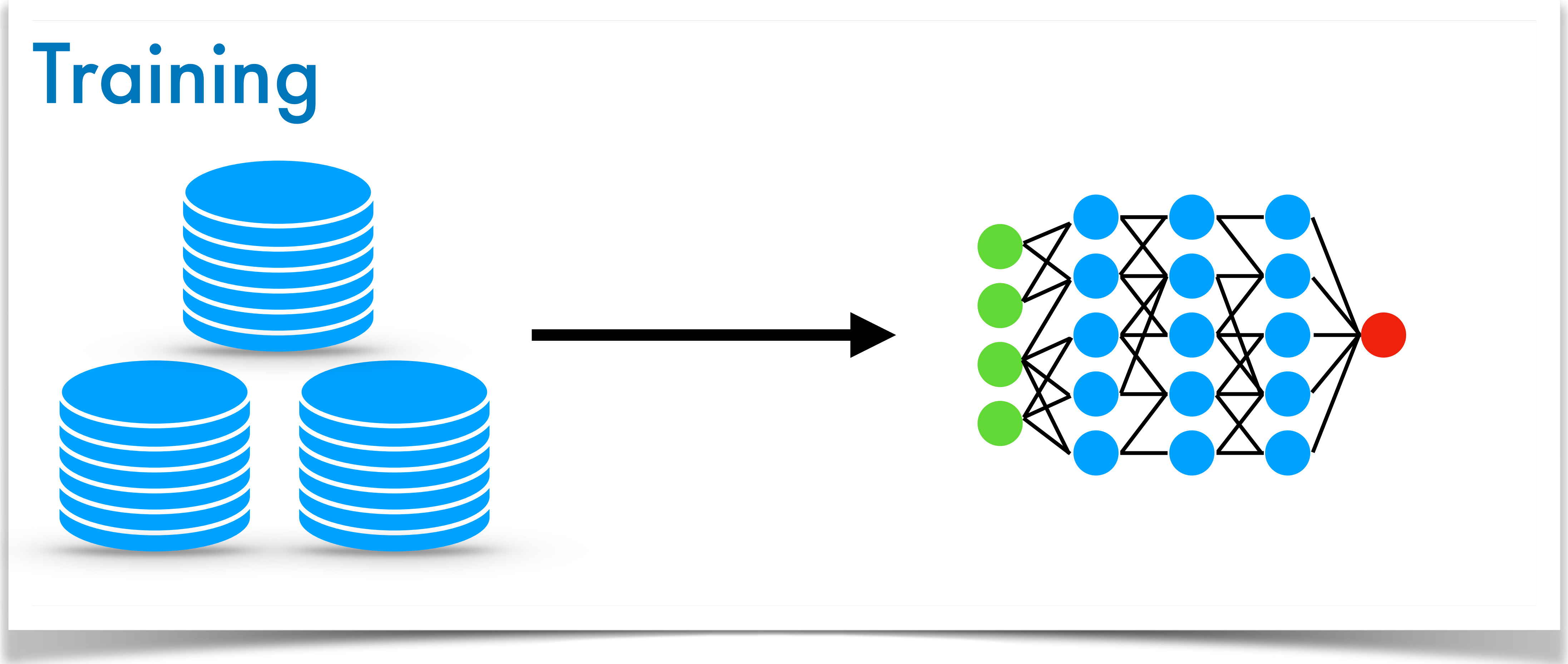
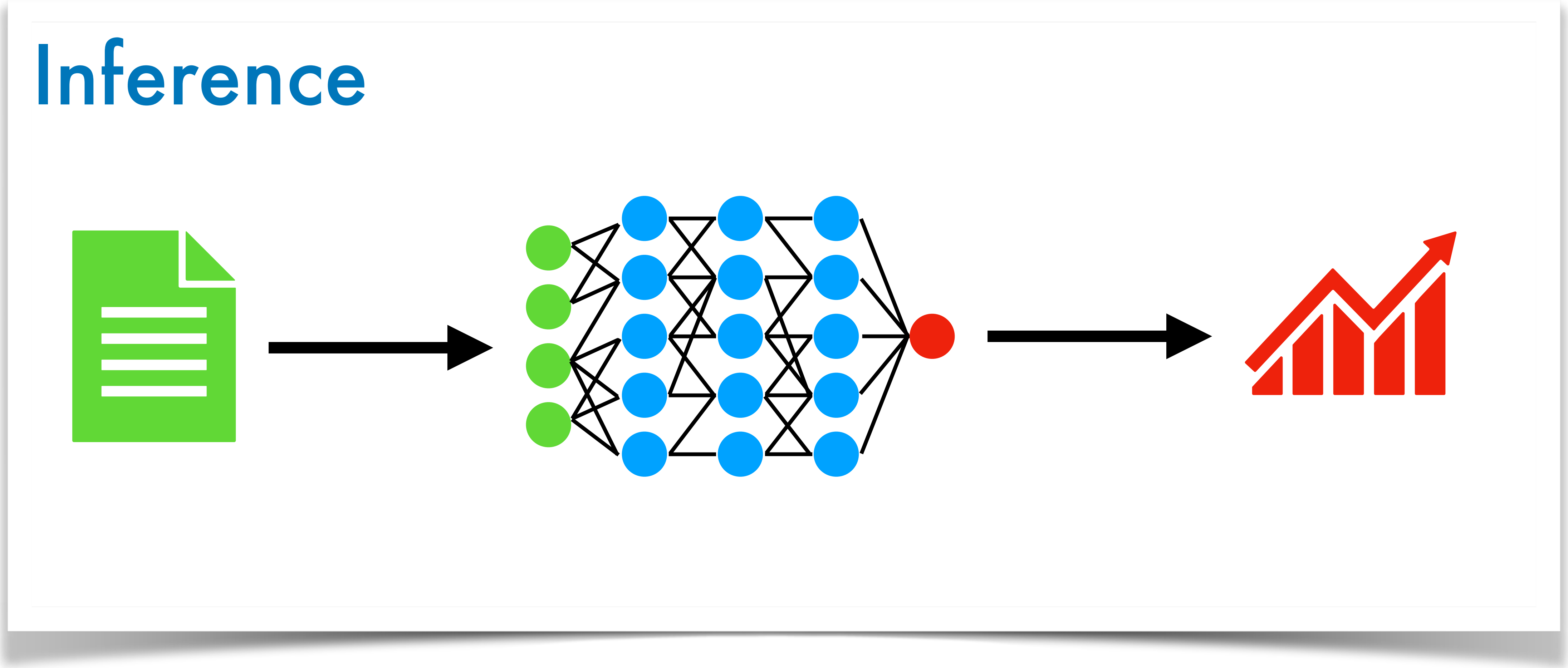
How can Cryptography help with AI regulation compliance?

*Sanjam Garg, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, **Guru-Vamsi Policharla**, and Mingyuan Wang*

THE UNIVERSITY OF CALIFORNIA BERKELEY · NTT Research · THE UNIVERSITY of WISCONSIN MADISON · Meta · UNIVERSITY of VIRGINIA

Machine Learning





Applications

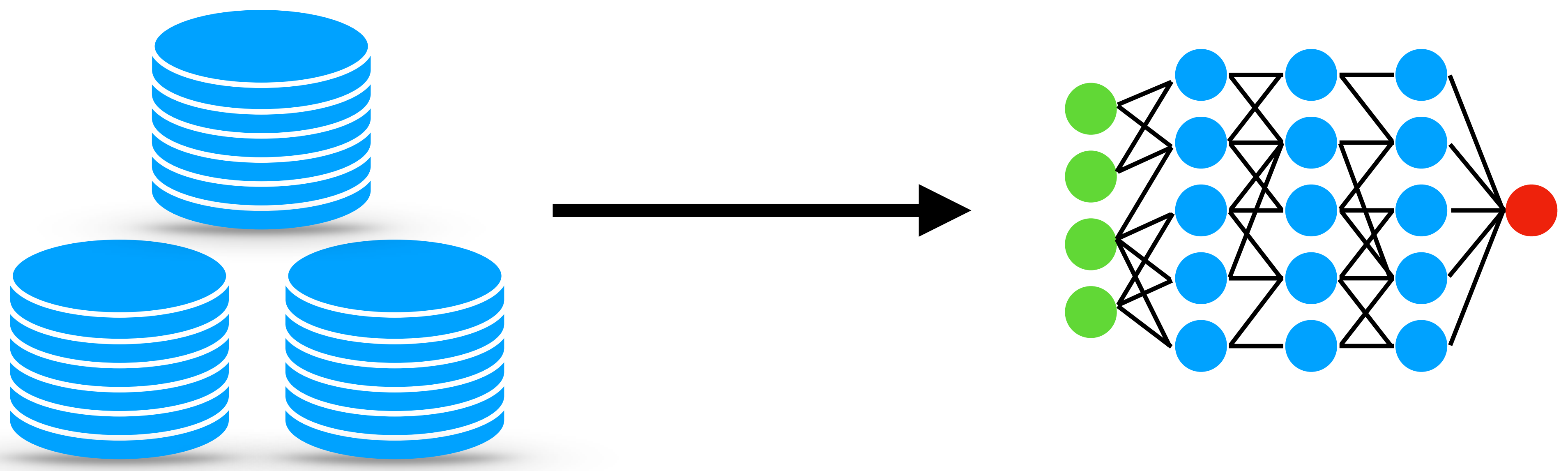
- Facial Recognition
- Grading Exams
- Resume Sorting
- Self Driving Cars

- Chatbots
- Manage Inventory
- Spam Filters
- Video Games

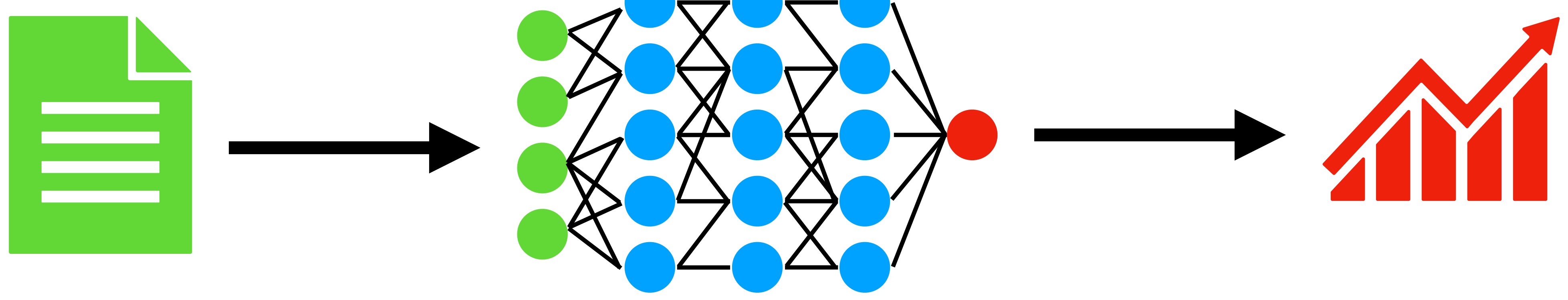
… many more

Machine Learning

Training



Inference



Applications

- Facial Recognition
- Grading Exams
- Resume Sorting
- Self Driving Cars

→ High Risk*

- Chatbots
- Manage Inventory
- Spam Filters
- Video Games

→ Low Risk*

… many more

As categorized by the EU AI Act

AI can cause serious harm

- Facial Recognition
- Grading Exams
- Resume Sorting
- Self Driving Cars

High Risk = Potential for serious harm

AI can cause serious harm

- Facial Recognition
- Grading Exams
- Resume Sorting
- Self Driving Cars

High Risk = Potential for serious harm

'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man

AI can cause serious harm

- Facial Recognition
- Grading Exams
- Resume Sorting
- Self Driving Cars

High Risk = Potential for serious harm

'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man

How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud

AI can cause serious harm

- Facial Recognition
- Grading Exams
- Resume Sorting
- Self Driving Cars

High Risk = Potential for serious harm

'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man

How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud

17 fatalities, 736 crashes: The shocking toll of Tesla's Autopilot

AI can cause serious harm

- Facial Recognition
- Grading Exams
- Resume Sorting
- Self Driving Cars

High Risk = Potential for serious harm

'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man

How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud

17 fatalities, 736 crashes: The shocking toll of Tesla's Autopilot

UnitedHealth uses AI model with 90% error rate to deny care, lawsuit alleges

For the largest health insurer in the US, AI's error rate is like a feature, not a bug.

Many more: 🔗 incidentdatabase.ai

Unanimous demand for regulation

Unanimous demand for regulation

- **EU:** Artificial Intelligence Act
- **NIST:** Risk Management Framework
- **USA:** Biden’s Executive Order
- **Canada**: AIDA
- **China:** AI Governance Initiative
- + local US state laws

… more incoming

Unanimous demand for regulation

- **EU:** Artificial Intelligence Act
- **NIST:** Risk Management Framework
- **USA:** Biden’s Executive Order
- **Canada**: AIDA
- **China:** AI Governance Initiative
- + local US state laws

… more incoming

Common Requirements*

*Technical details unspecified

Unanimous demand for regulation

- **EU:** Artificial Intelligence Act
- **NIST:** Risk Management Framework
- **USA:** Biden's Executive Order
- **Canada**: AIDA
- **China:** AI Governance Initiative
- + local US state laws

… more incoming

Common Requirements*

- **High quality datasets**
    - Eg: Facial Recognition dataset — demographic diversity, such as age, gender, race etc.

*Technical details unspecified

Unanimous demand for regulation

- **EU:** Artificial Intelligence Act
- **NIST:** Risk Management Framework
- **USA:** Biden's Executive Order
- **Canada**: AIDA
- **China:** AI Governance Initiative
- + local US state laws

… more incoming

Common Requirements*

- **High quality datasets**
  - Eg: Facial Recognition dataset — demographic diversity, such as age, gender, race etc.
- **Procedural Regularity for Decisions**
  - Same algorithm used for all individuals
  - Reproducible, including randomness

*Technical details unspecified

Unanimous demand for regulation

- **EU:** Artificial Intelligence Act
- **NIST:** Risk Management Framework
- **USA:** Biden's Executive Order
- **Canada**: AIDA
- **China:** AI Governance Initiative
- + local US state laws

… more incoming

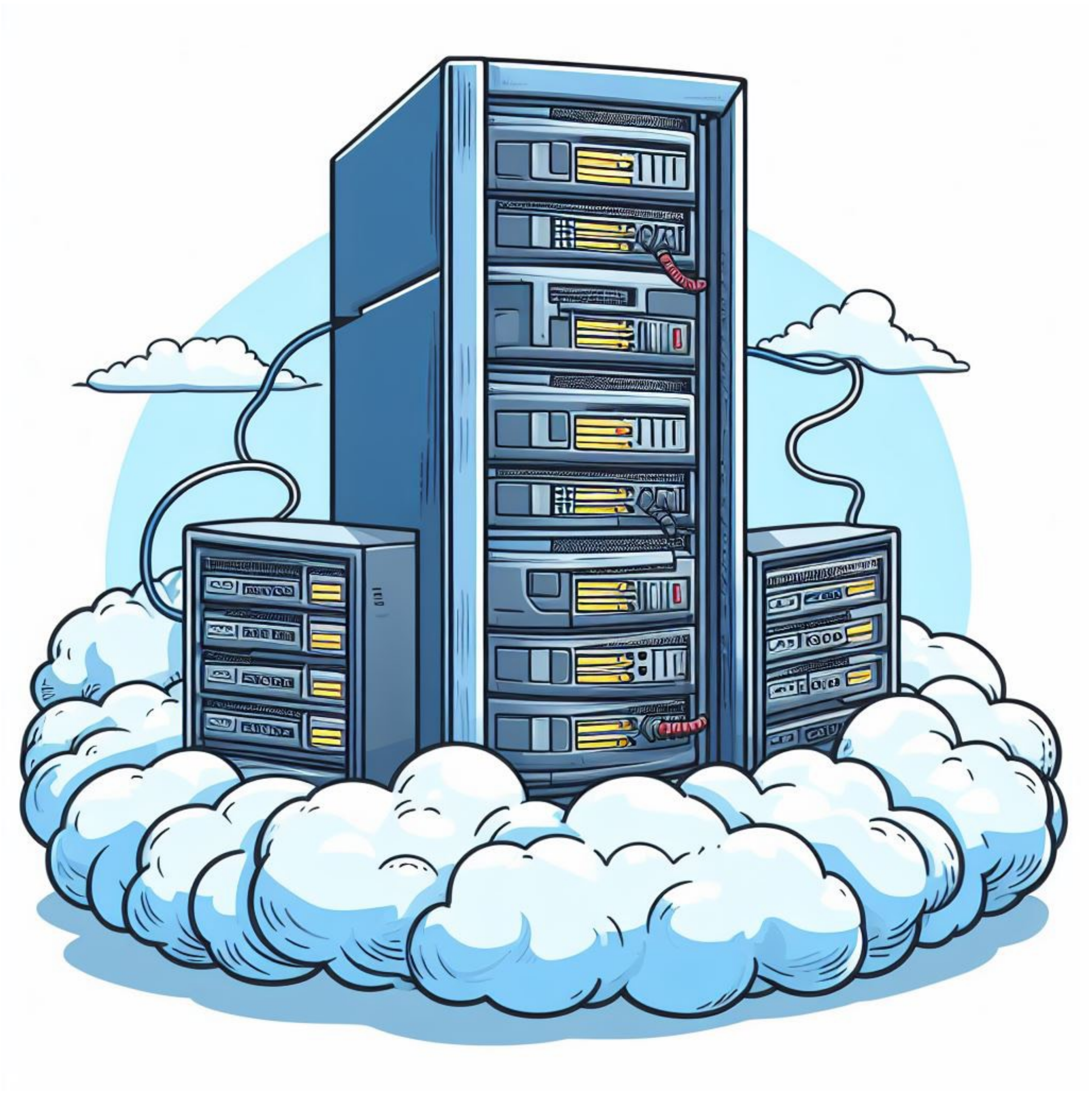
Common Requirements*

- **High quality datasets**
  - Eg: Facial Recognition dataset — demographic diversity, such as age, gender, race etc.
- **Procedural Regularity for Decisions**
  - Same algorithm used for all individuals
  - Reproducible, including randomness
- **Privacy**
  - Need to preserve privacy of data and model to comply with data privacy laws
  - Companies may not want to leak IP
  - Prevent gaming of system

*Technical details unspecified

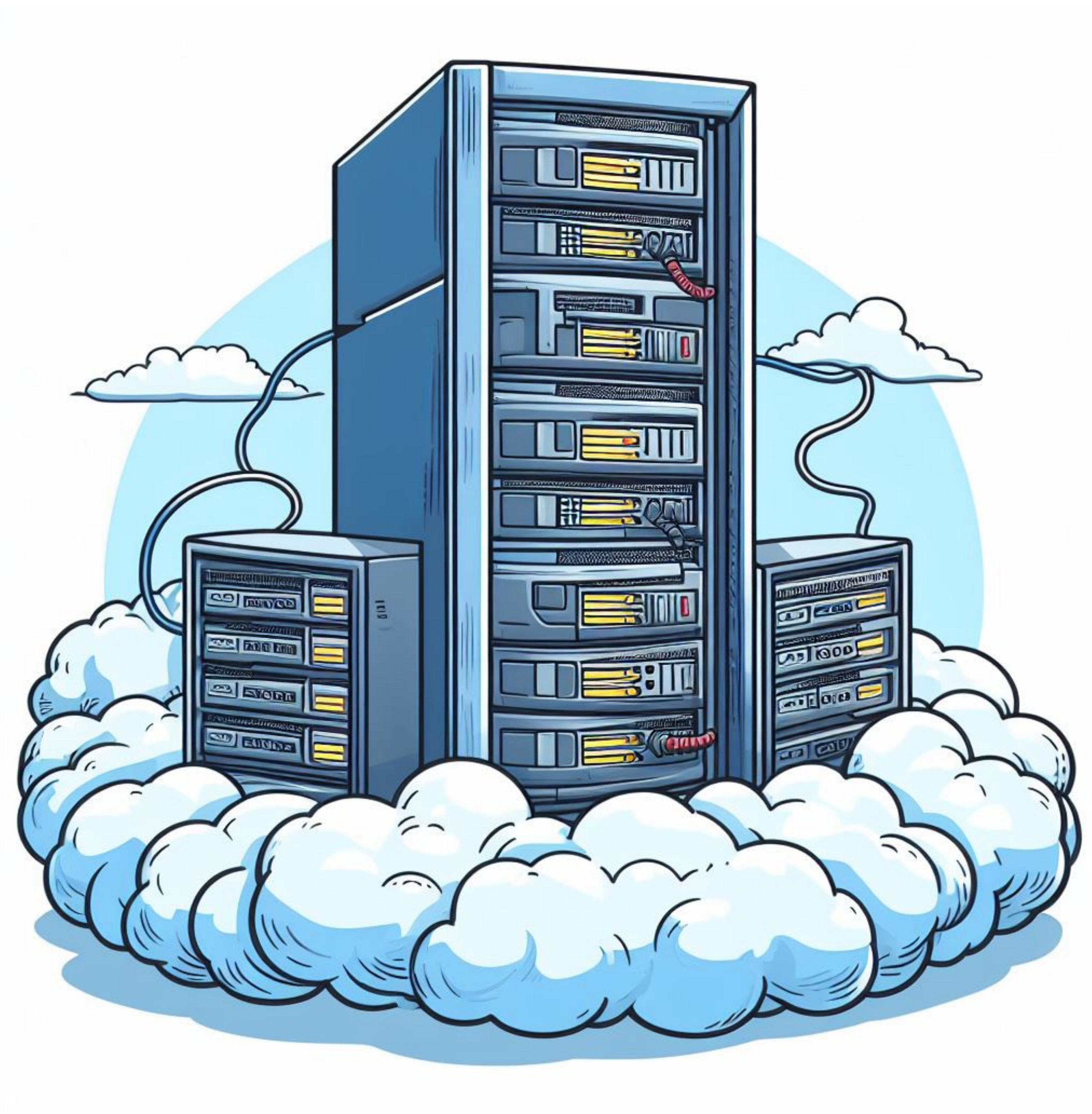
How does AI + Regulation look?

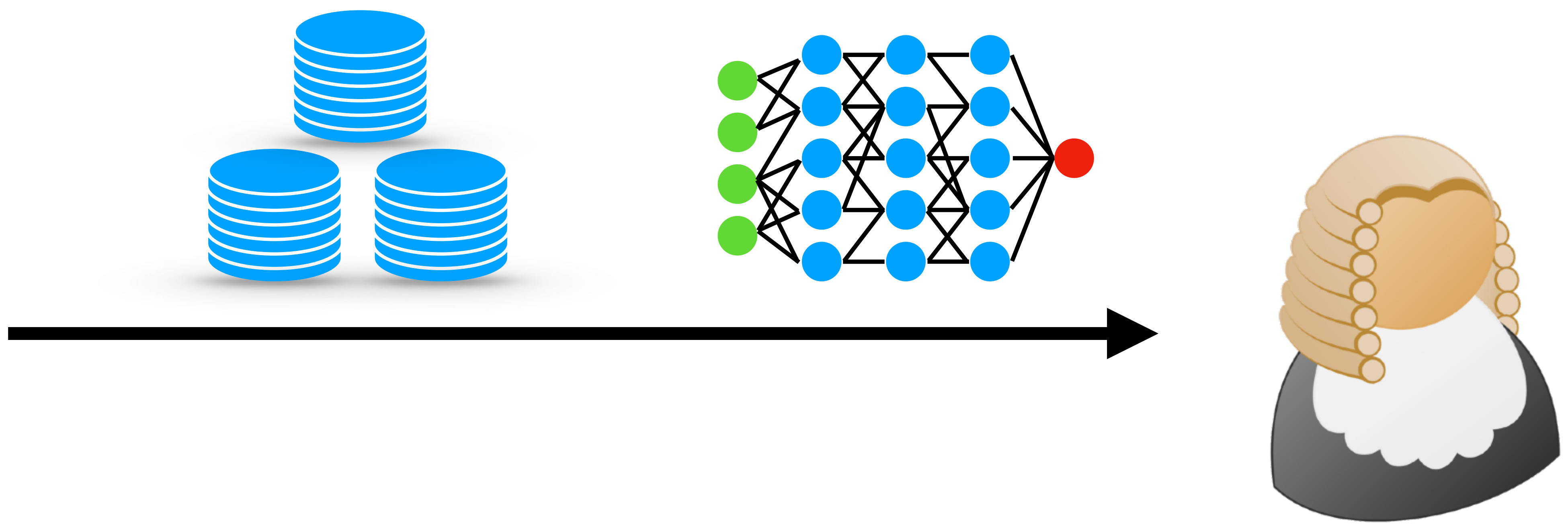
Potential approach: Independent auditor certifies compliance



How does AI + Regulation look?

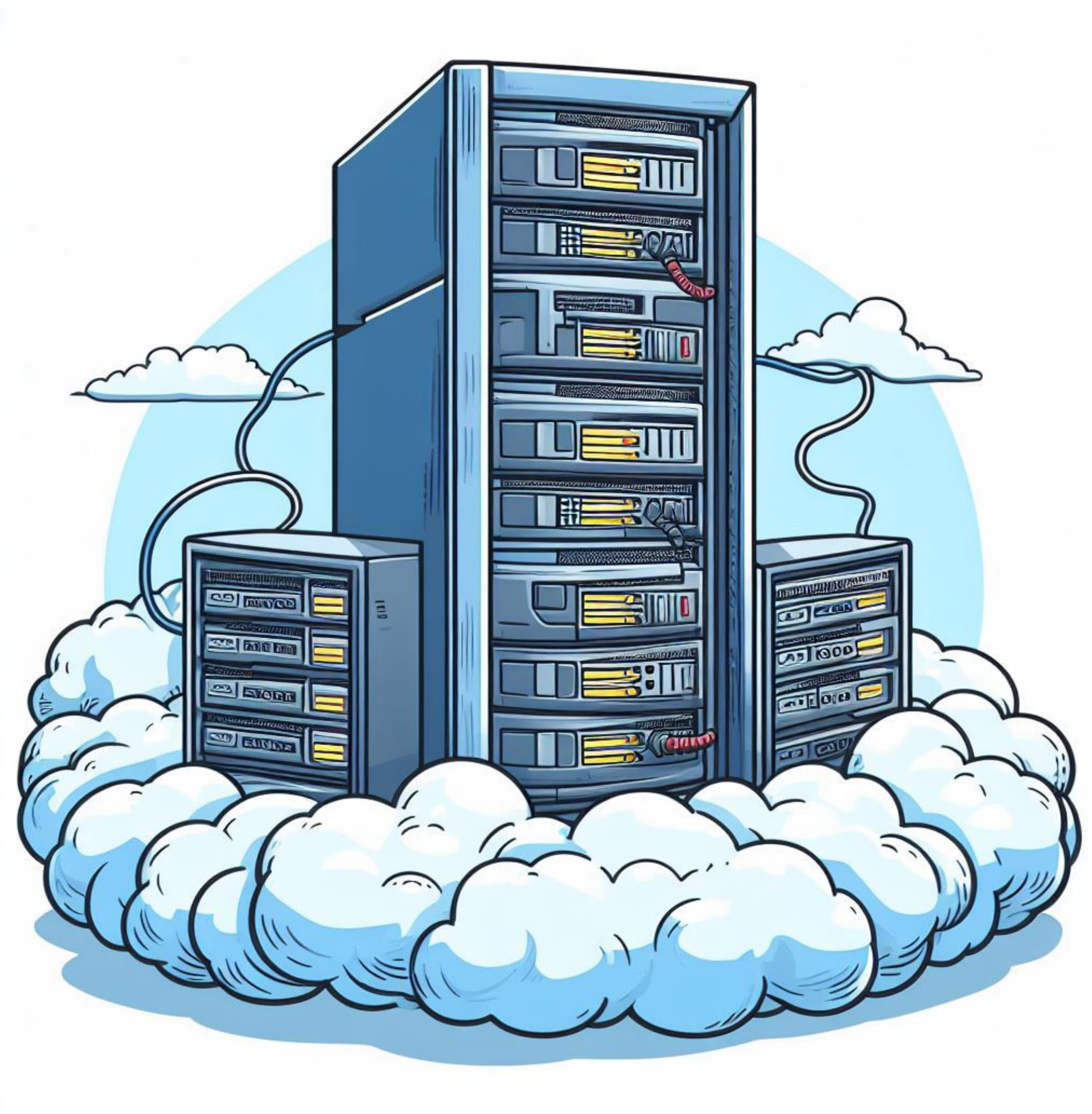
Potential approach: Independent auditor certifies compliance

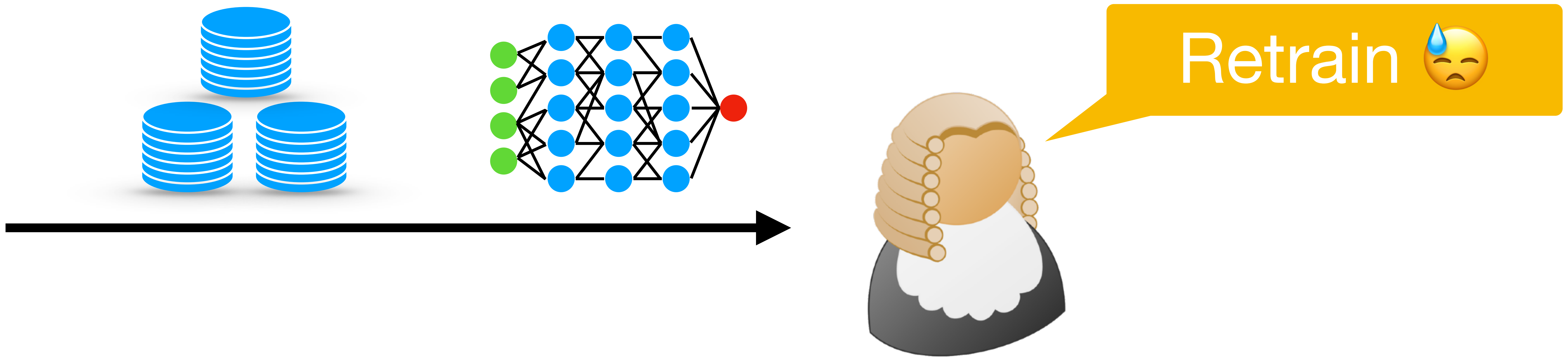




How does AI + Regulation look?

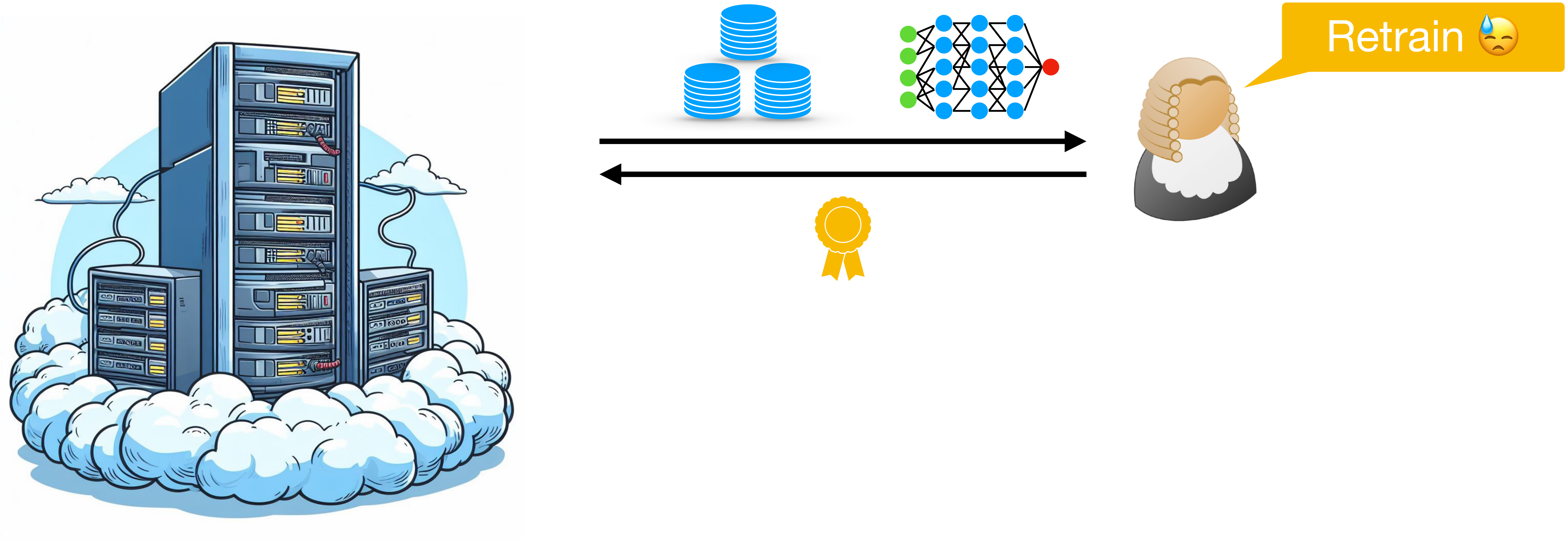
Potential approach: Independent auditor certifies compliance





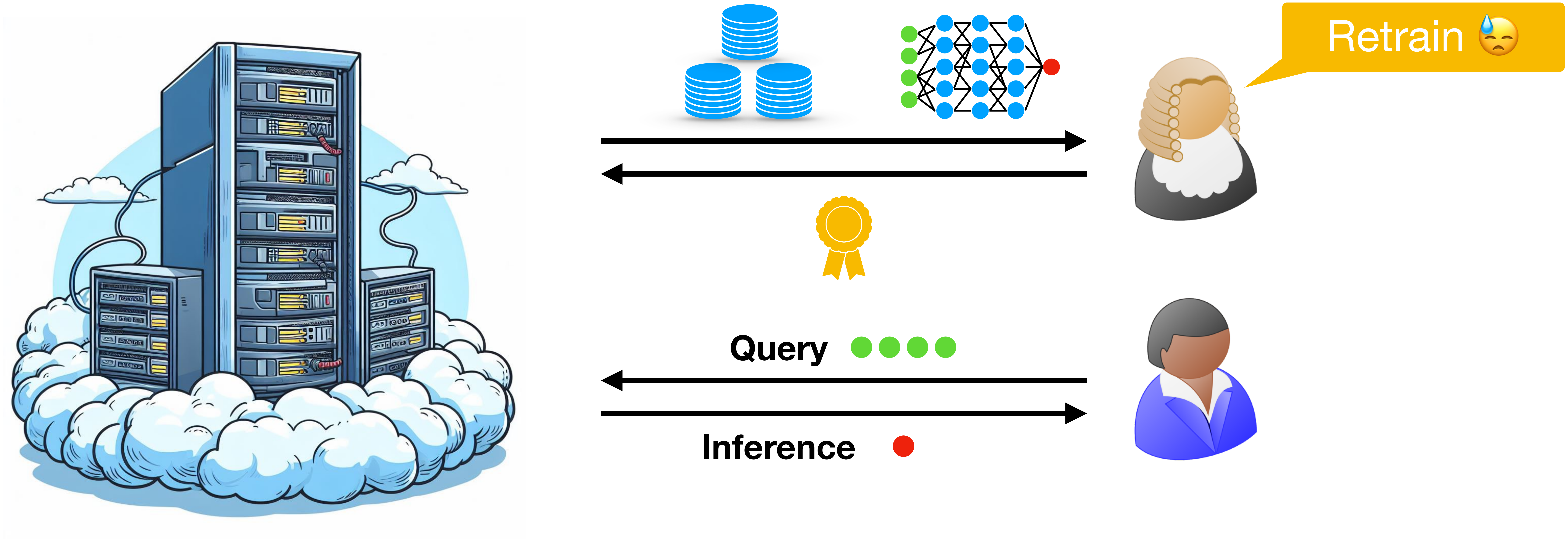
How does AI + Regulation look?

Potential approach: Independent auditor certifies compliance



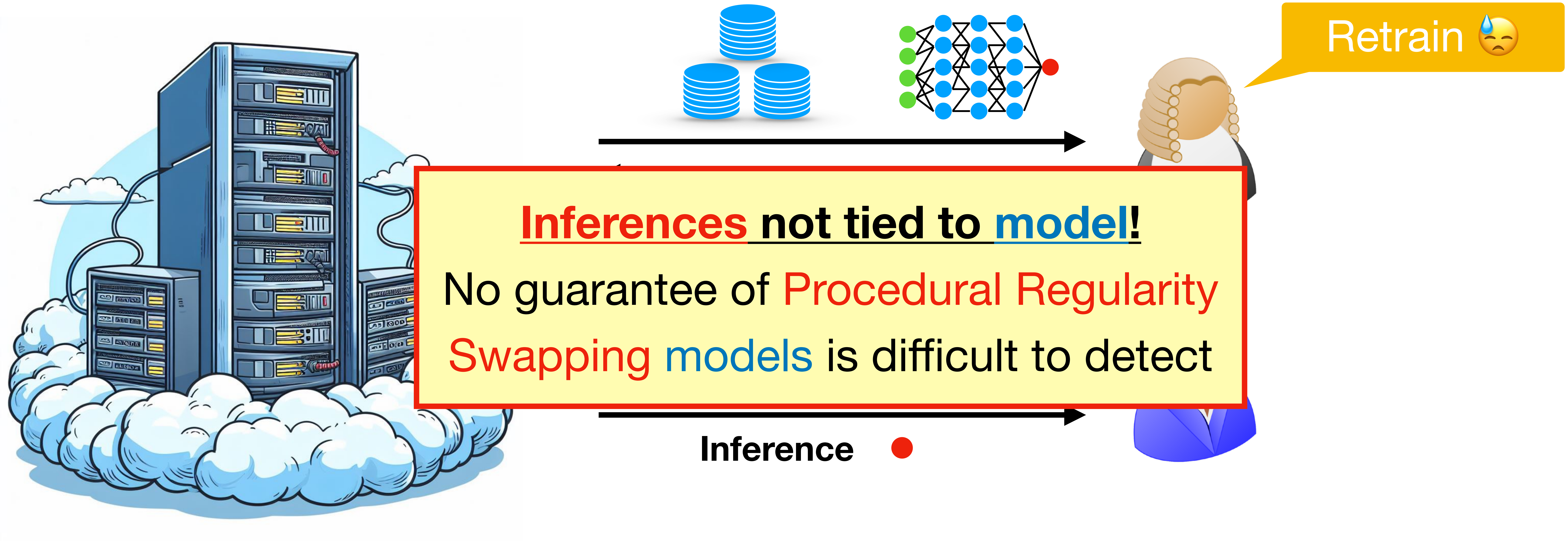
How does AI + Regulation look?

Potential approach: Independent auditor certifies compliance



How does AI + Regulation look?

Potential approach: Independent auditor certifies compliance



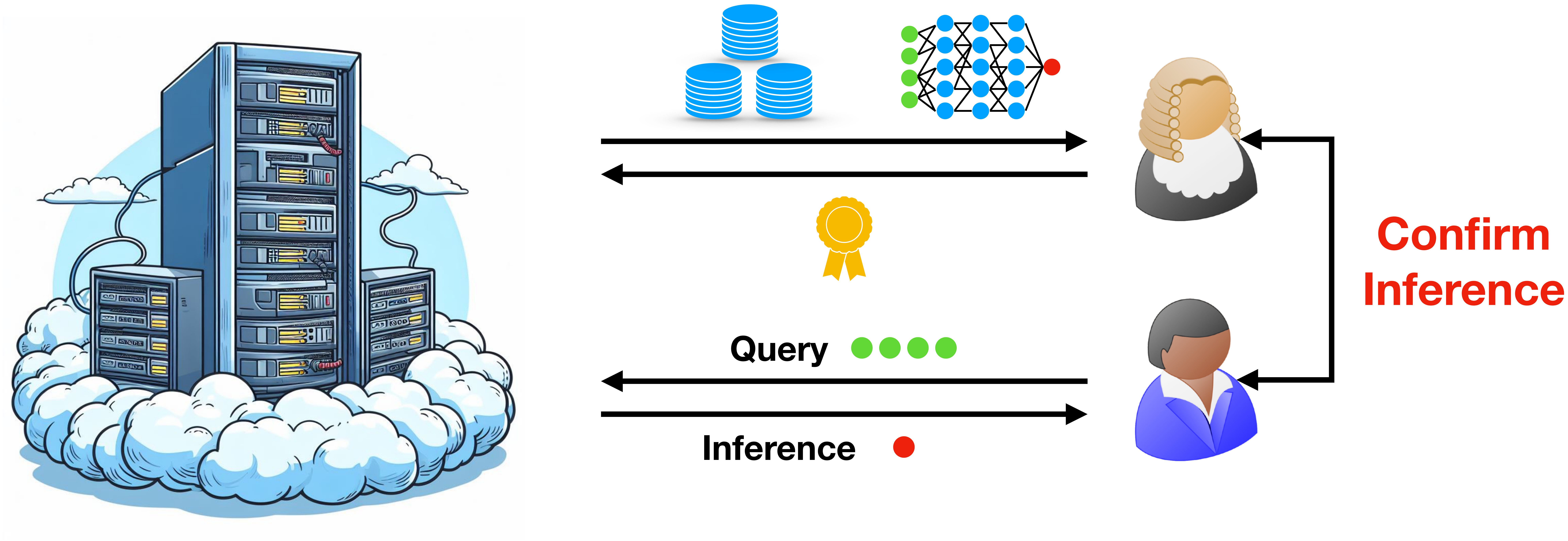
Inferences not tied to model!

No guarantee of Procedural Regularity

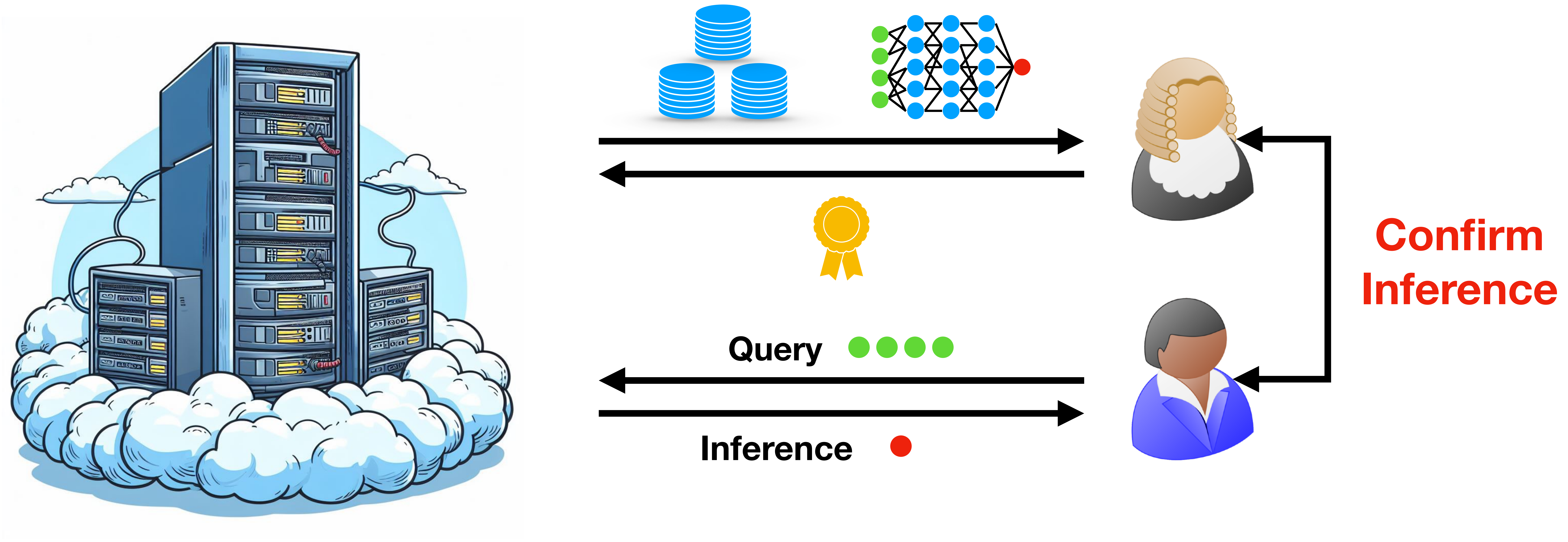
Swapping models is difficult to detect

How does AI + Regulation look?

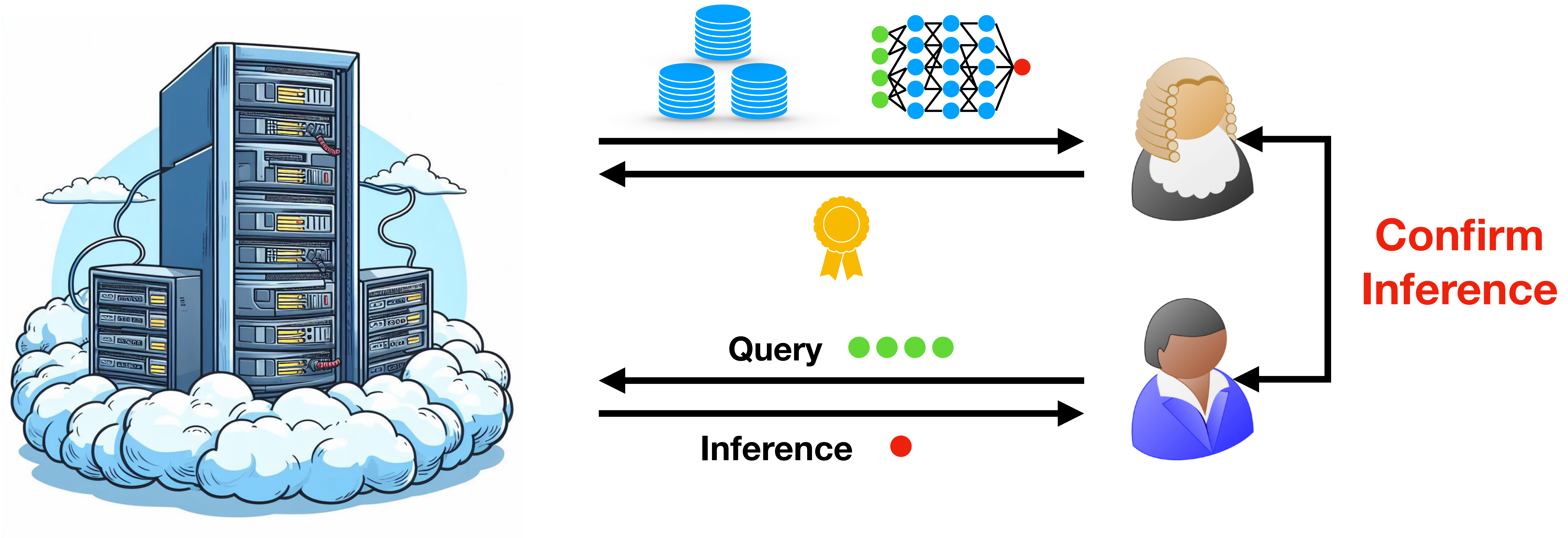
Potential approach: Independent auditor certifies compliance



How does AI + Regulation look?

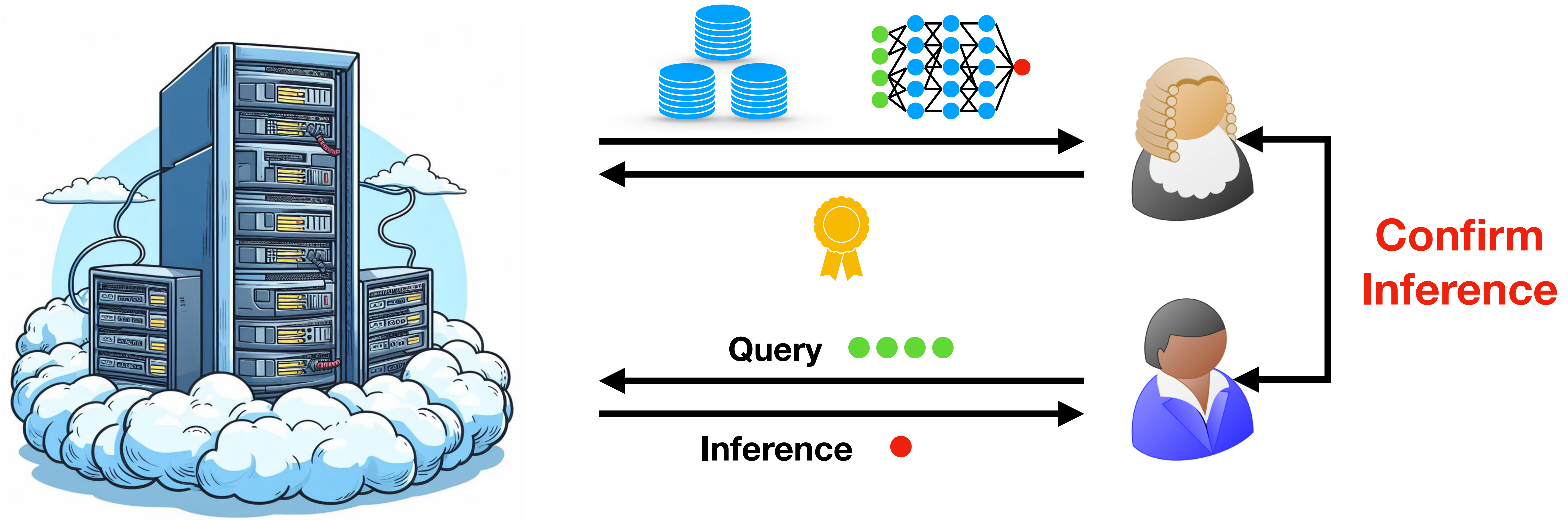


How does AI + Regulation look?



Other Problems:

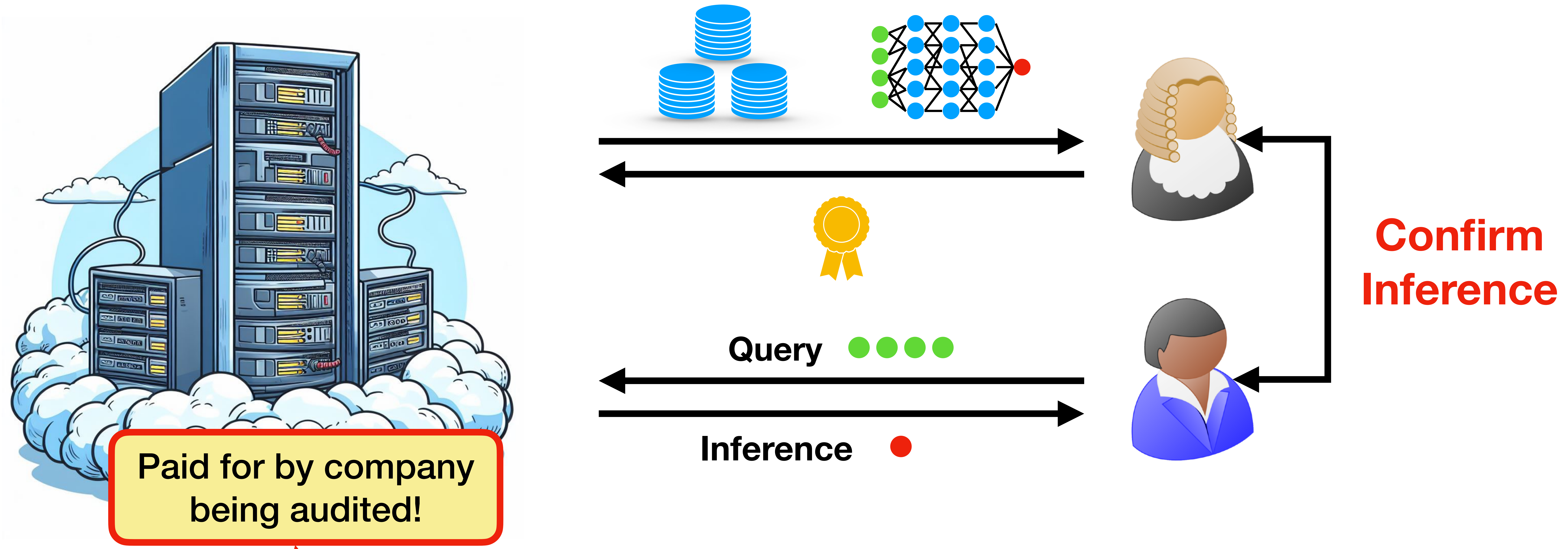
How does AI + Regulation look?



Other Problems:

- Confirmation is interactive. Auditor stores model and re-runs inference.

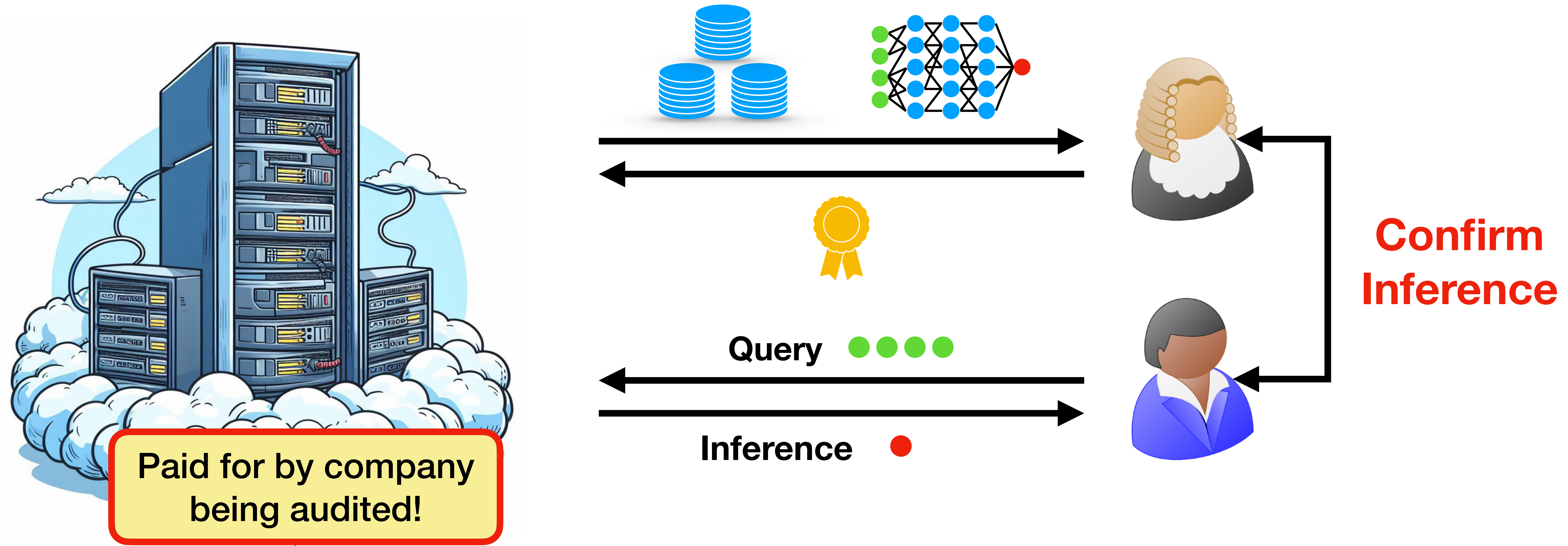
How does AI + Regulation look?



Other Problems:

- Confirmation is interactive. Auditor stores model and re-runs inference.
- What if the auditor was coerced? Want public verifiability.

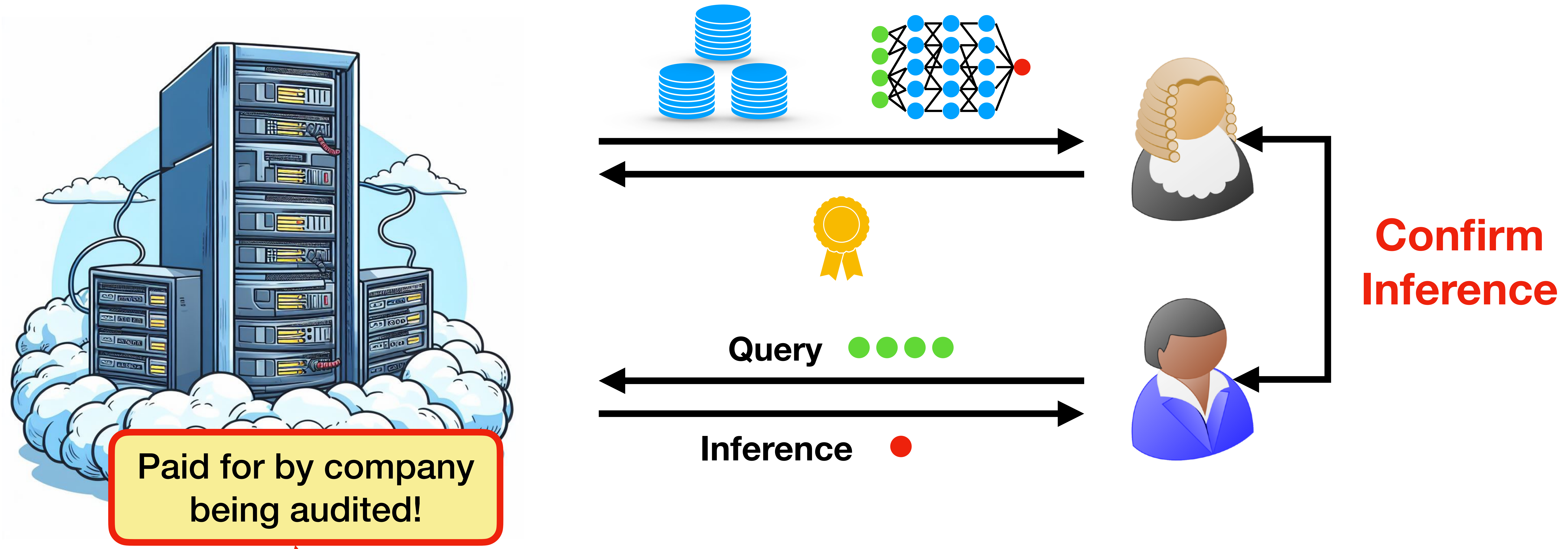
How does AI + Regulation look?



Other Problems:

- Confirmation is interactive. Auditor stores model and re-runs inference.
- What if the auditor was coerced? Want public verifiability.
- Models continuously change — auditors are expensive.

How does AI + Regulation look?



Other Problems:

- Confirmation is interactive. Auditor stores model and re-runs inference.
- What if the auditor was coerced? Want public verifiability.
- Models continuously change — auditors are expensive.
- Companies may not want to reveal model, even to auditors.

Financial vs AI Compliance

Financial vs AI Compliance

Finance:

AI:

Financial vs AI Compliance

Finance:

The “final” product (balance sheet) is certified by auditors.

AI:

Financial vs AI Compliance

Finance:

The “final” product (balance sheet) is certified by auditors.
Everyone sees the same thing.

AI:

Financial vs AI Compliance

Finance:

The "final" product (balance sheet) is certified by auditors.
Everyone sees the same thing.

AI:

The "final" product (inference) is a function of model certified by auditors.

Financial vs AI Compliance

Finance:

The "final" product (balance sheet) is certified by auditors.
Everyone sees the same thing.

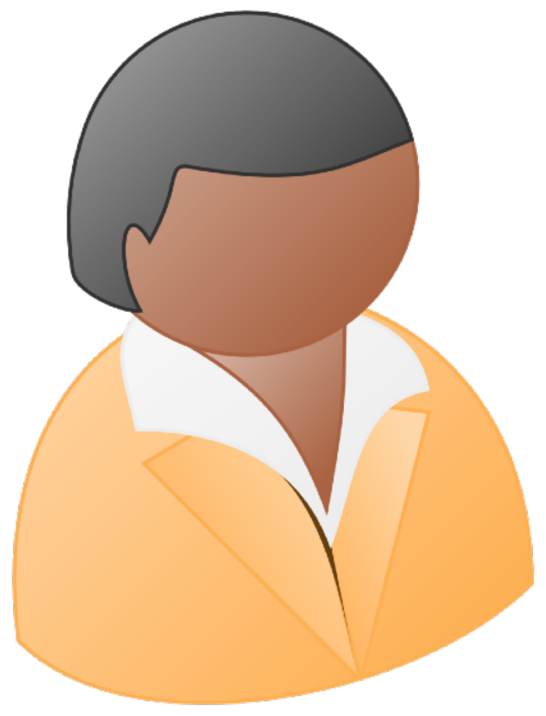
Guarantees provided by an auditor is <u>strictly</u> weaker in AI Compliance!

AI:

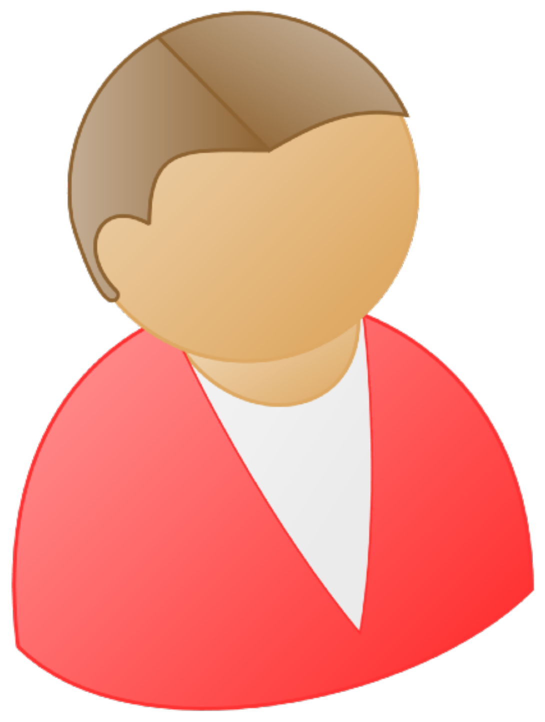
The "final" product (inference) is a function of model certified by auditors.

How can Cryptography help with AI regulation compliance?

ZK Proofs to the rescue



Prover



Verifier

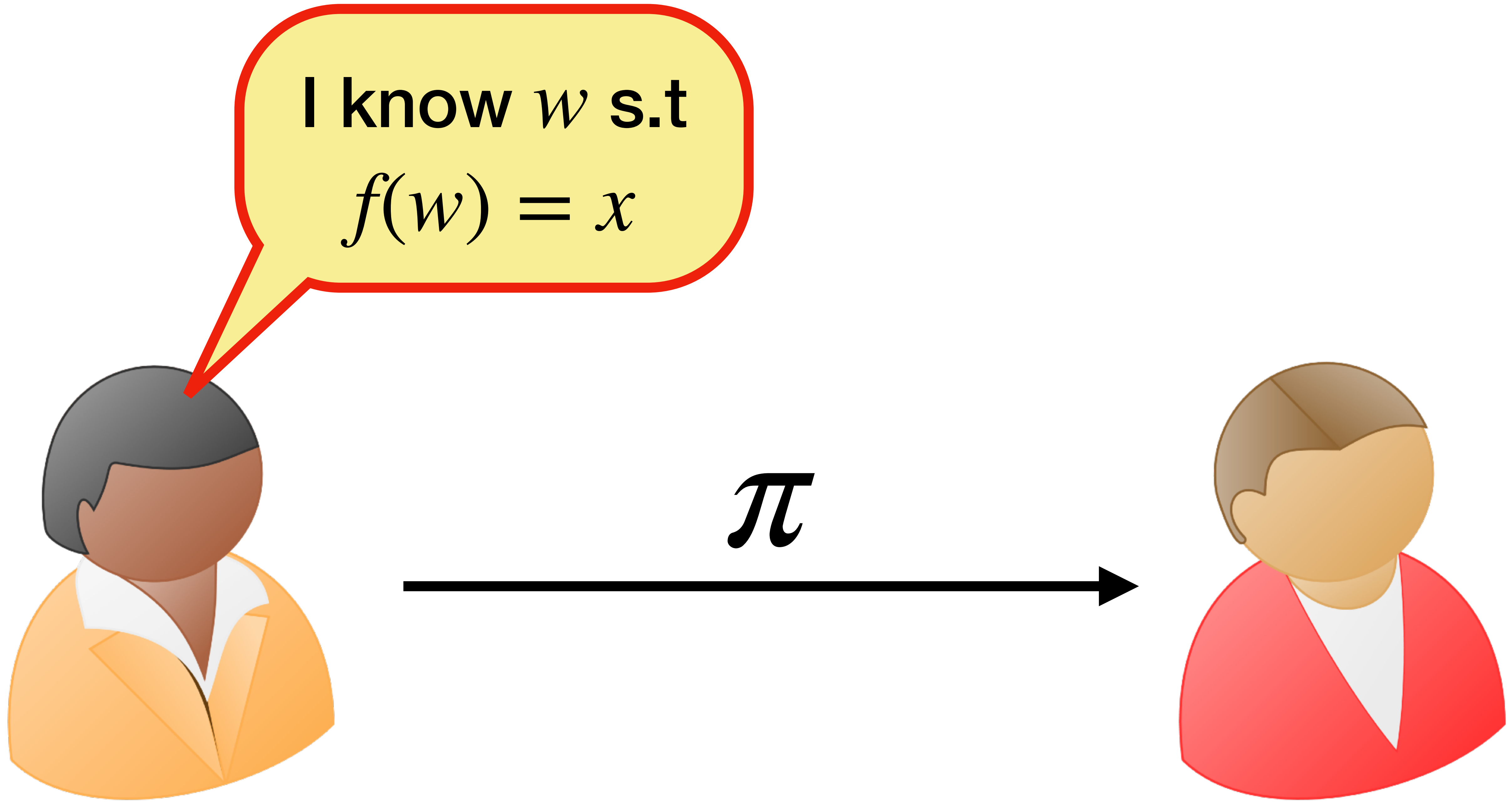
ZK Proofs to the rescue

Prover

Verifier

Public: $f(\cdot)$, output x

ZK Proofs to the rescue

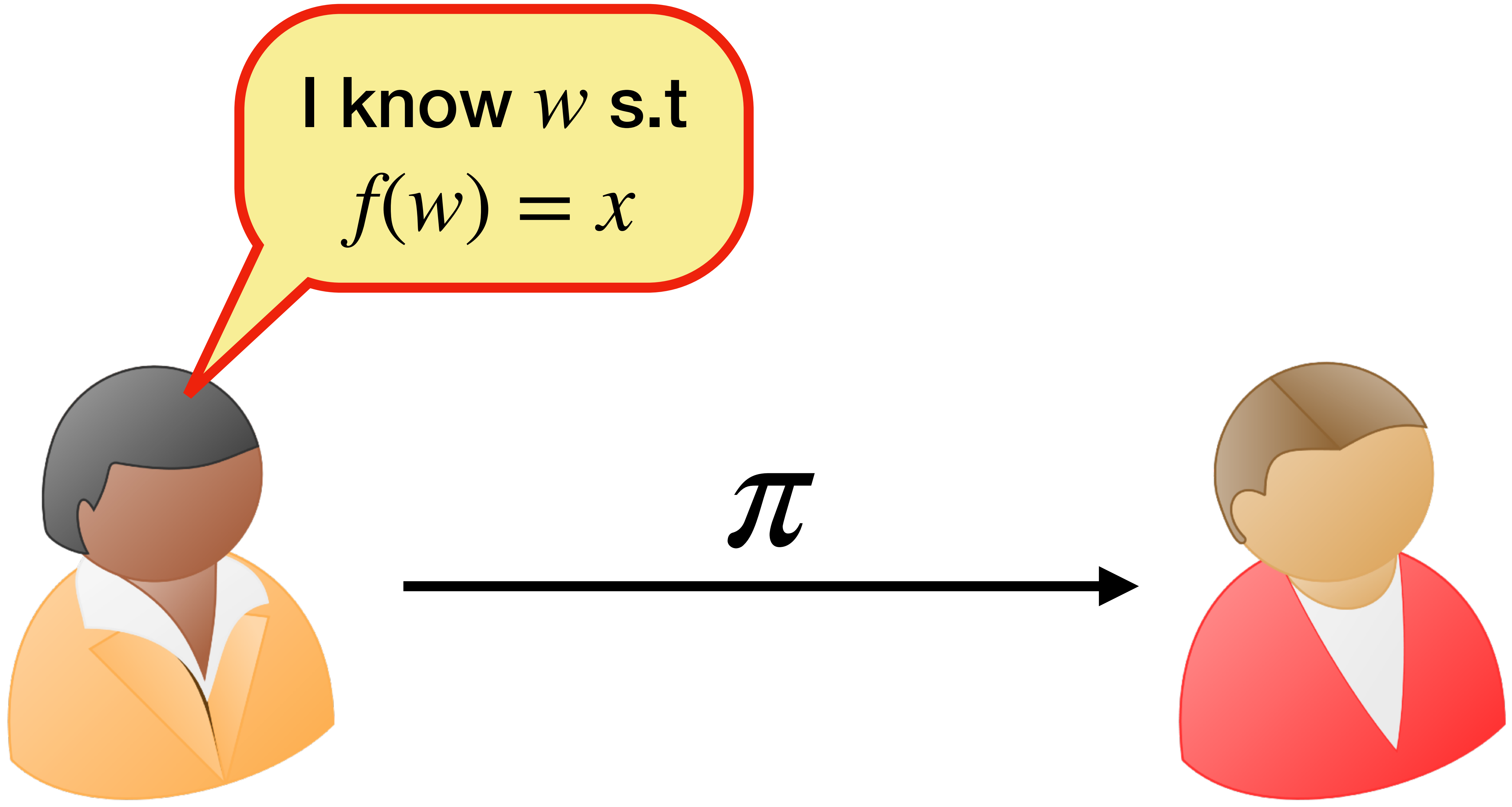


Prover **Verifier**

Public: $f(\cdot)$, output x

ZK Proofs to the rescue

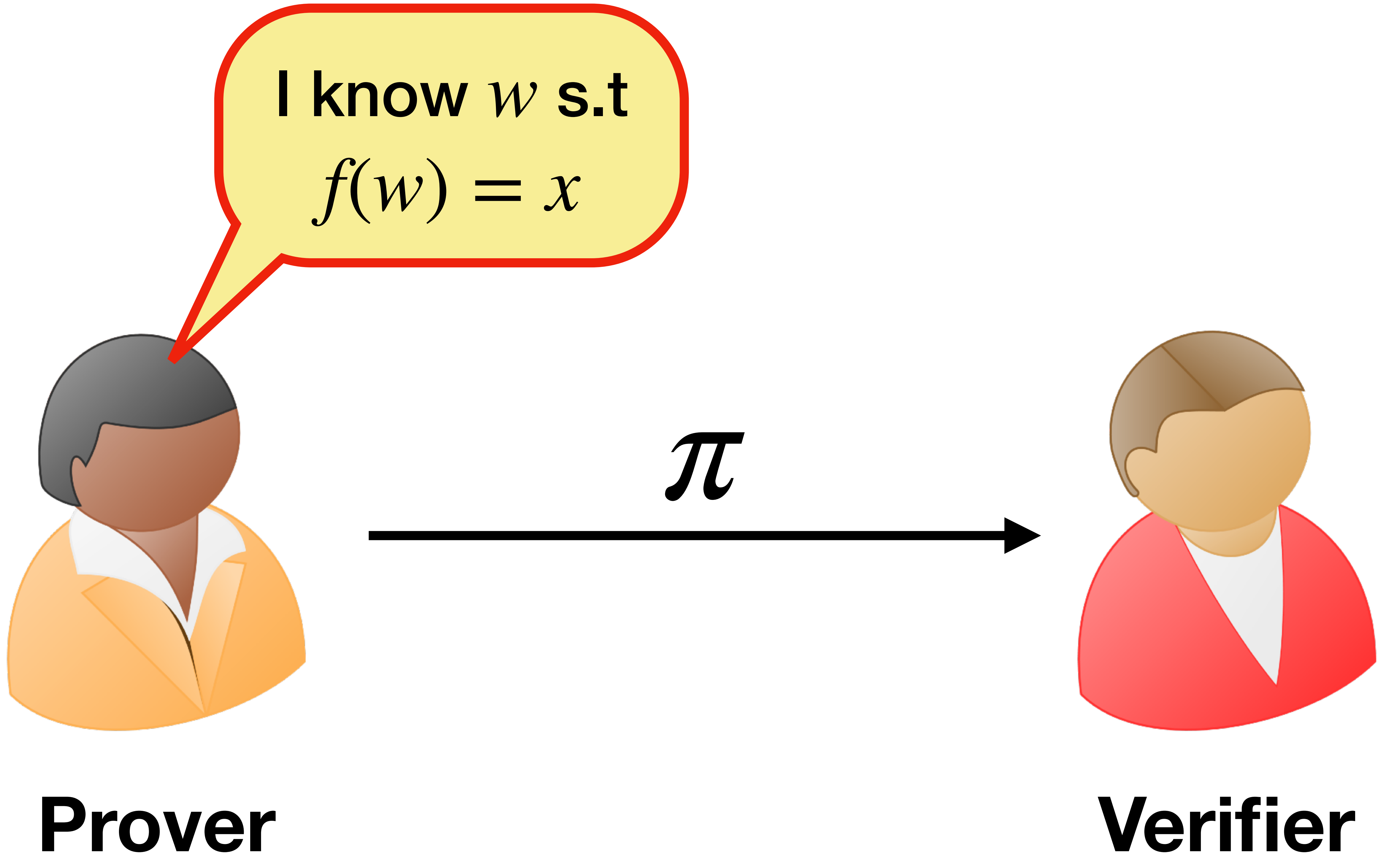
f can be any function (ML training)



Prover **Verifier**

Public: $f(\,\cdot\,)$, output x

ZK Proofs to the rescue



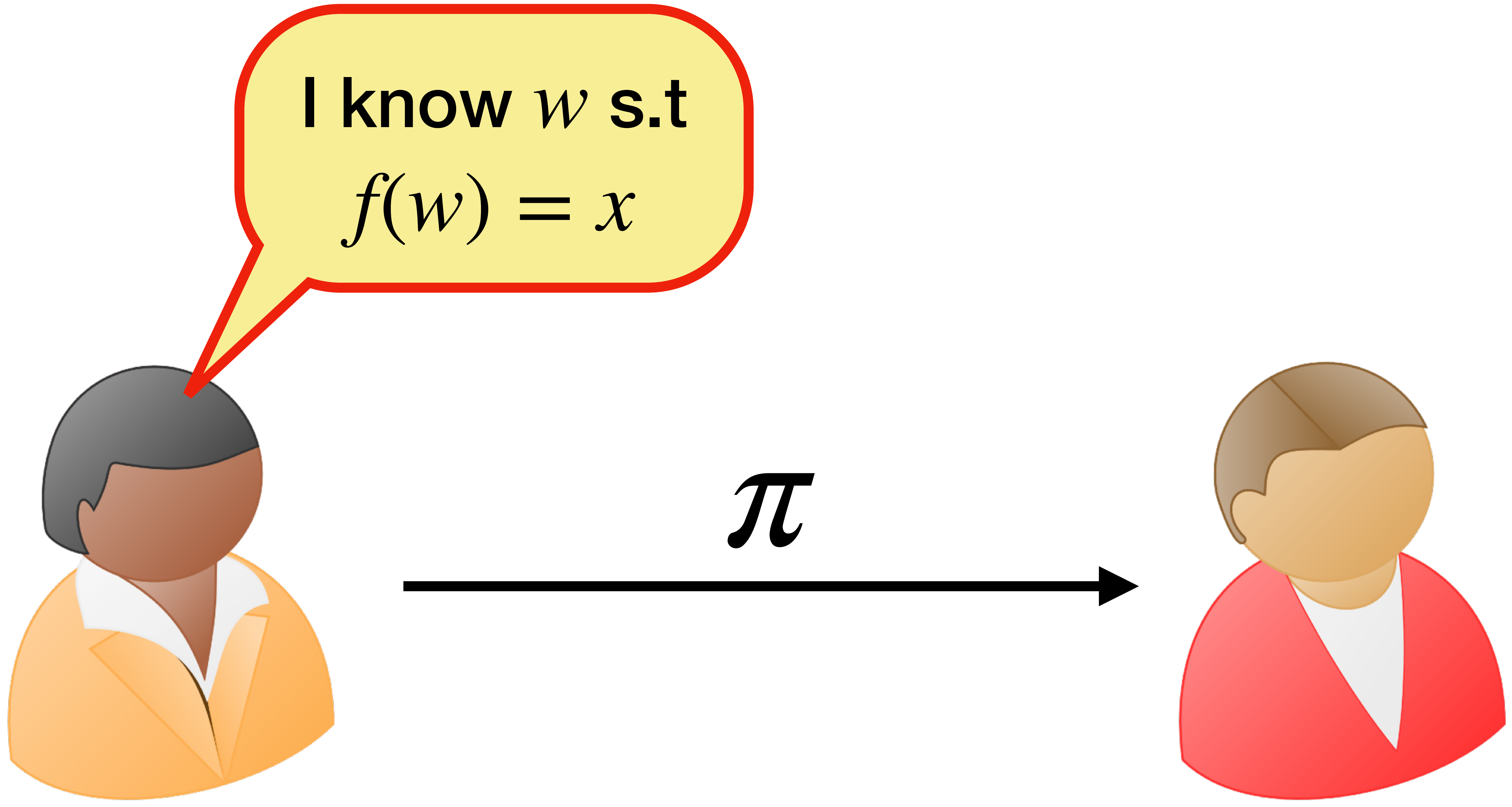
f can be any function (ML training)

Zero-Knowledge

Verifier learns nothing about w

Public: $f(\cdot)$, output x

ZK Proofs to the rescue



Prover **Verifier**

Public: $f(\,\cdot\,)$, output x

f can be any function (ML training)

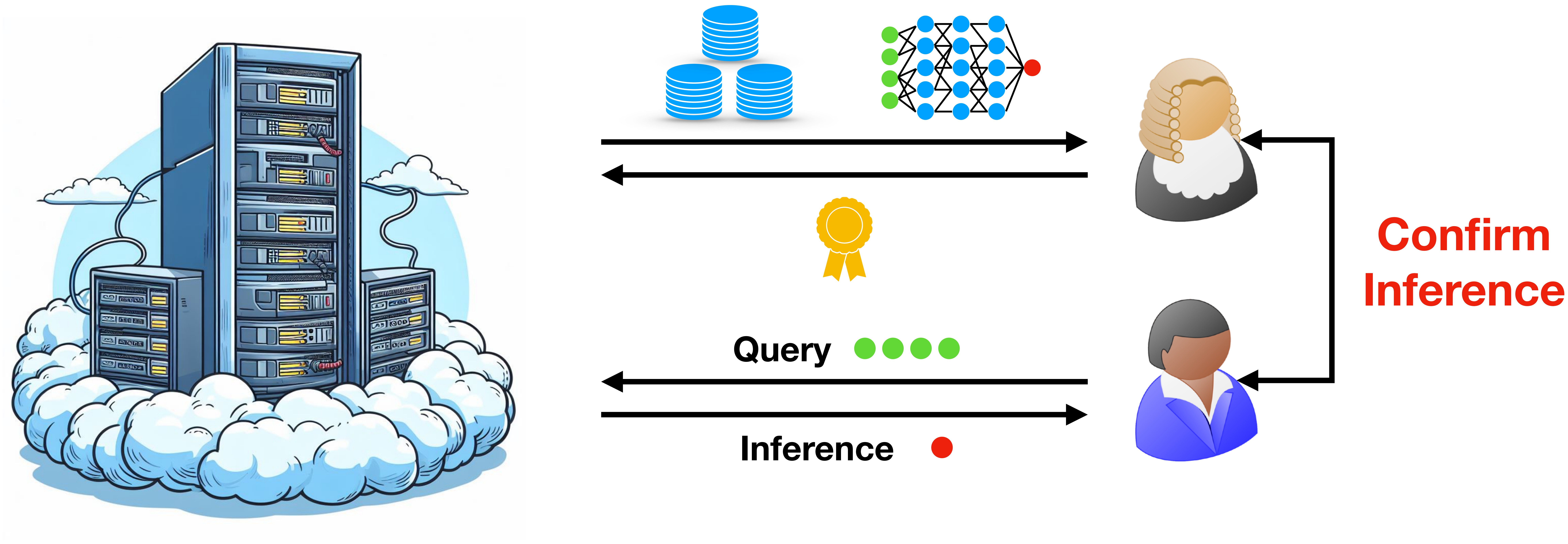
Zero-Knowledge

Verifier learns nothing about w

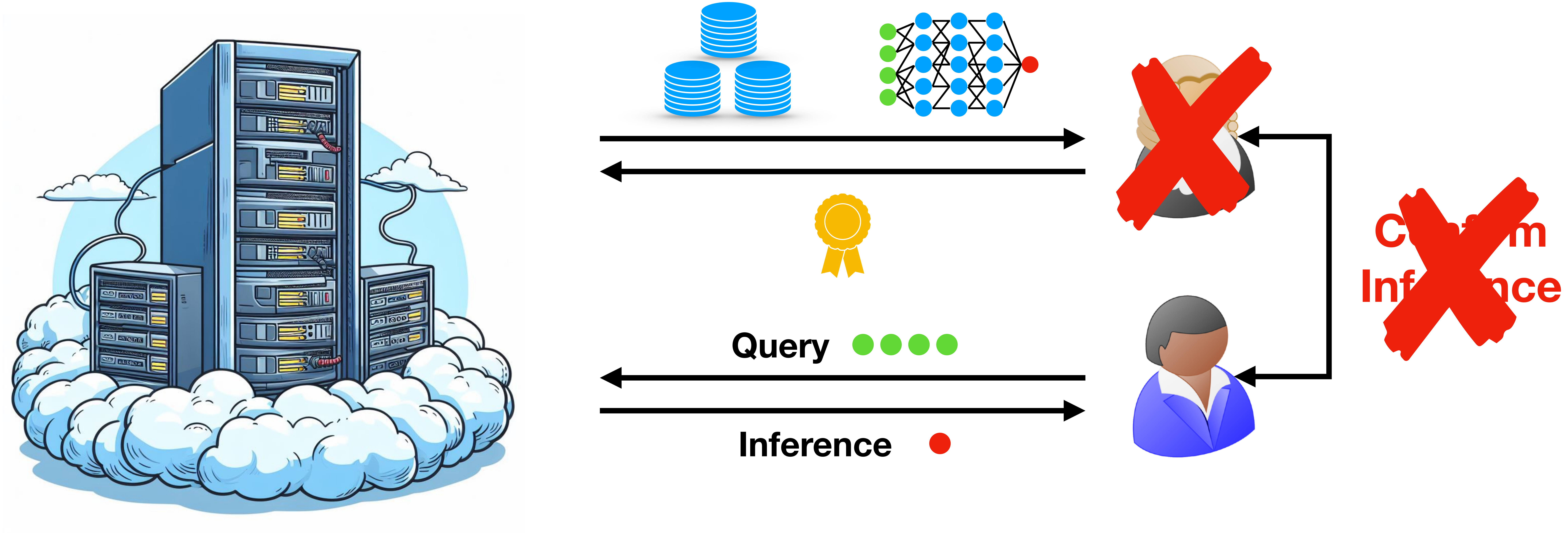
Soundness

Cheating prover cannot produce π if they don't know $w : f(w) = x$

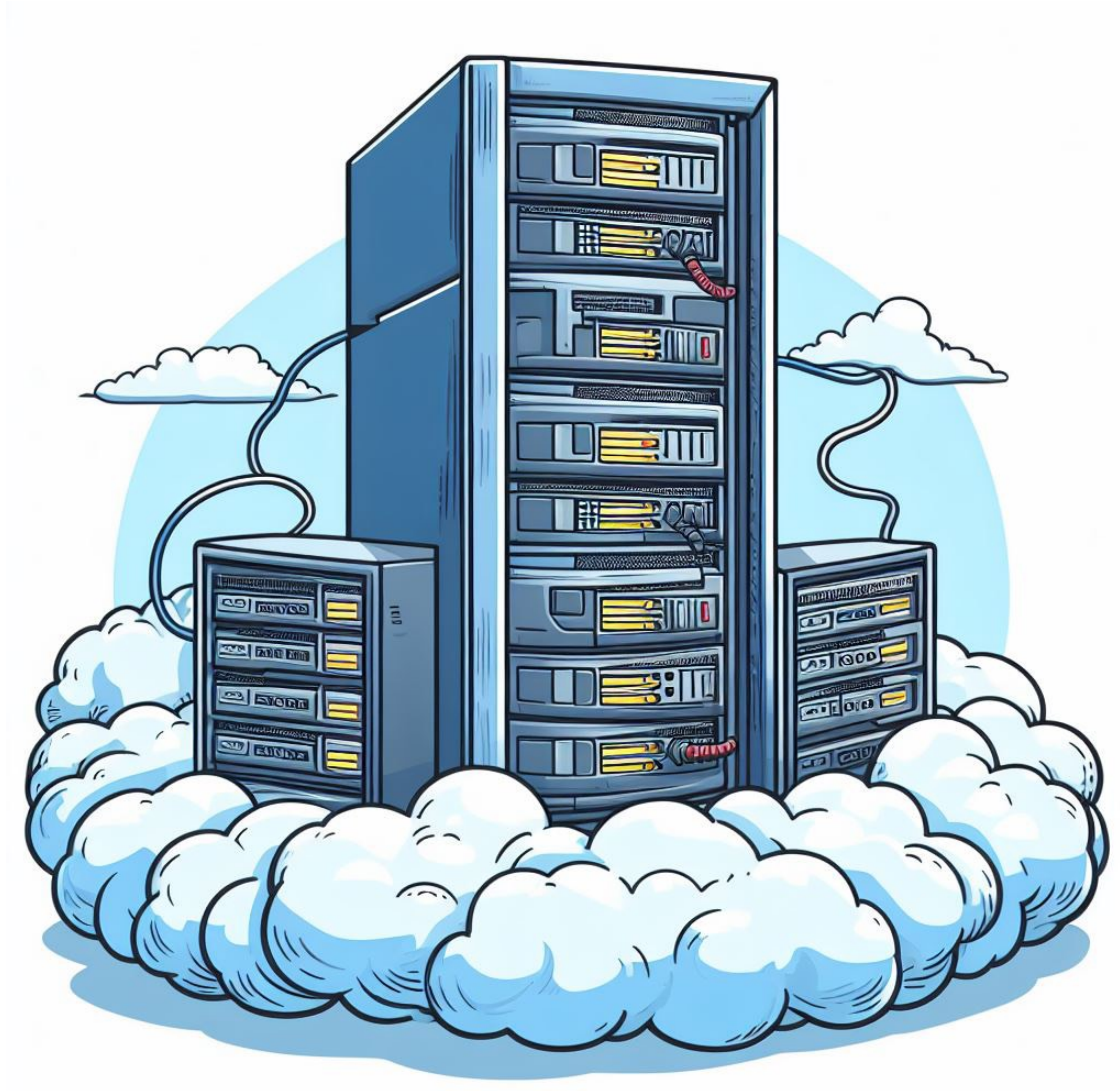
AI + Regulation + ZKPs

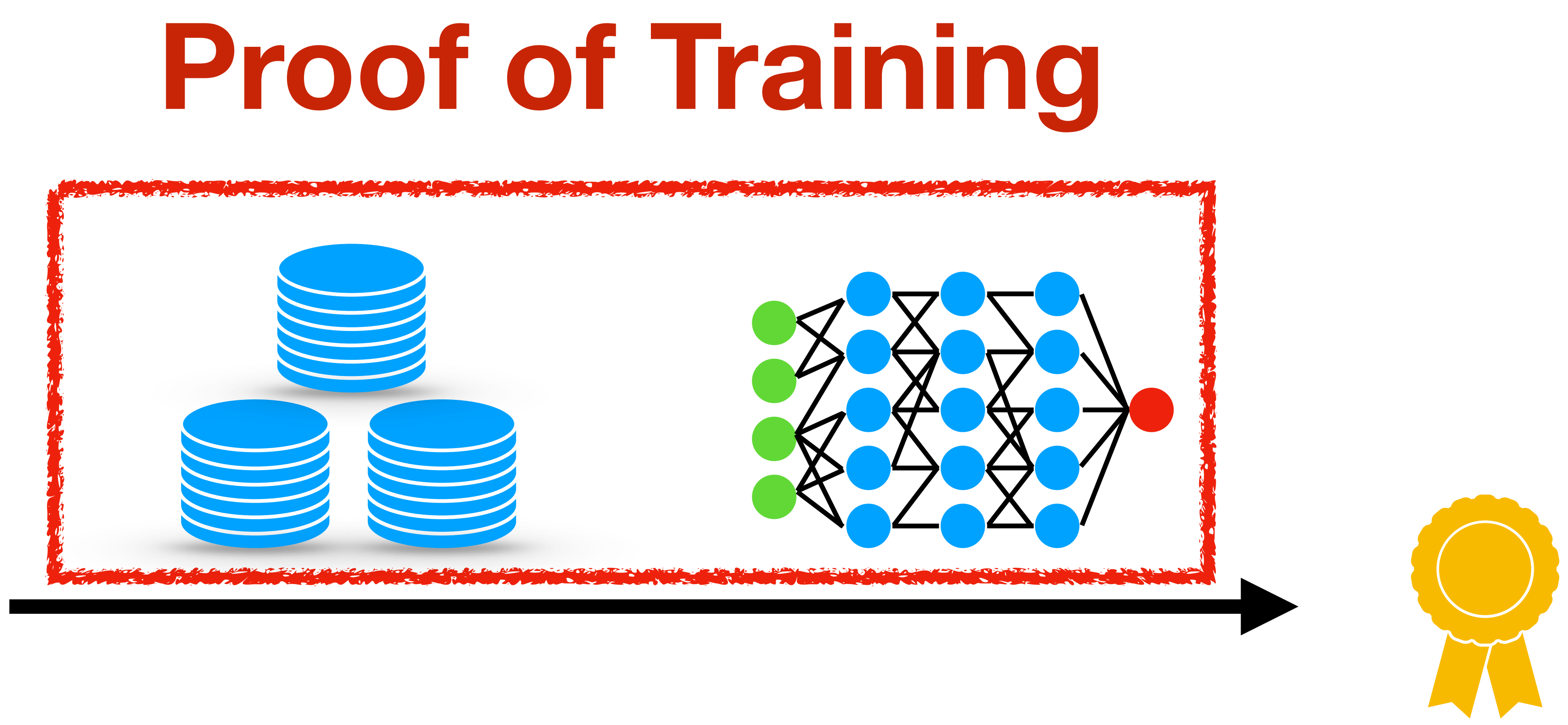


AI + Regulation + ZKPs

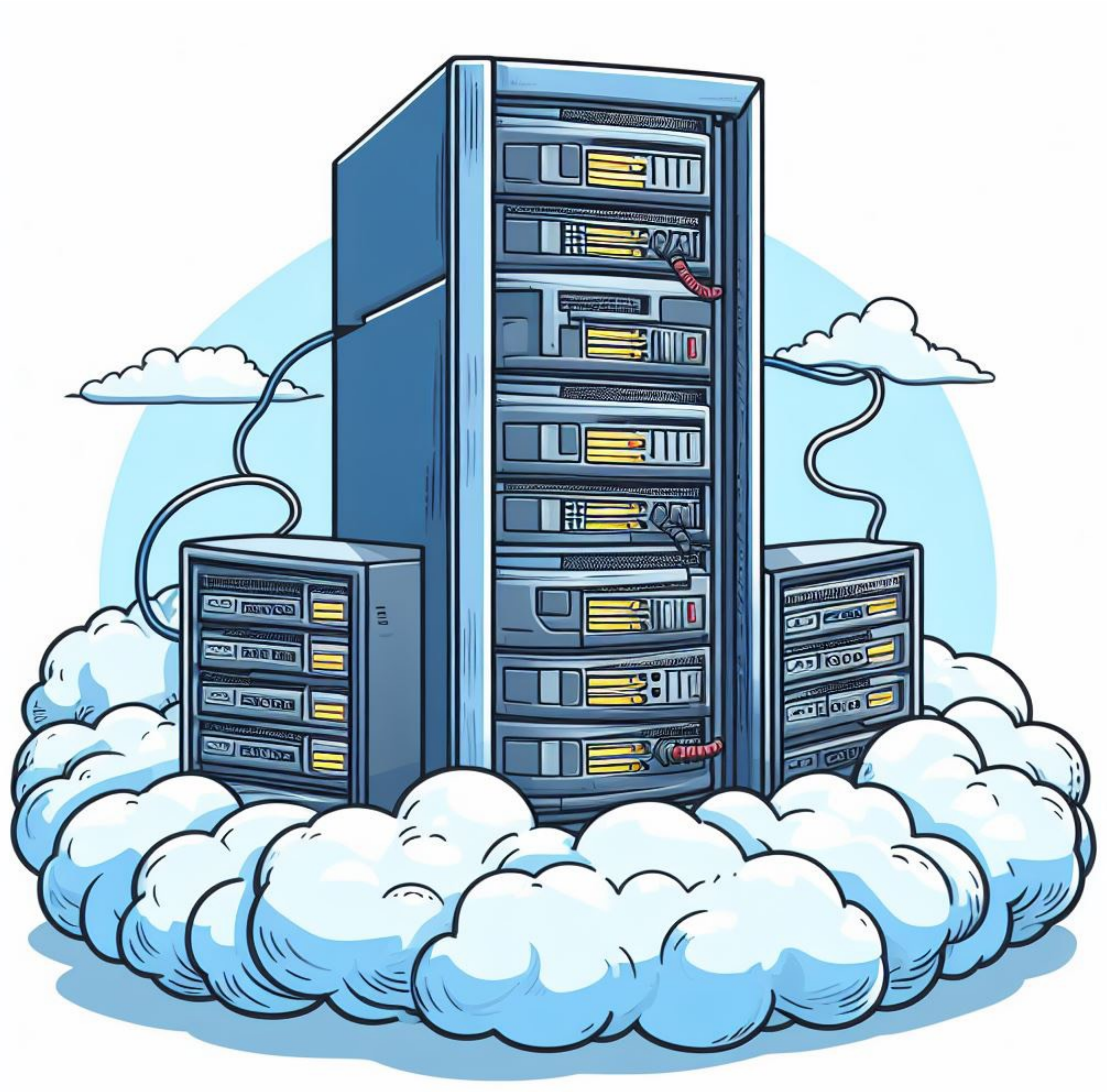


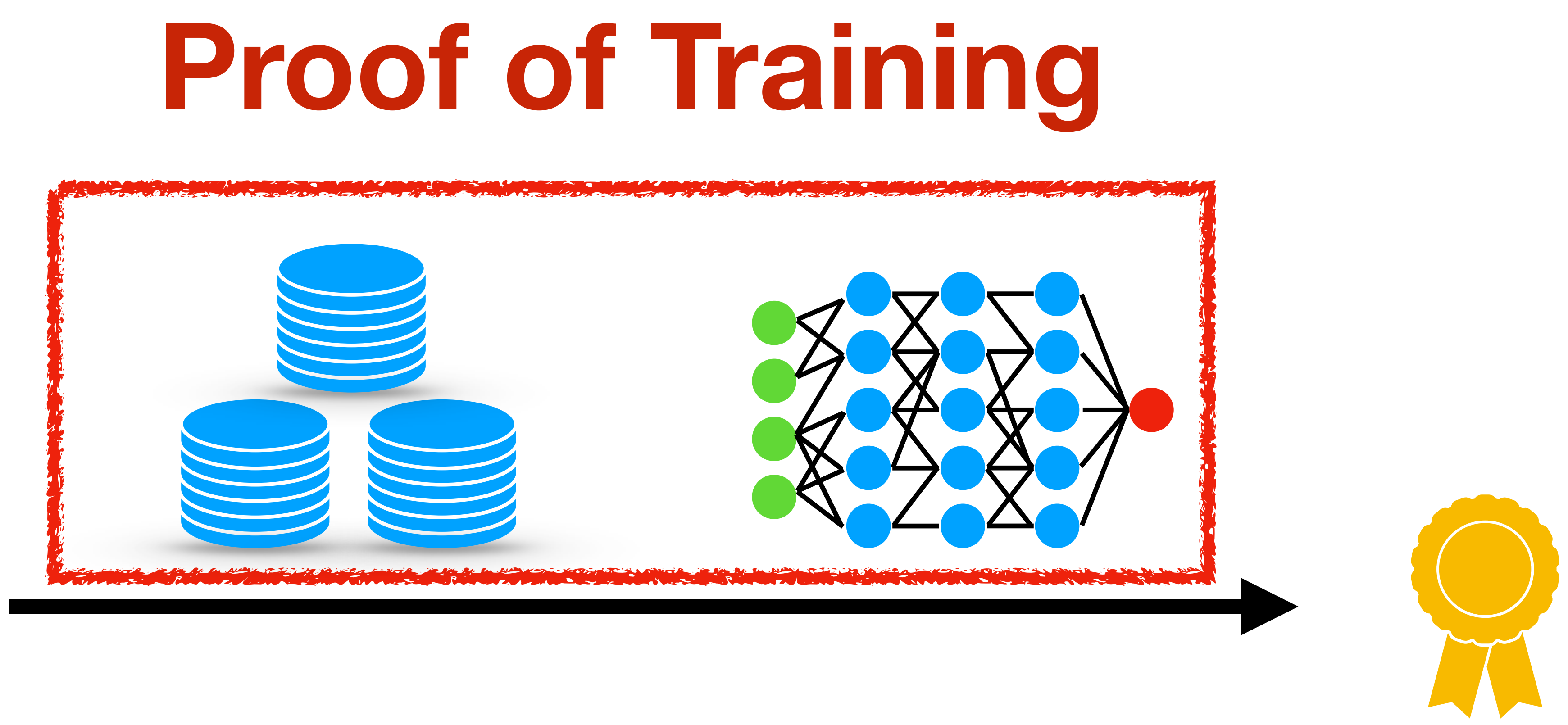
AI + Regulation + ZKPs





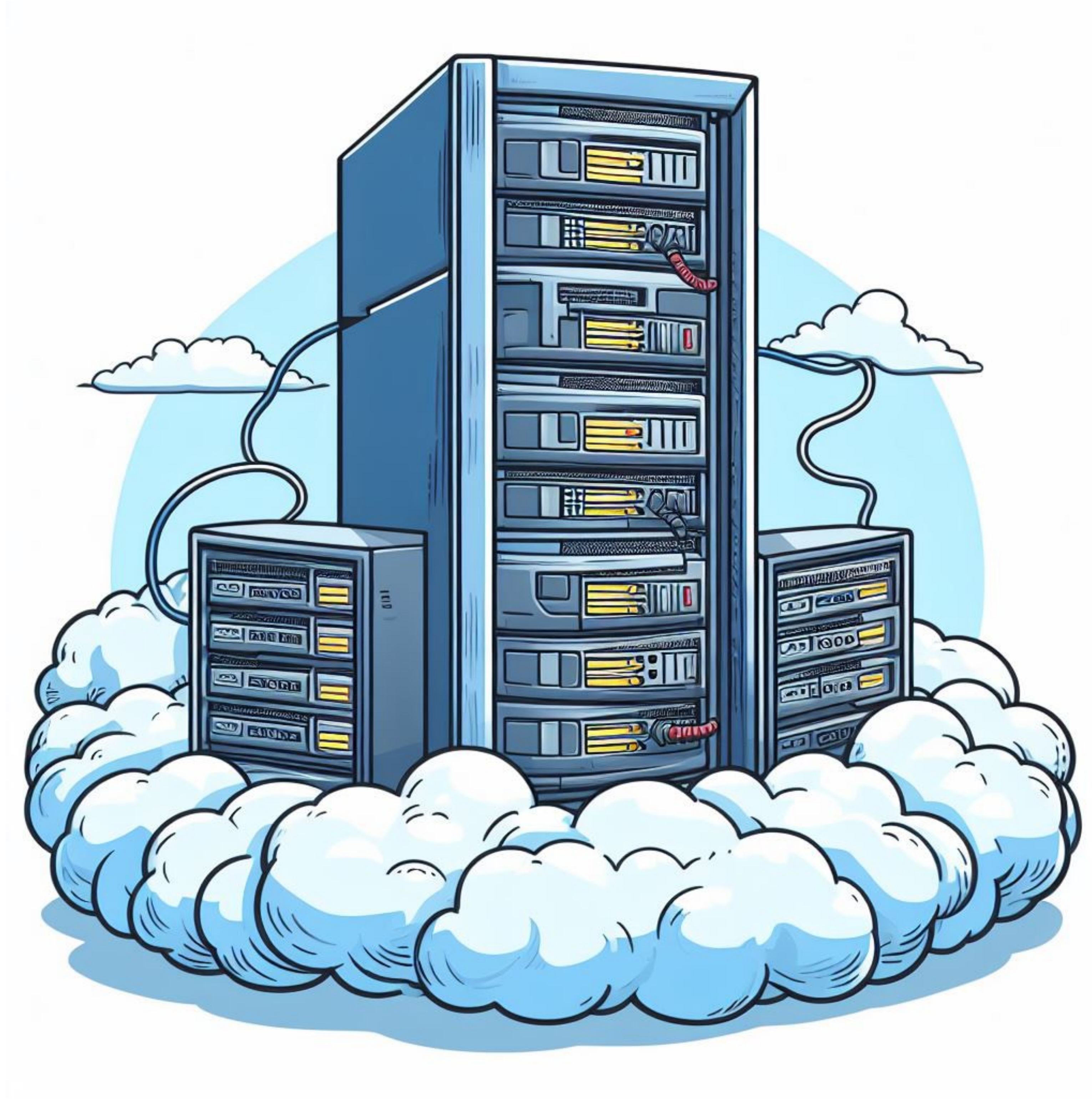
AI + Regulation + ZKPs



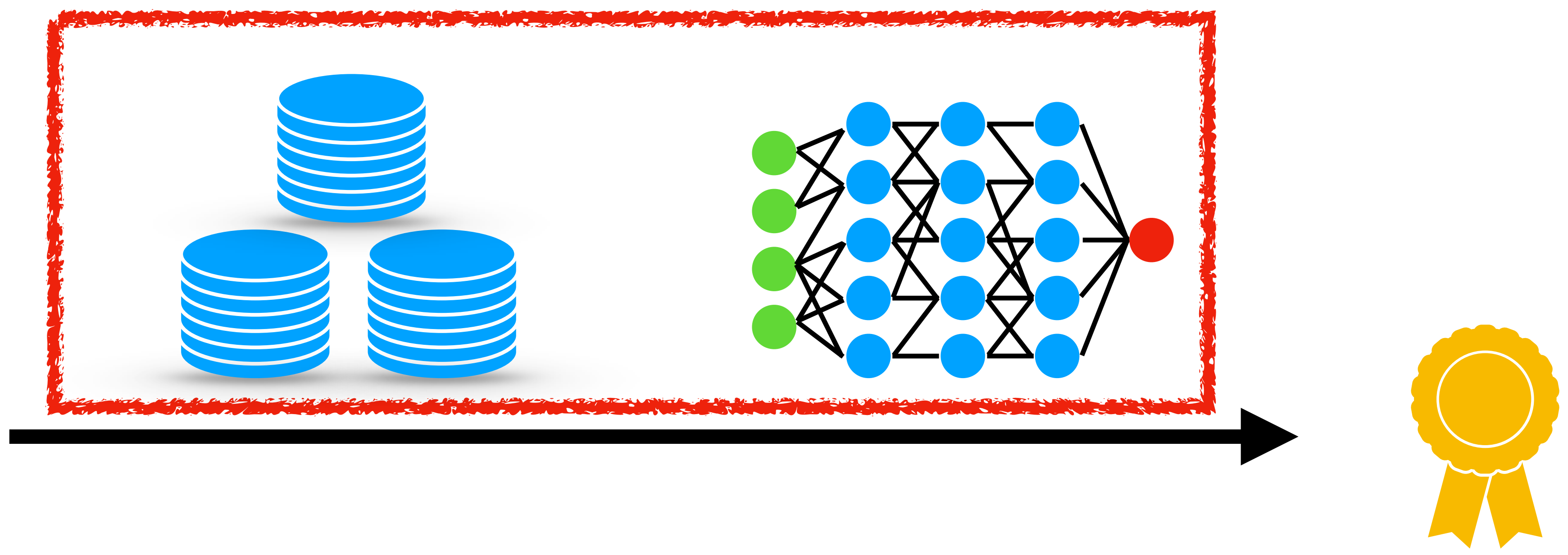


- Prover knows some training data and training results in some model

AI + Regulation + ZKPs

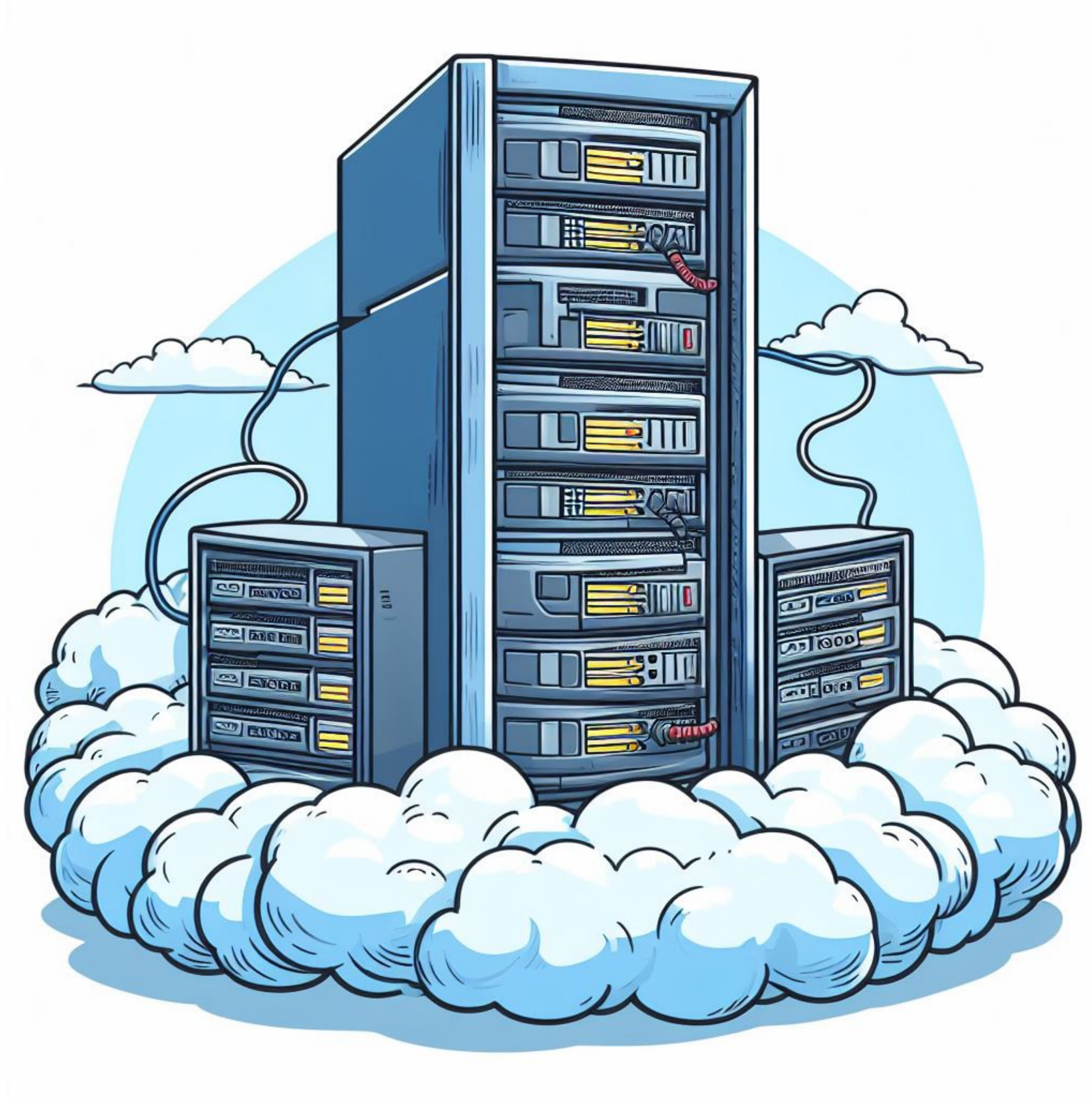


Proof of Training

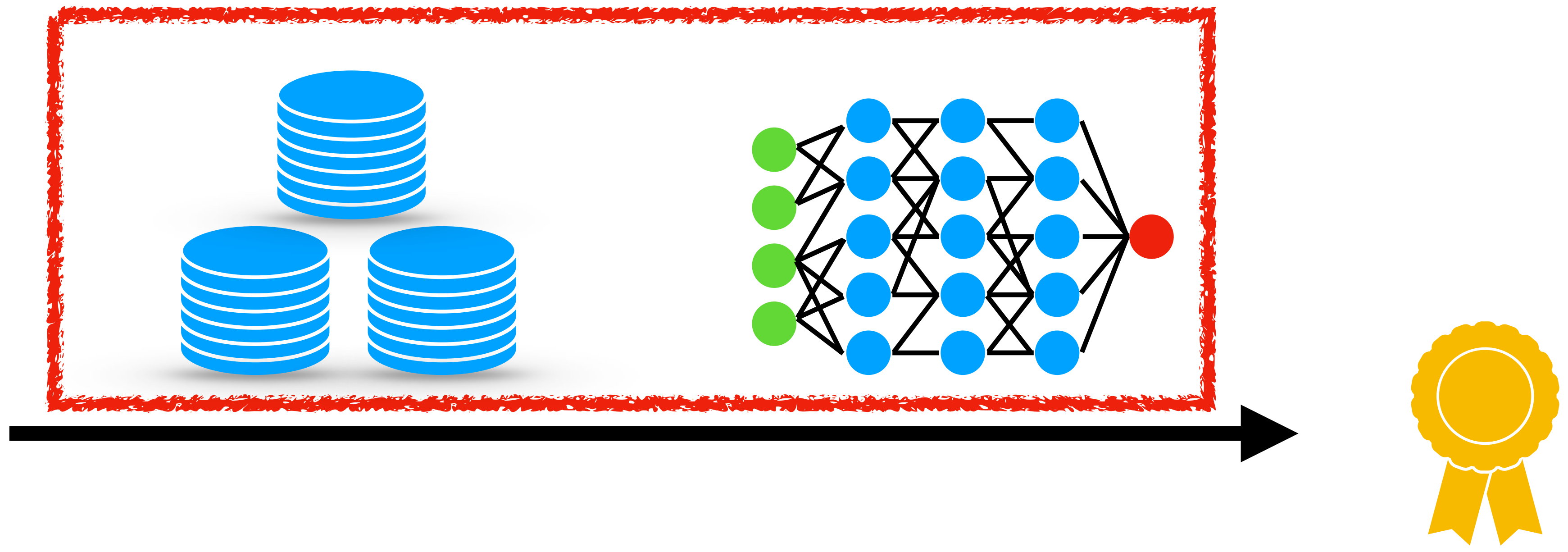


- Prover knows some training data and training results in some model
- Training data satisfies desired statistical properties

AI + Regulation + ZKPs

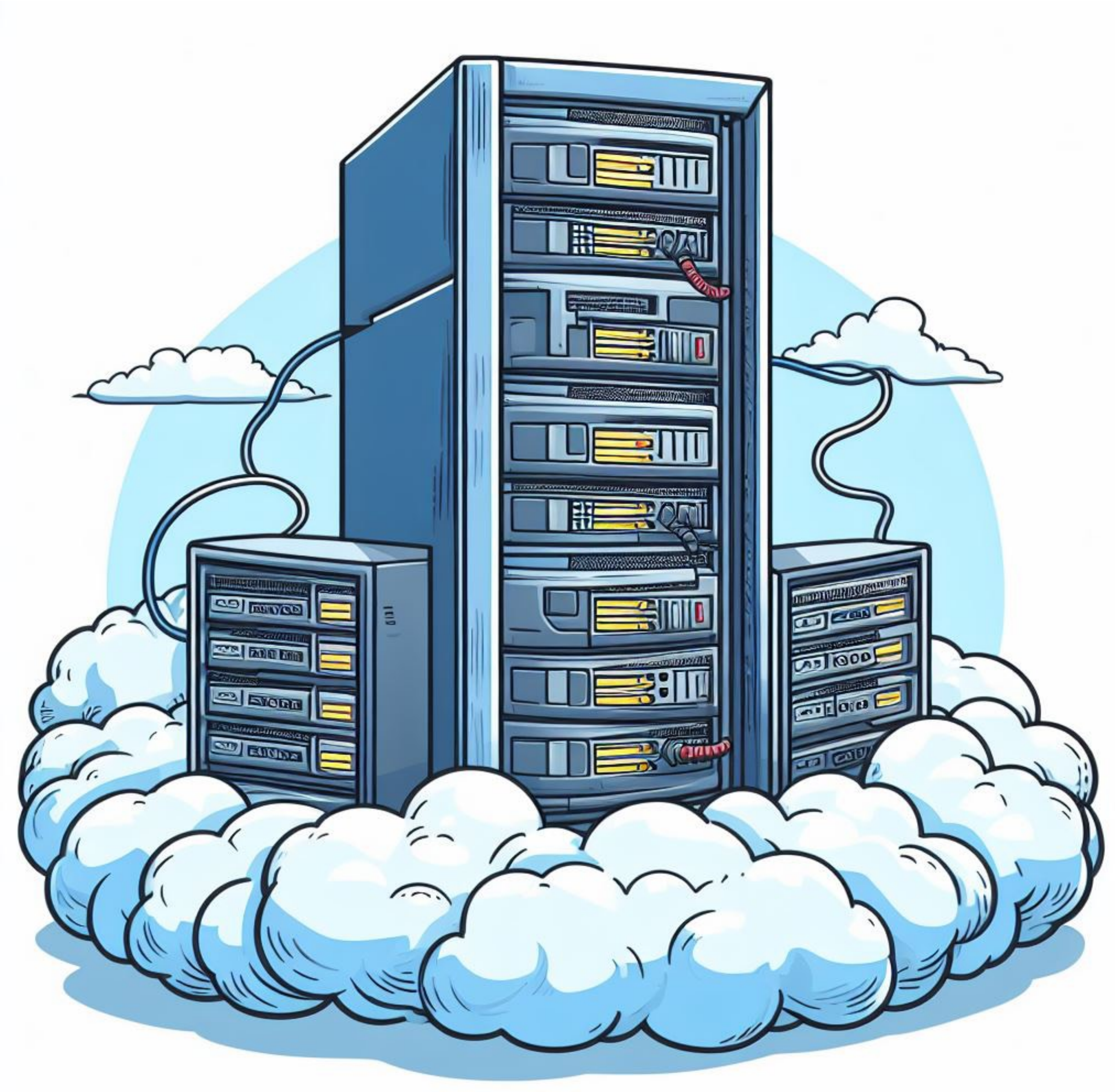


Proof of Training

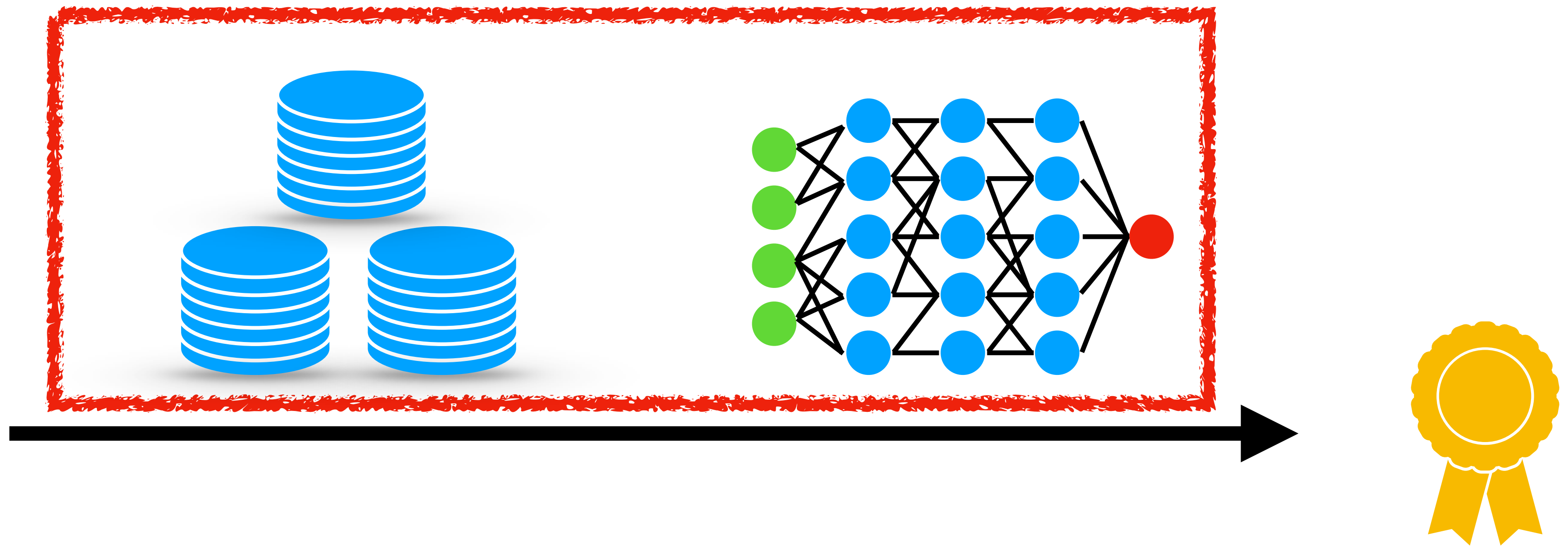


- Prover knows some training data and training results in some model
- Training data satisfies desired statistical properties
- + any other guarantees. e.g. copyright secured

AI + Regulation + ZKPs

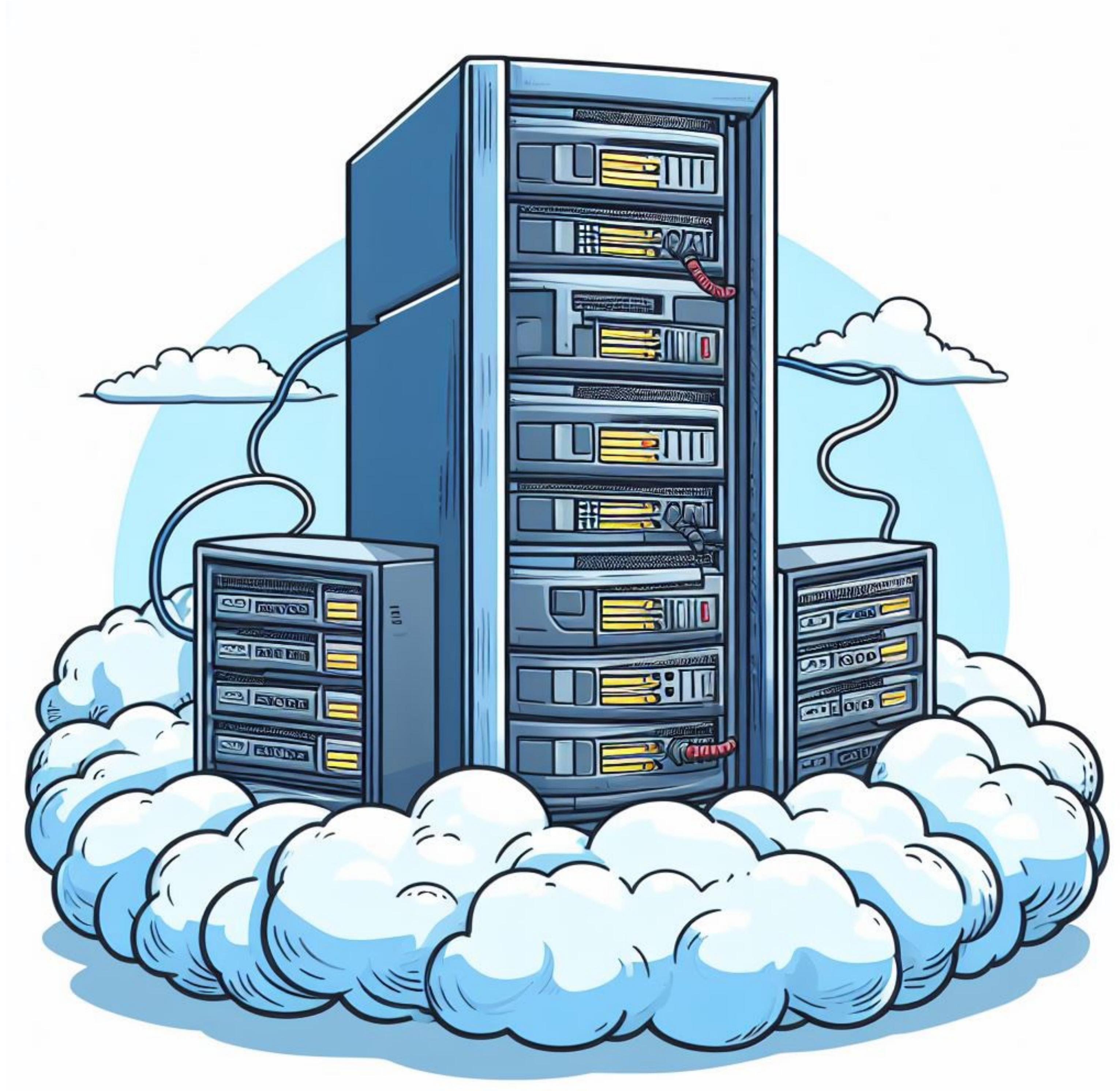


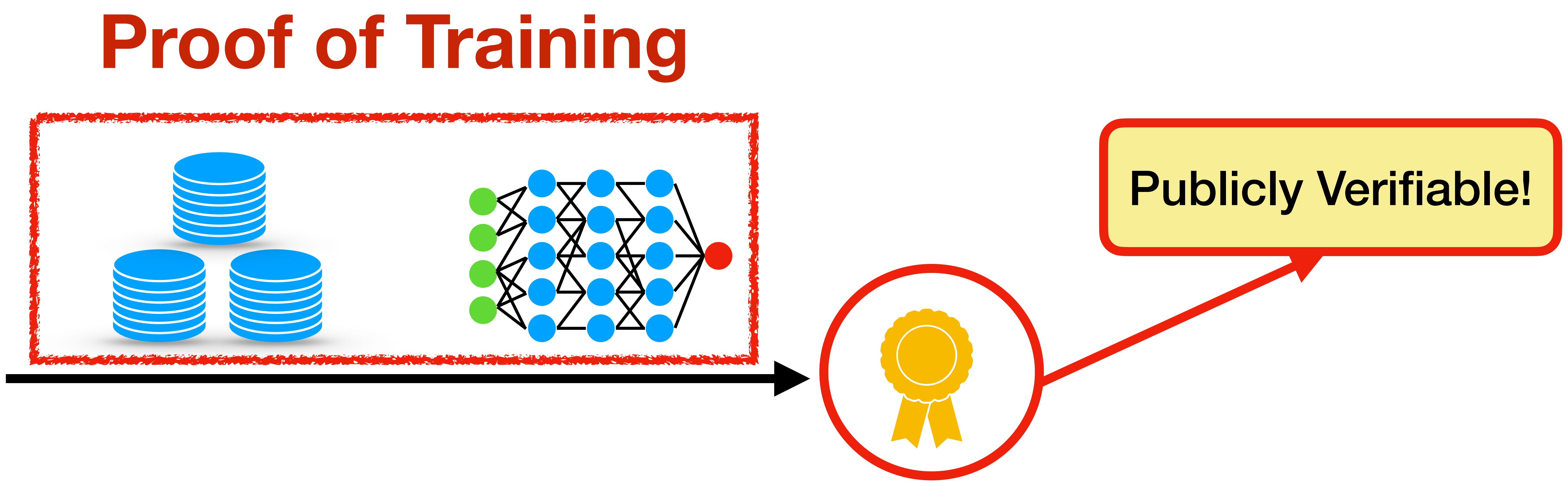
Proof of Training



- Prover knows some training data and training results in some model
- Training data satisfies desired statistical properties
- + any other guarantees. e.g. copyright secured
- ZK ⇒ No information about model/data leaked

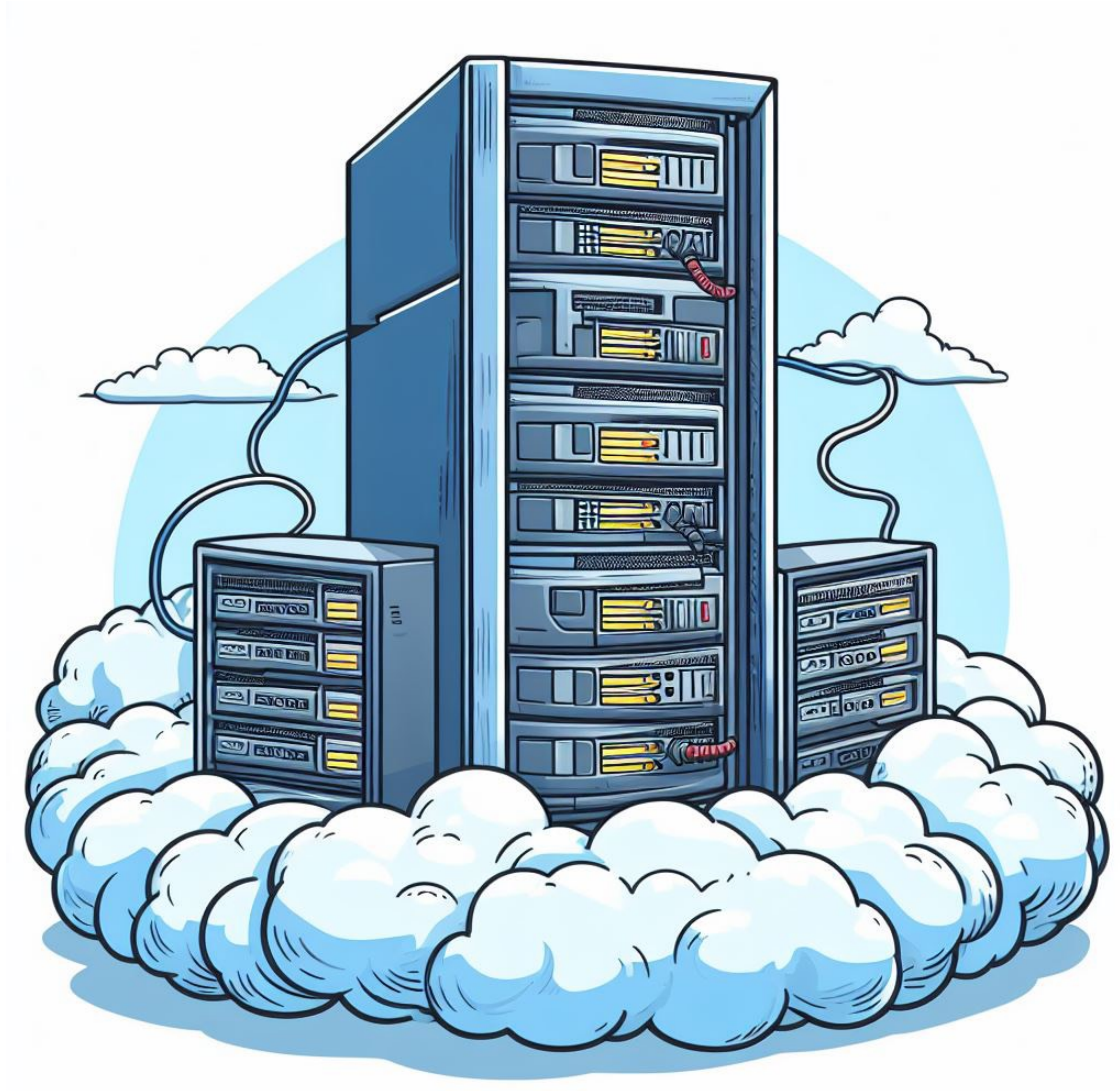
AI + Regulation + ZKPs



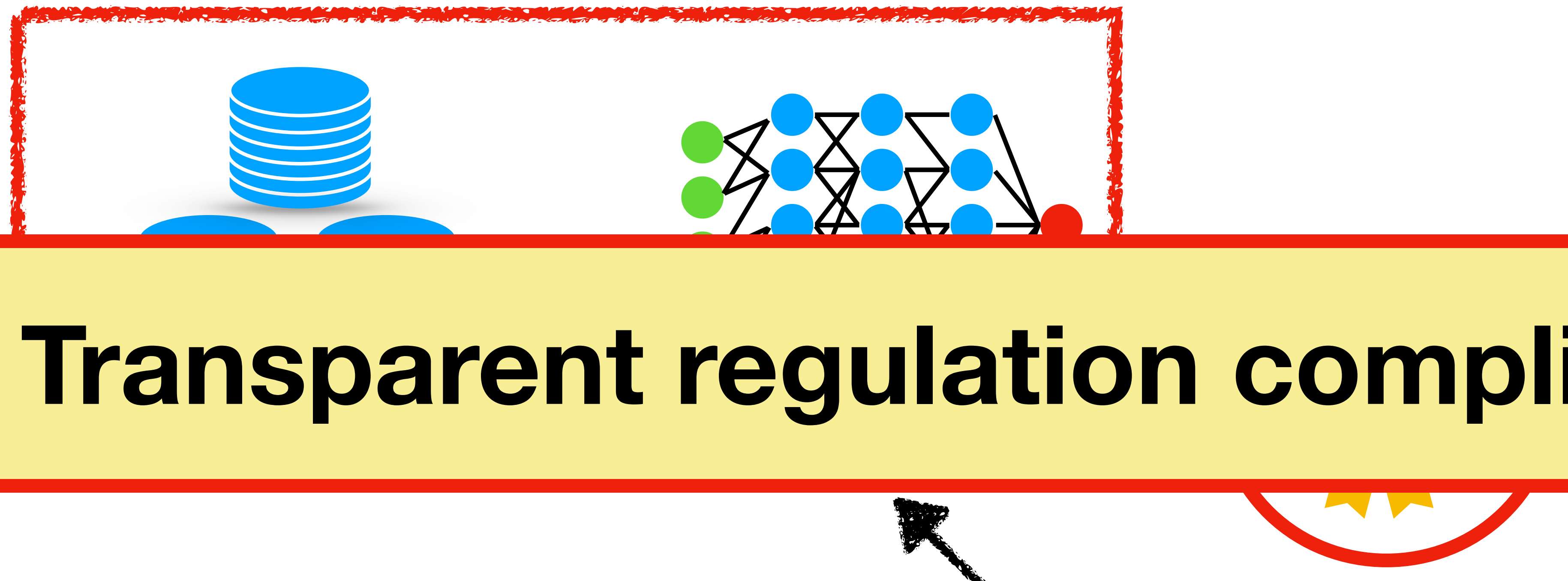


- Prover knows some training data and training results in some model
- Training data satisfies desired statistical properties
- + any other guarantees. e.g. copyright secured
- ZK ⇒ No information about model/data leaked

AI + Regulation + ZKPs



Proof of Training

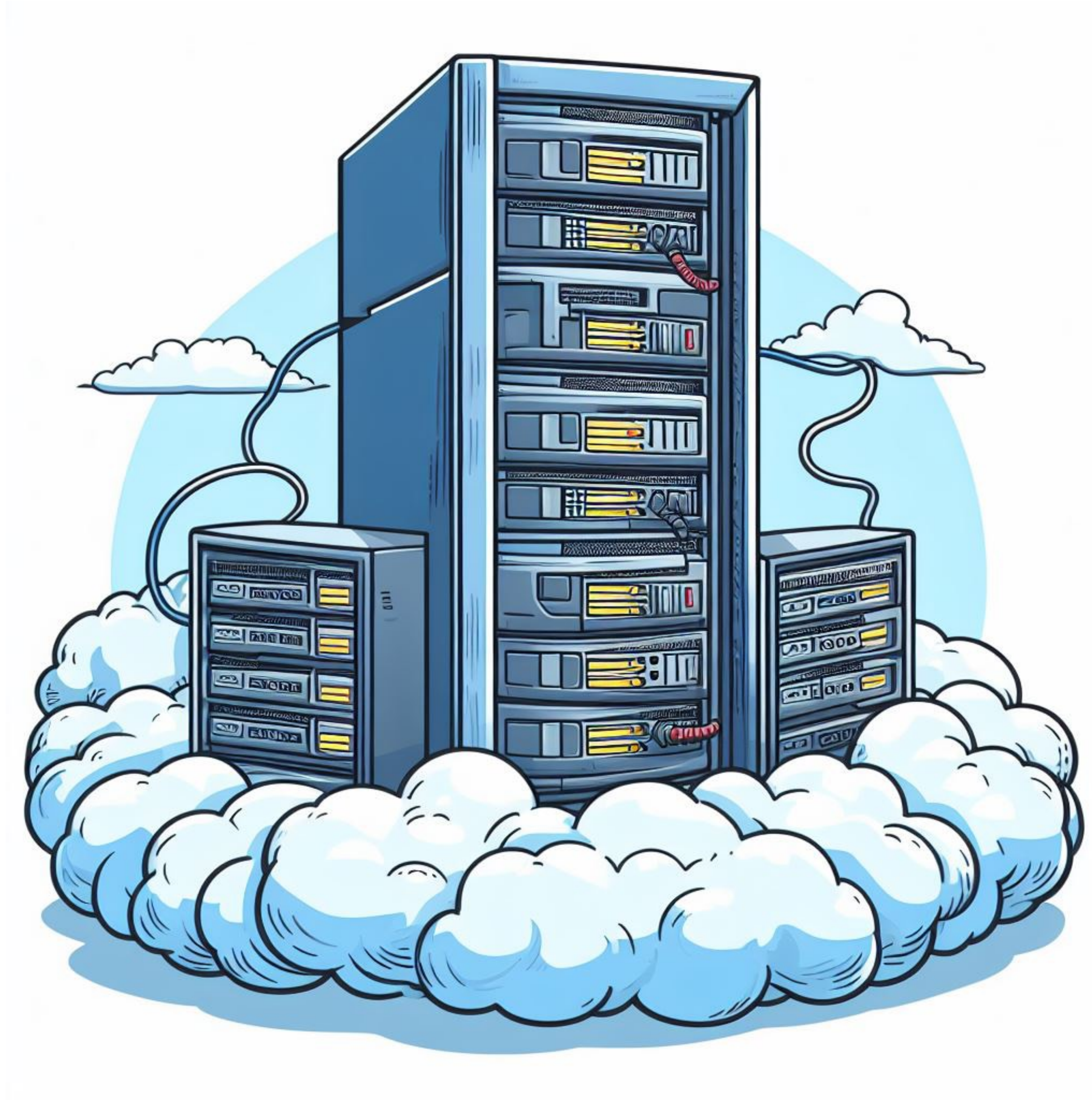


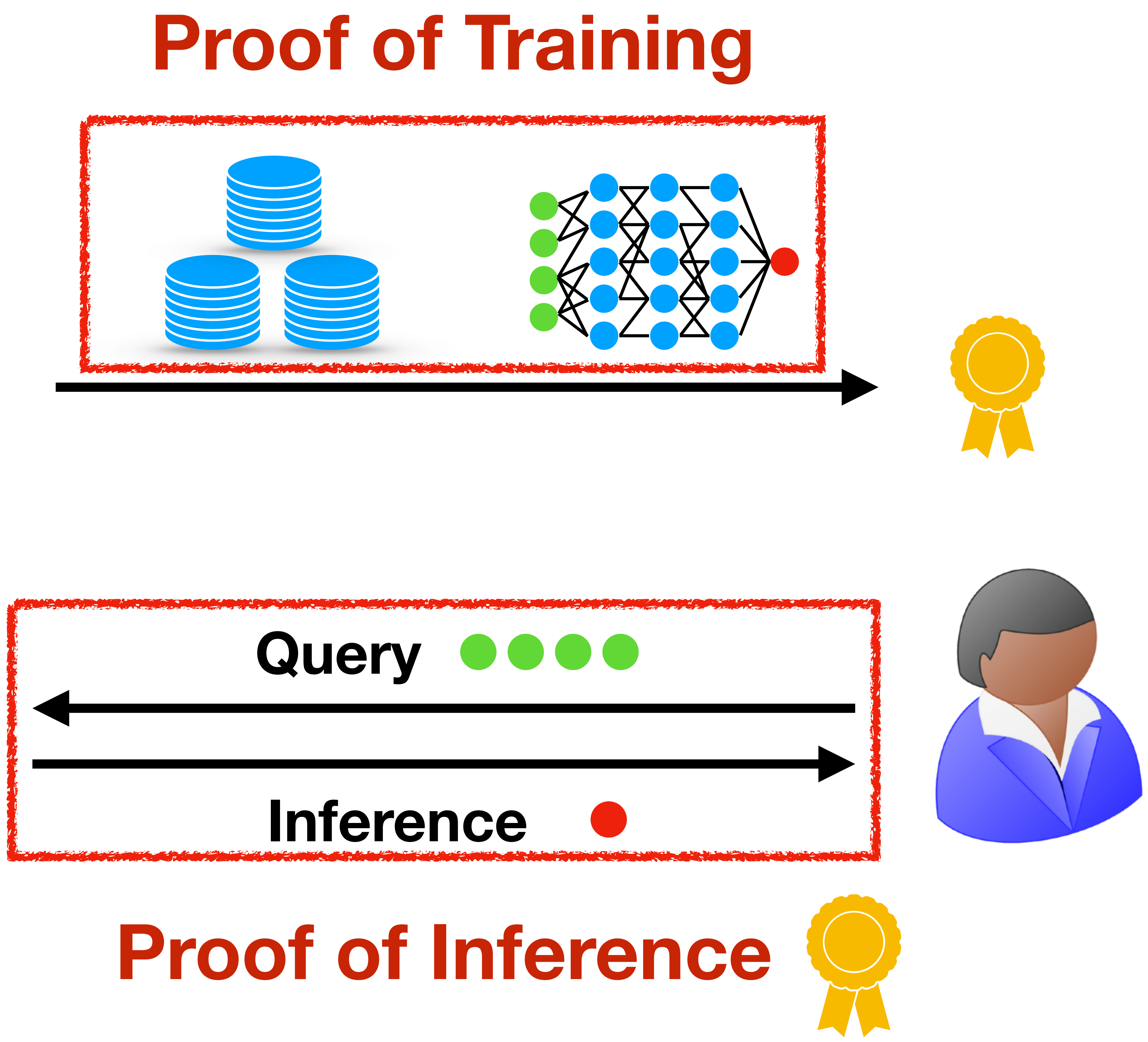
Publicly Verifiable!

Transparent regulation compliance!

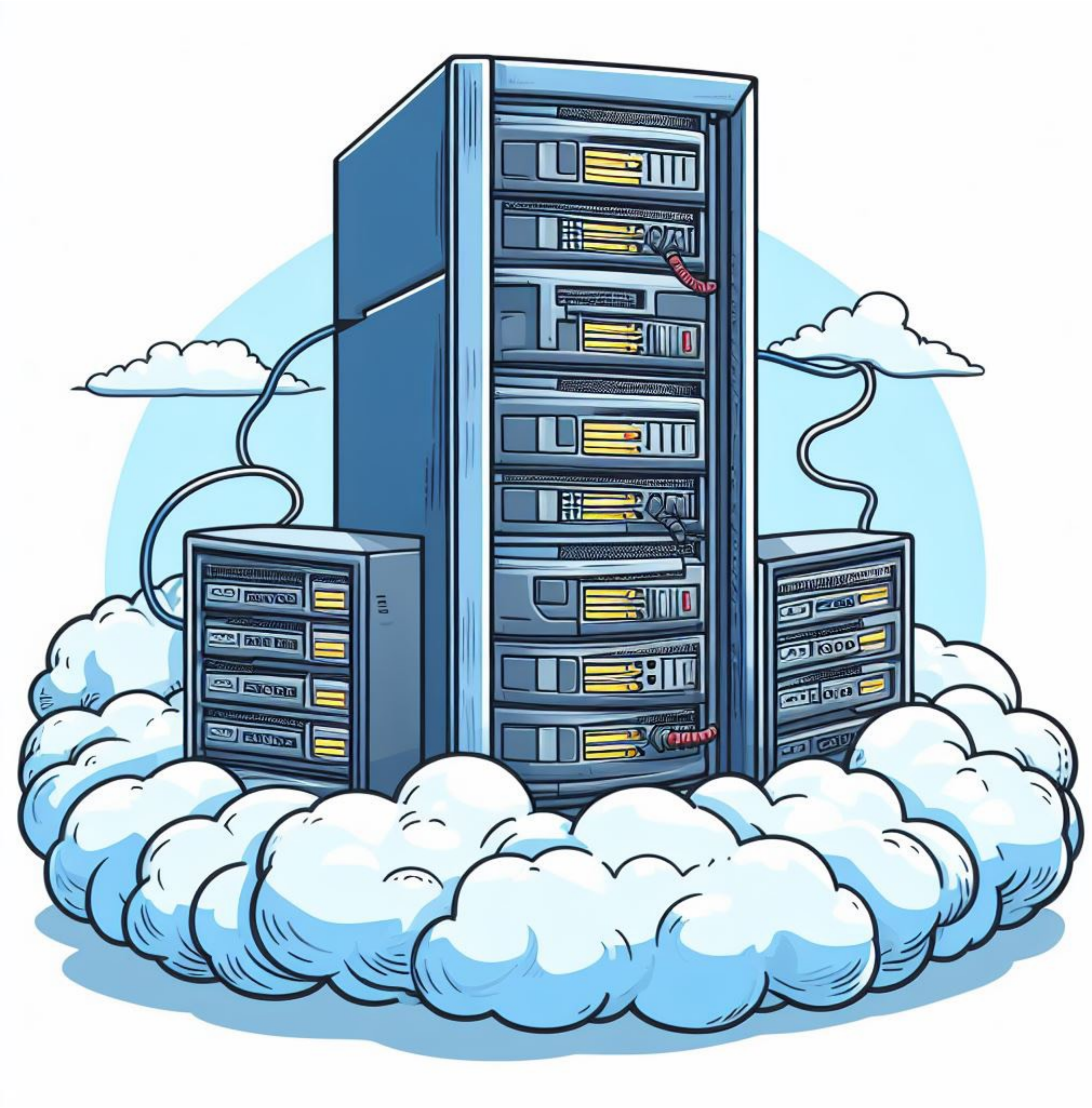
- Prover knows some training data and training results in some model
- Training data satisfies desired statistical properties
- + any other guarantees. e.g. copyright secured
- ZK ⇒ No information about model/data leaked

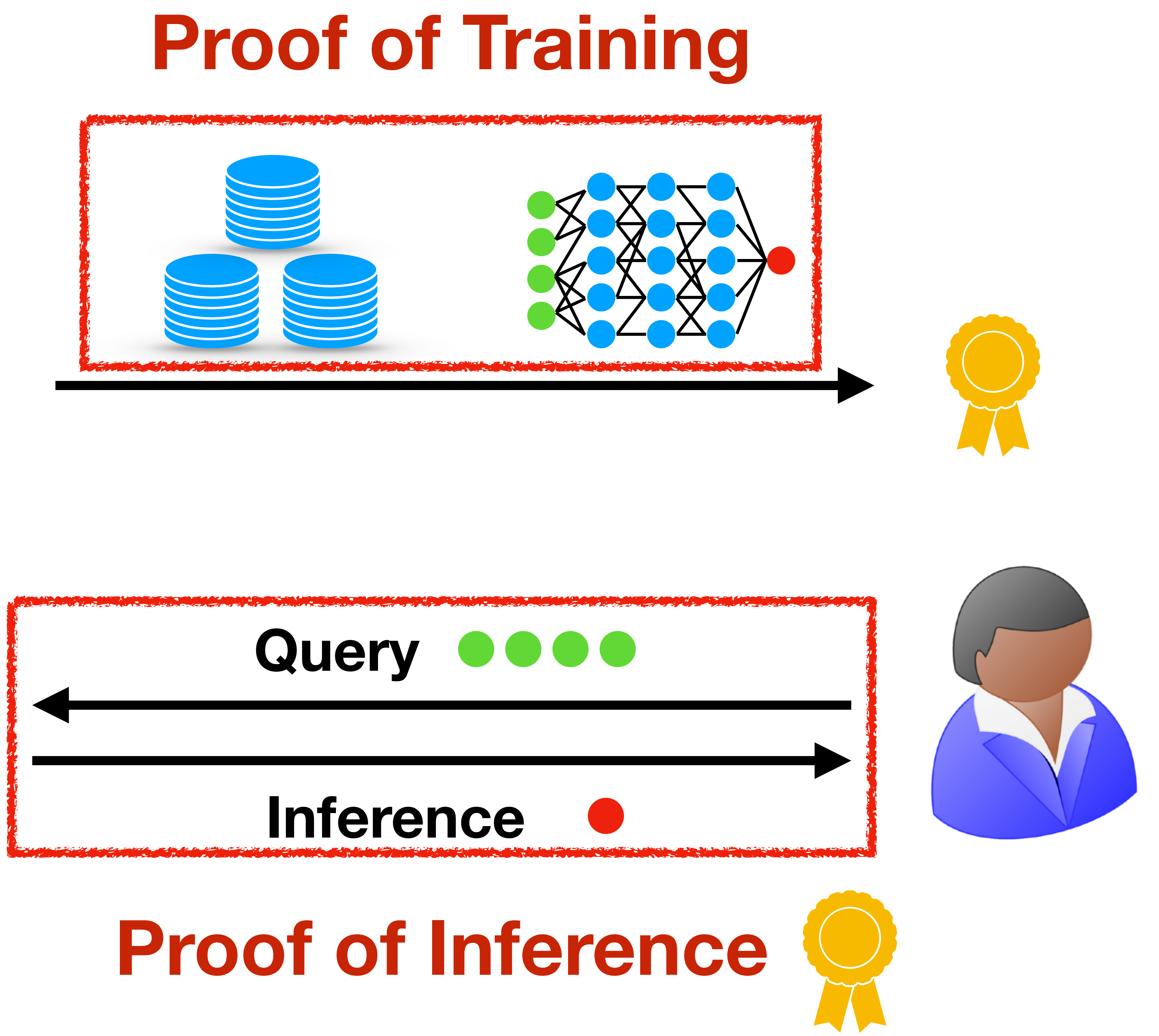
AI + Regulation + ZKPs





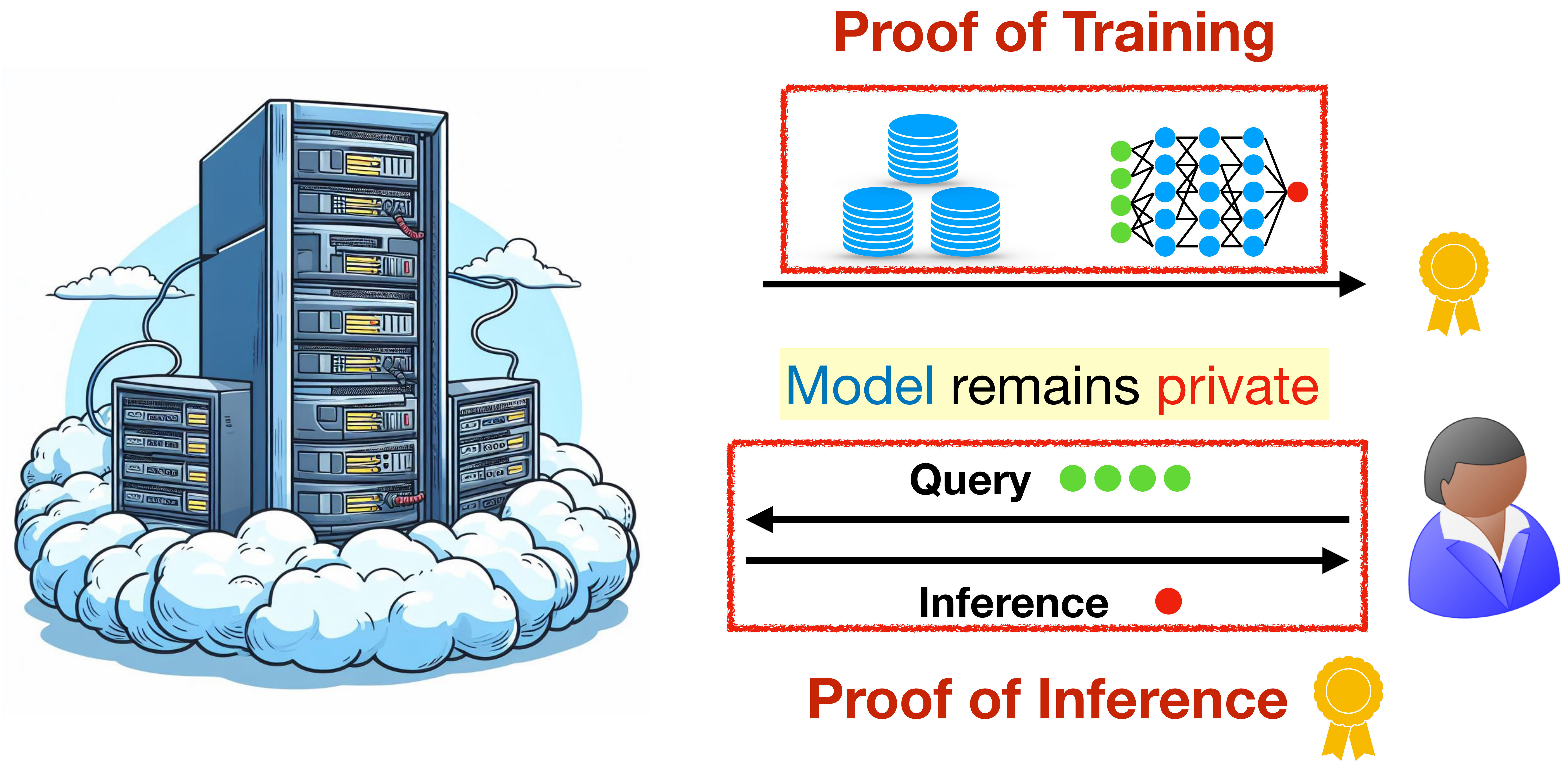
AI + Regulation + ZKPs





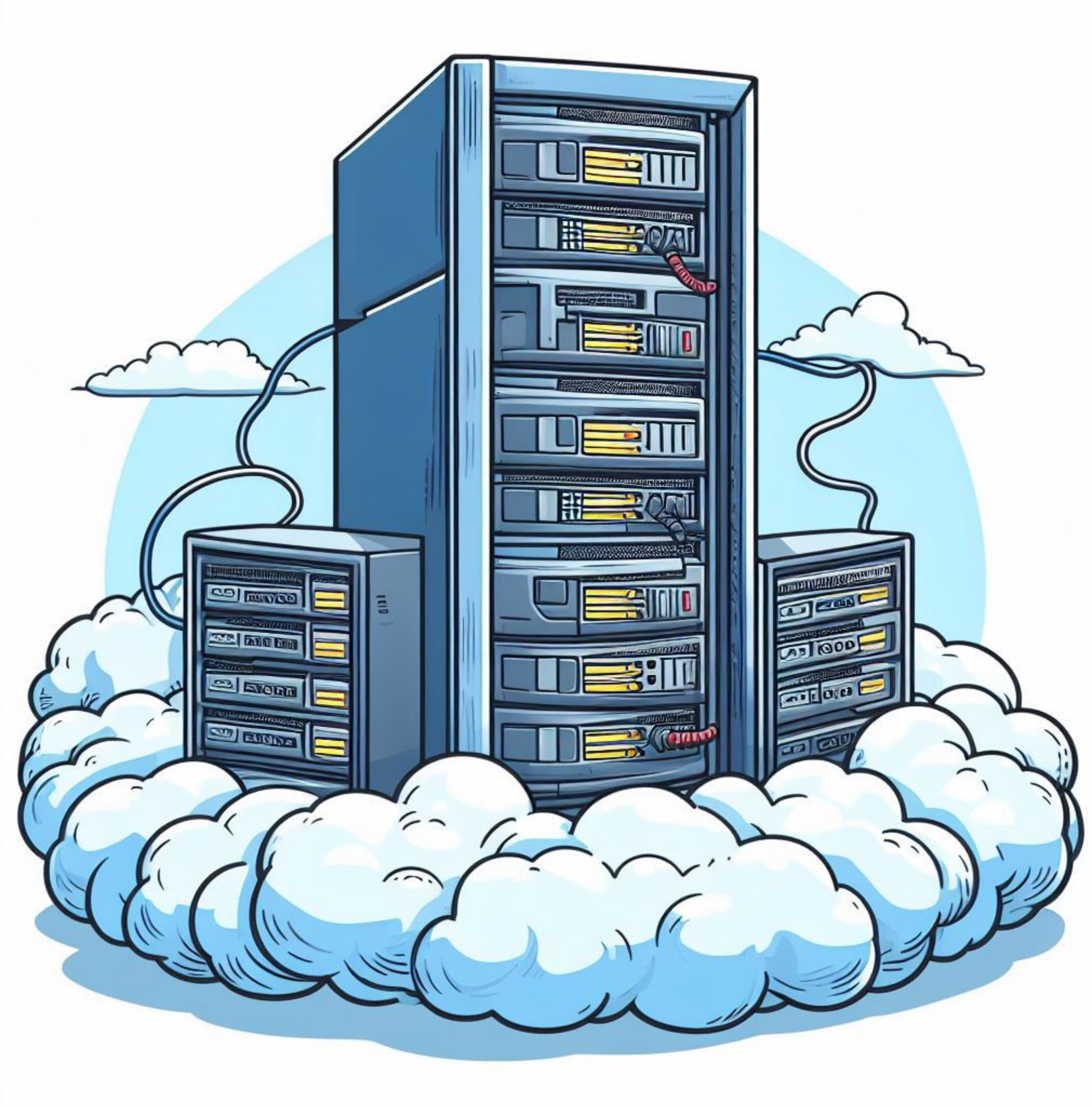
Guarantees same model is used for inference

AI + Regulation + ZKPs

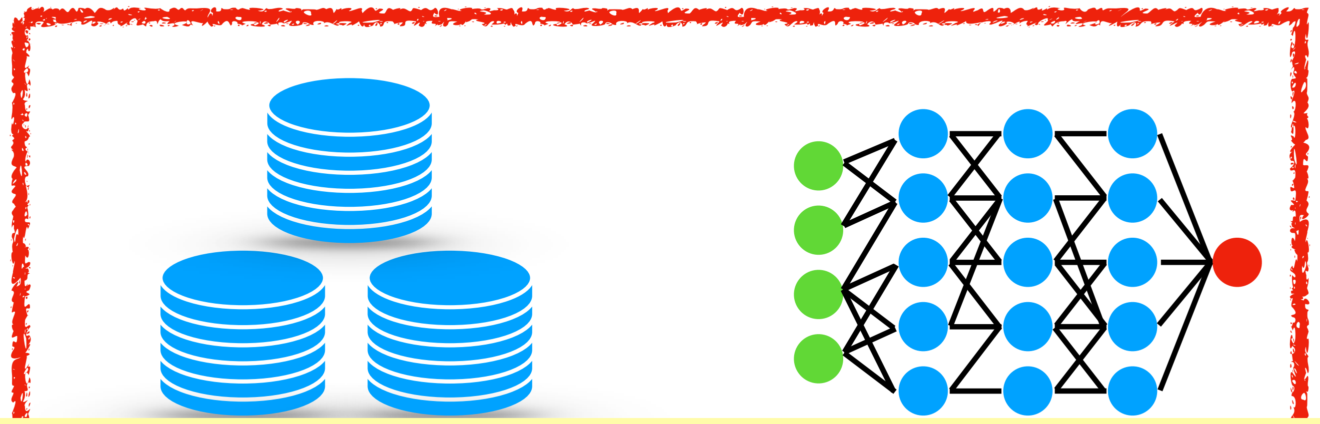


Guarantees same model is used for inference

AI + Regulation + ZKPs



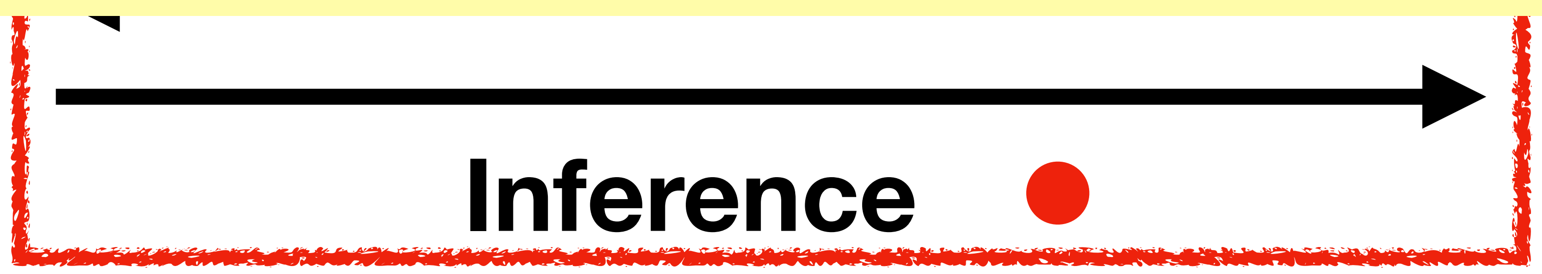
Proof of Training



Ties together

Data ⟺ Model ⟺ Inference

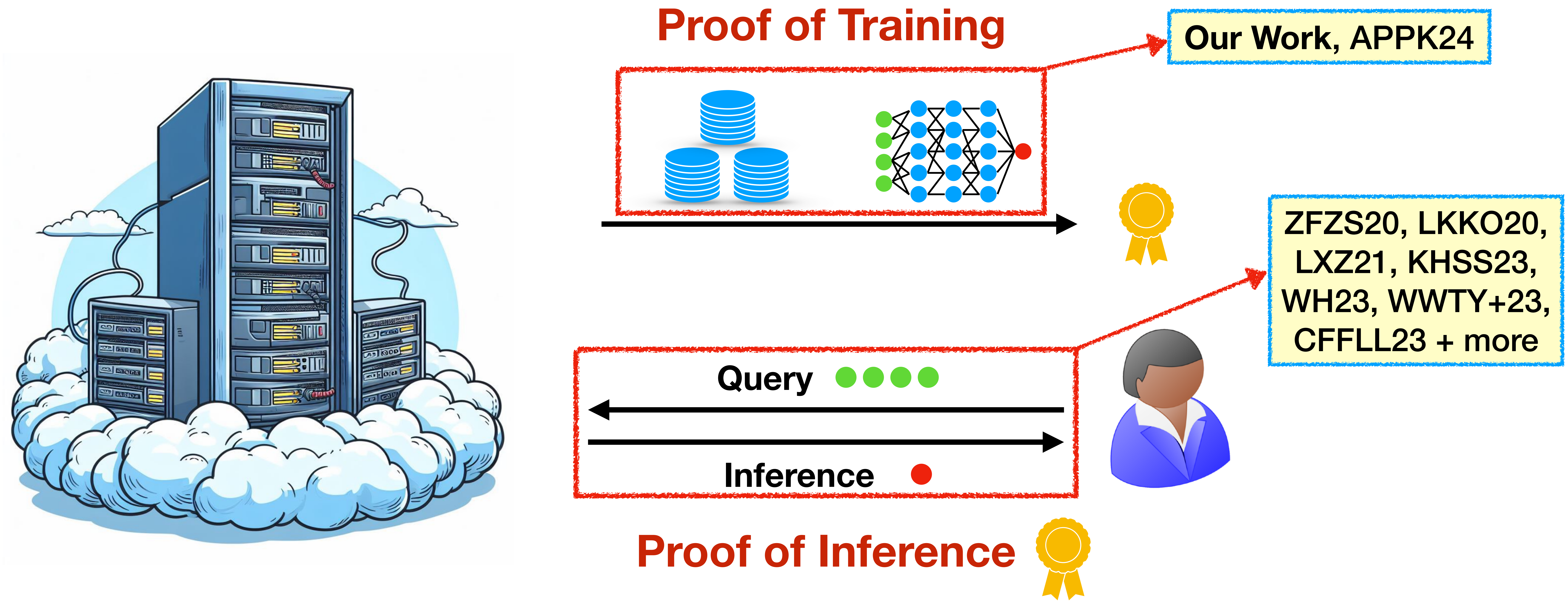
(Not really possible w/o crypto)



Proof of Inference

Guarantees same model is used for inference

AI + Regulation + ZKPs



Brief Overview of Our Work

Some Considerations

Some Considerations

1. Training is already very expensive

Some Considerations

1. Training is already very expensive
   - Minimize prover overhead and need to be streaming friendly (massive circuits)

Some Considerations

1. Training is already very expensive
   - Minimize prover overhead and need to be streaming friendly (massive circuits)
2. Verification of training happens only once per model

Some Considerations

1. Training is already very expensive
    - Minimize prover overhead and need to be streaming friendly (massive circuits)
2. Verification of training happens only once per model
    - Can increase verifier work/proof size to get a faster prover

Some Considerations

1. Training is already very expensive
   - Minimize prover overhead and need to be streaming friendly (massive circuits)
2. Verification of training happens only once per model
   - Can increase verifier work/proof size to get a faster prover
3. Machine Learning involves 32/64-bit floating point operations

Some Considerations

1. Training is already very expensive
   - Minimize prover overhead and need to be streaming friendly (massive circuits)
2. Verification of training happens only once per model
   - Can increase verifier work/proof size to get a faster prover
3. Machine Learning involves 32/64-bit floating point operations
   - Avoid very large fields — unnecessary overhead

Some Considerations

1. Training is already very expensive
   - Minimize prover overhead and need to be streaming friendly (massive circuits)
2. Verification of training happens only once per model
   - Can increase verifier work/proof size to get a faster prover
3. Machine Learning involves 32/64-bit floating point operations
   - Avoid very large fields — unnecessary overhead
   - Need to handle floating point algebra

Off the shelf ZKPs insufficient

Two *okay* candidates:	zkSNARKs [BCCT12, Groth16, Plonk…]	MPC-in-the-Head [IKOS07 …]

*table has changed now!

Off the shelf ZKPs insufficient

Two *okay* candidates:	zkSNARKs [BCCT12, Groth16, Plonk…]	MPC-in-the-Head [IKOS07 …]
Proof Size	Small	Large

*table has changed now!

Off the shelf ZKPs insufficient

Two *okay* candidates:	zkSNARKs [BCCT12, Groth16, Plonk…]	MPC-in-the-Head [IKOS07 …]
Proof Size	Small	Large
Verification	Fast	Slow

*table has changed now!

Off the shelf ZKPs insufficient

Two *okay* candidates:	zkSNARKs [BCCT12, Groth16, Plonk…]	MPC-in-the-Head [IKOS07 …]
Proof Size	Small	Large
Verification	Fast	Slow
Proof Generation	Slow	Fast

*table has changed now!

Off the shelf ZKPs insufficient

Two *okay* candidates:	zkSNARKs [BCCT12, Groth16, Plonk…]	MPC-in-the-Head [IKOS07 …]
Proof Size	Small	Large
Verification	Fast	Slow
Proof Generation	Slow	Fast
Streaming Friendly	No*	Yes

***table has changed now!**

Off the shelf ZKPs insufficient

Two *okay* candidates:	zkSNARKs [BCCT12, Groth16, Plonk…]	MPC-in-the-Head [IKOS07 …]
Proof Size	Small	Large
Verification	Fast	Slow
Proof Generation	Slow	Fast
Streaming Friendly	No*	Yes
Small Field Support	No*	Yes

*table has changed now!

Our approach: Best of both worlds

	zkSNARKs + MPC-in-the-Head
Proof Size	Medium
Verification	Medium
Proof Generation	Medium
Streaming Friendly	Yes
Small Field Support	Yes

Our approach: Best of both worlds

	zkSNARKs + MPC-in-the-Head
Proof Size	Medium
Verification	Medium
Proof Generation	Medium
Streaming Friendly	Yes
Small Field Support	Yes

Independent of number of features

Our approach: Best of both worlds

	zkSNARKs + MPC-in-the-Head
Proof Size	Medium
Verification	Medium
Proof Generation	Medium
Streaming Friendly	Yes
Small Field Support	Yes

Independent of number of features

Medium → Offline Phase (Before data known) + Online Phase (Data dependent)

Evaluation

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features
- With caveats: floating → fixed point, approximate activation function

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features
- With caveats: floating → fixed point, approximate activation function
- Proof Size: < 10% of dataset size

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features
- With caveats: floating → fixed point, approximate activation function
- Proof Size: < 10% of dataset size

Better than sending all the data and privacy!

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features
- With caveats: floating → fixed point, approximate activation function
- Proof Size: < 10% of dataset size

Better than sending all the data and privacy!

- Online Phase (single thread):

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features
- With caveats: floating → fixed point, approximate activation function
- Proof Size: < 10% of dataset size
- Online Phase (single thread):
  - Prover ~ 1 hour

Better than sending all the data and privacy!

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features
- With caveats: floating → fixed point, approximate activation function
- Proof Size: < 10% of dataset size

Better than sending all the data and privacy!

- Online Phase (single thread):
  - Prover ~ 1 hour
  - Verifier ~ few minutes

Evaluation

- Logistic Regression on a realistic dataset — 250K records, 1024 features
- With caveats: floating → fixed point, approximate activation function
- Proof Size: < 10% of dataset size

Better than sending all the data and privacy!

- Online Phase (single thread):
    - Prover ~ 1 hour
    - Verifier ~ few minutes
    - Training ~ 2-3 seconds

Other Applications

- Proof of Training for fine-tuning foundational models
- Also solves open problems in other papers:
    - [DDKYSA23] — “Data Property Attestation”
    - [JBVGSTD23] — “Tying models to the dataset”

DDKYSA23: https://arxiv.org/abs/2308.09552
JBVGSTD23: https://arxiv.org/abs/2303.07476

Key Takeaways

Key Takeaways

- Incoming AI regulation can benefit greatly from ZKPs
  - Proofs of Training and Inferences are core building blocks

Key Takeaways

- Incoming AI regulation can benefit greatly from ZKPs
  - Proofs of Training and Inferences are core building blocks
- Make the job of Regulators easier!
  - Easy to use and deploy > fastest scheme.
  - Try to build on top of popular tooling. Better community adoption.

Key Takeaways

- Incoming AI regulation can benefit greatly from ZKPs
  - Proofs of Training and Inferences are core building blocks
- Make the job of Regulators easier!
  - Easy to use and deploy > fastest scheme.
  - Try to build on top of popular tooling. Better community adoption.
- ZKPs for ML Training can be practical!

Thank you!

Paper: ia.cr/2023/1345

Code: https://github.com/guruvamsi-policharla/zkpot

Blogpost: