Patterns in Data Provenance

Cryptographic Magic Included

Benedict Lau // March 23, 2024 // Causal Islands LA 2024

Benedict Lau

Data Provenance
@ Hypha Worker Co-op

HYPHA

Data Integrity

Decentralized Preservation

Verifiable Computing

AI Model Lineage

Hardware Attestations

Media Authentication

Benedict Lau

Data Provenance
@ Hypha Worker Co-op

HYPHA

I want to ensure my data, and its associated metadata, is not tampered.

Data

Data

Metadata

author	Inge Wallumrød
license	CC-BY-SA 4.0

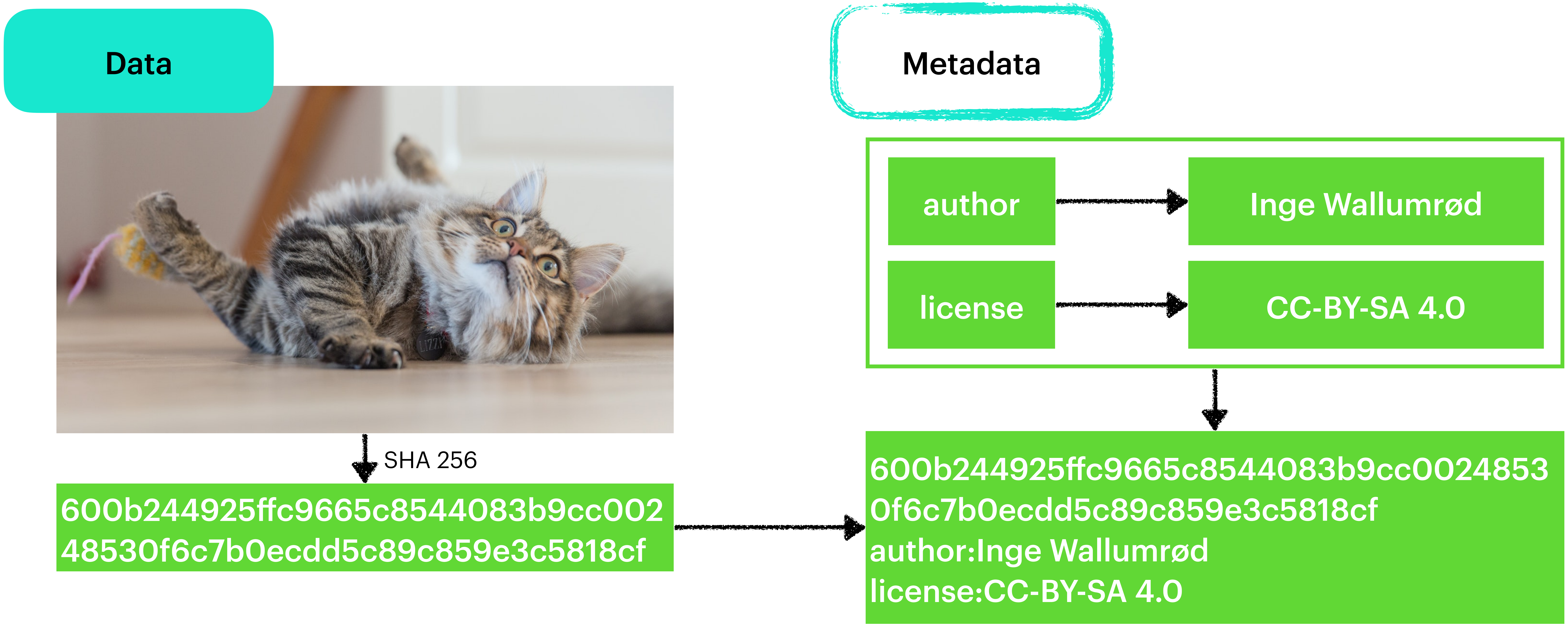
Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author	→	Inge Wallumrød
license	→	CC-BY-SA 4.0



Data

SHA 256

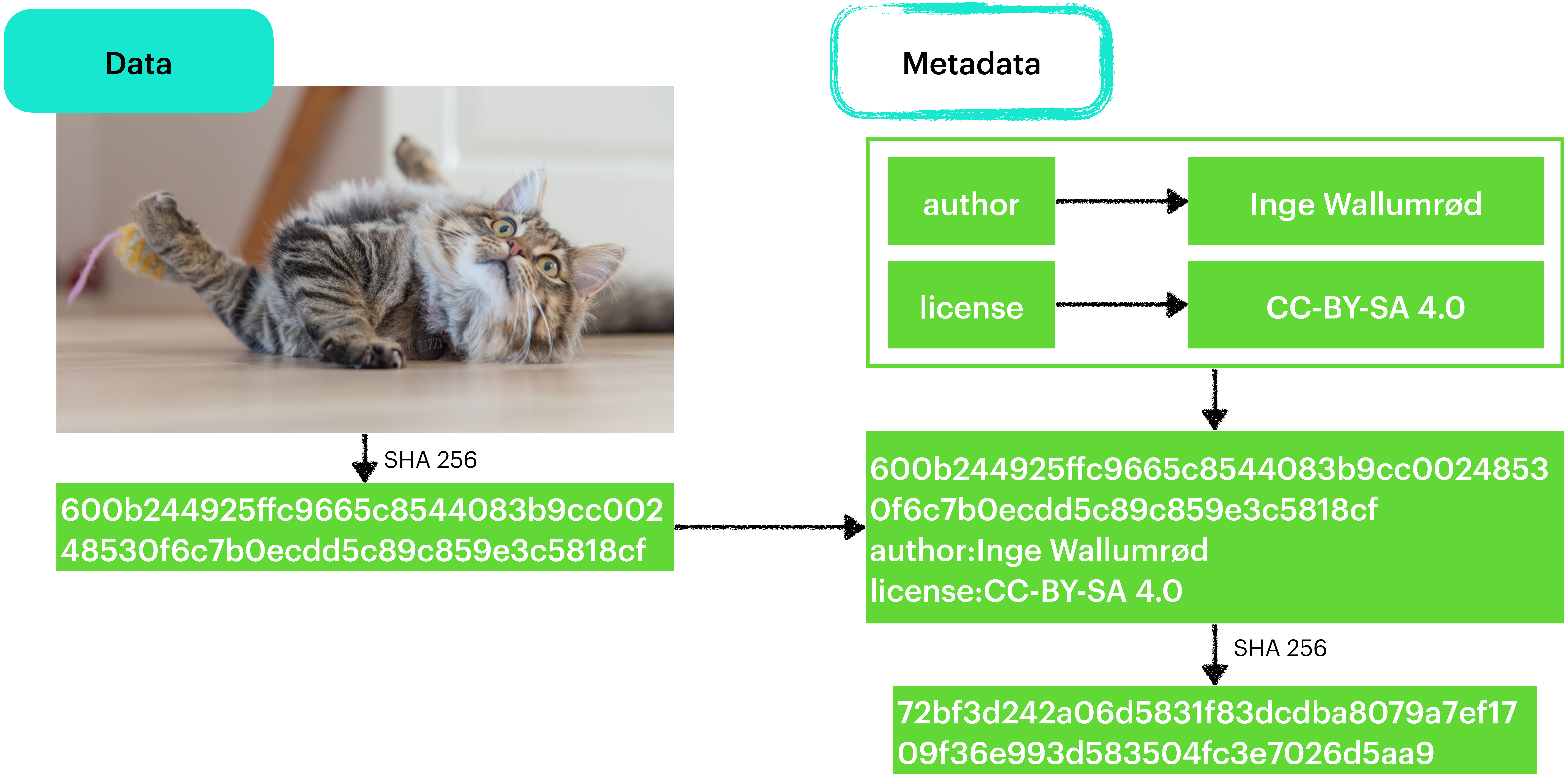
600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author → Inge Wallumrød

license → CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0



Data

Metadata

author → Inge Wallumrød

license → CC-BY-SA 4.0

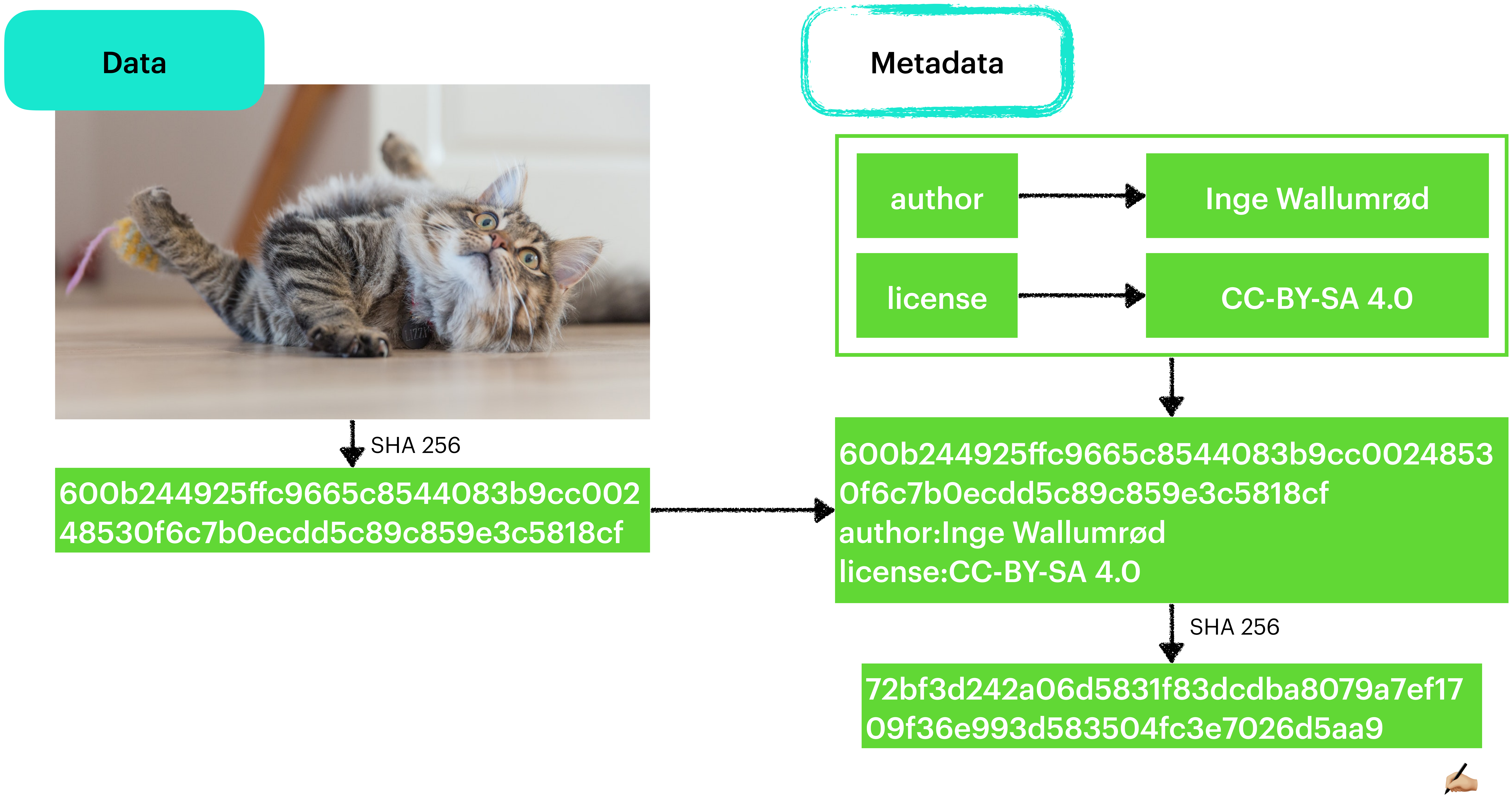
SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

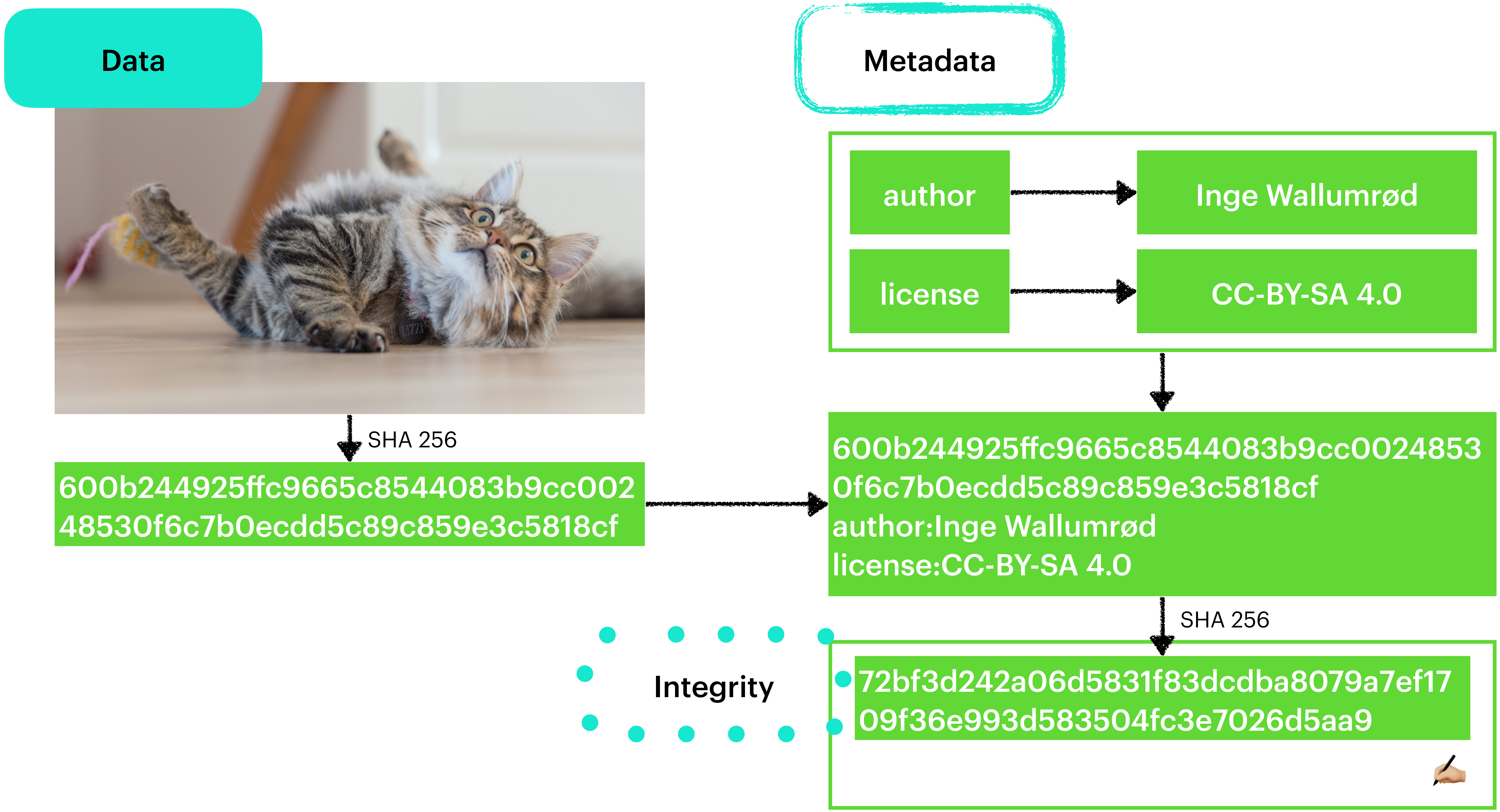
author	Inge Wallumrød
license	CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

✍



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author	Inge Wallumrød
license	CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

✍

Integrity

Data

Metadata

Integrity

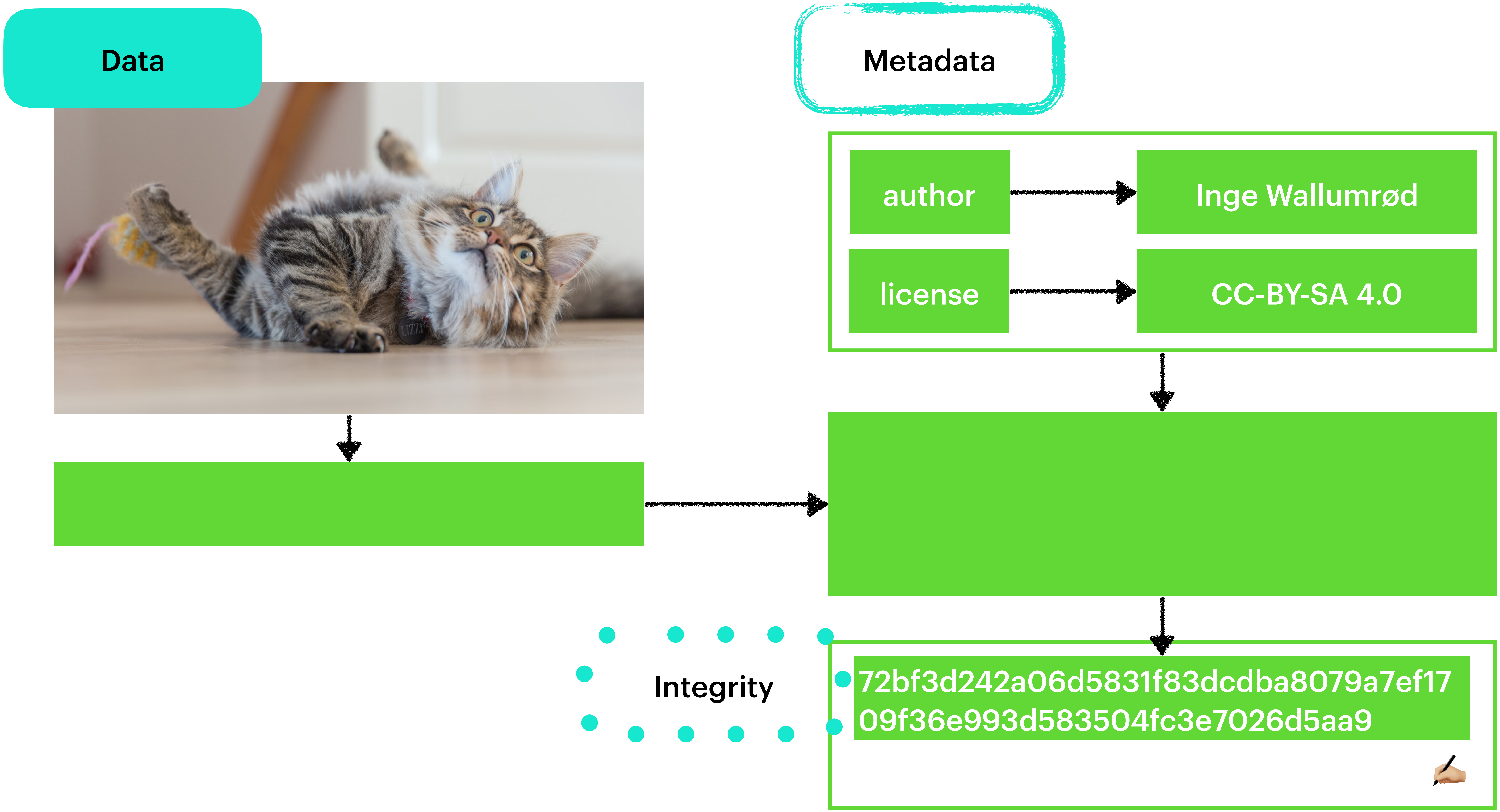
Let’s try verifying

- A verifier is given:

Data

Metadata

Integrity



Data

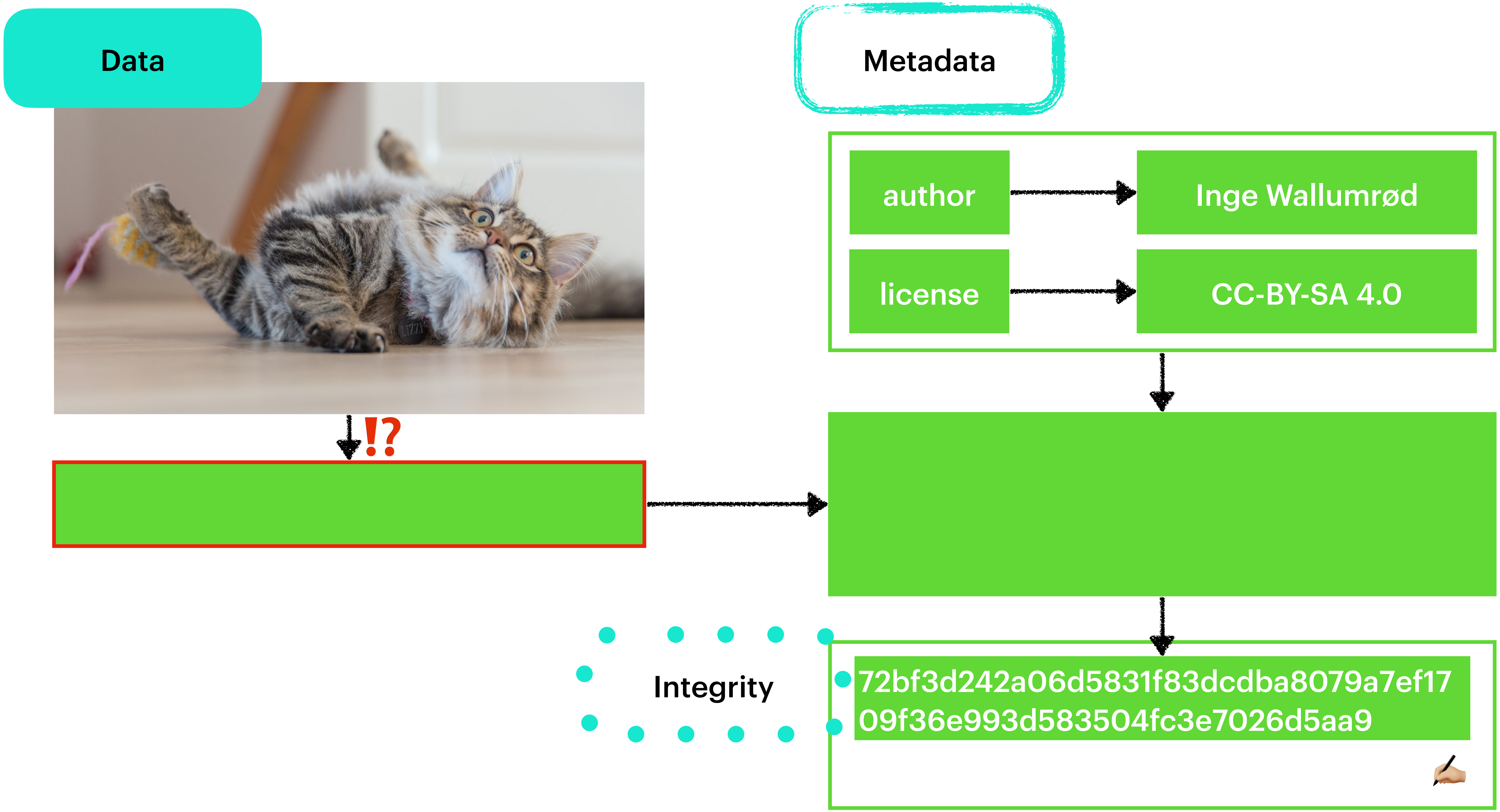
Metadata

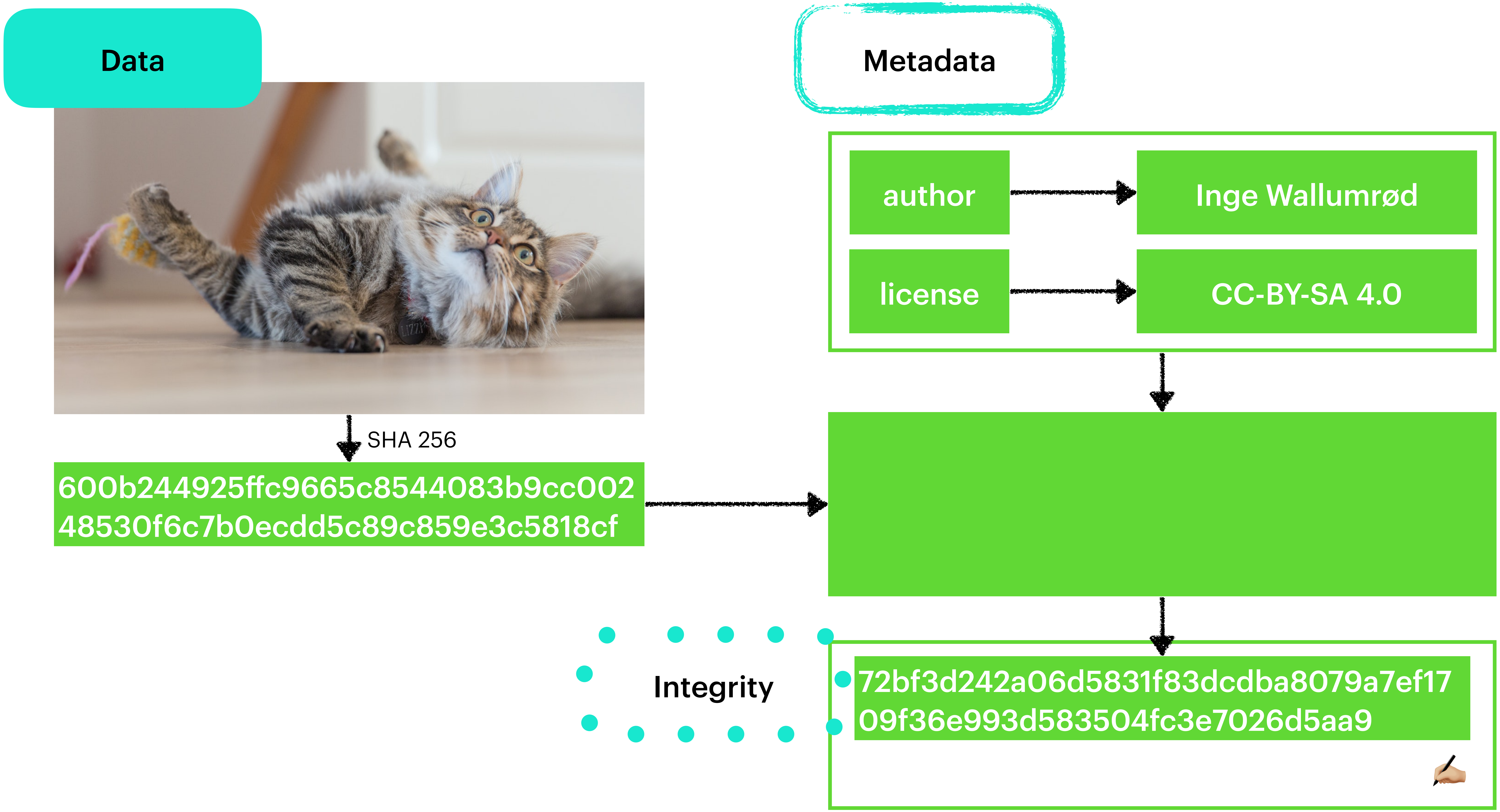
author	Inge Wallumrød
license	CC-BY-SA 4.0

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

✍

Integrity


Data
Metadata
author
Inge Wallumrød
license
CC-BY-SA 4.0
⁉
Integrity
72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9
✍



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author	Inge Wallumrød
license	CC-BY-SA 4.0

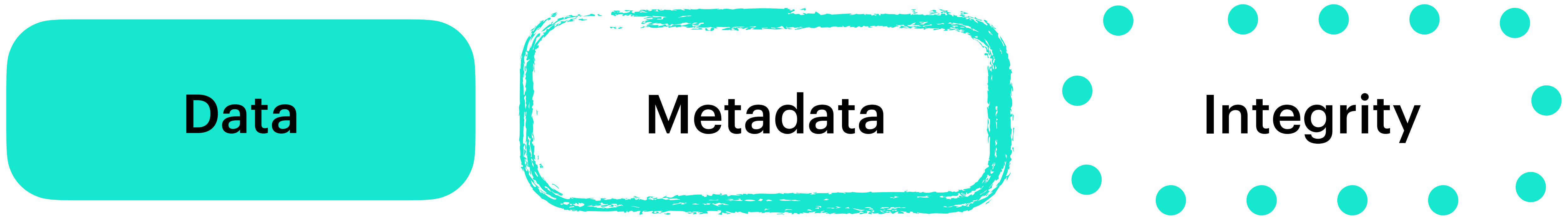
Integrity

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

✍

Let's try verifying

- A verifier is given:

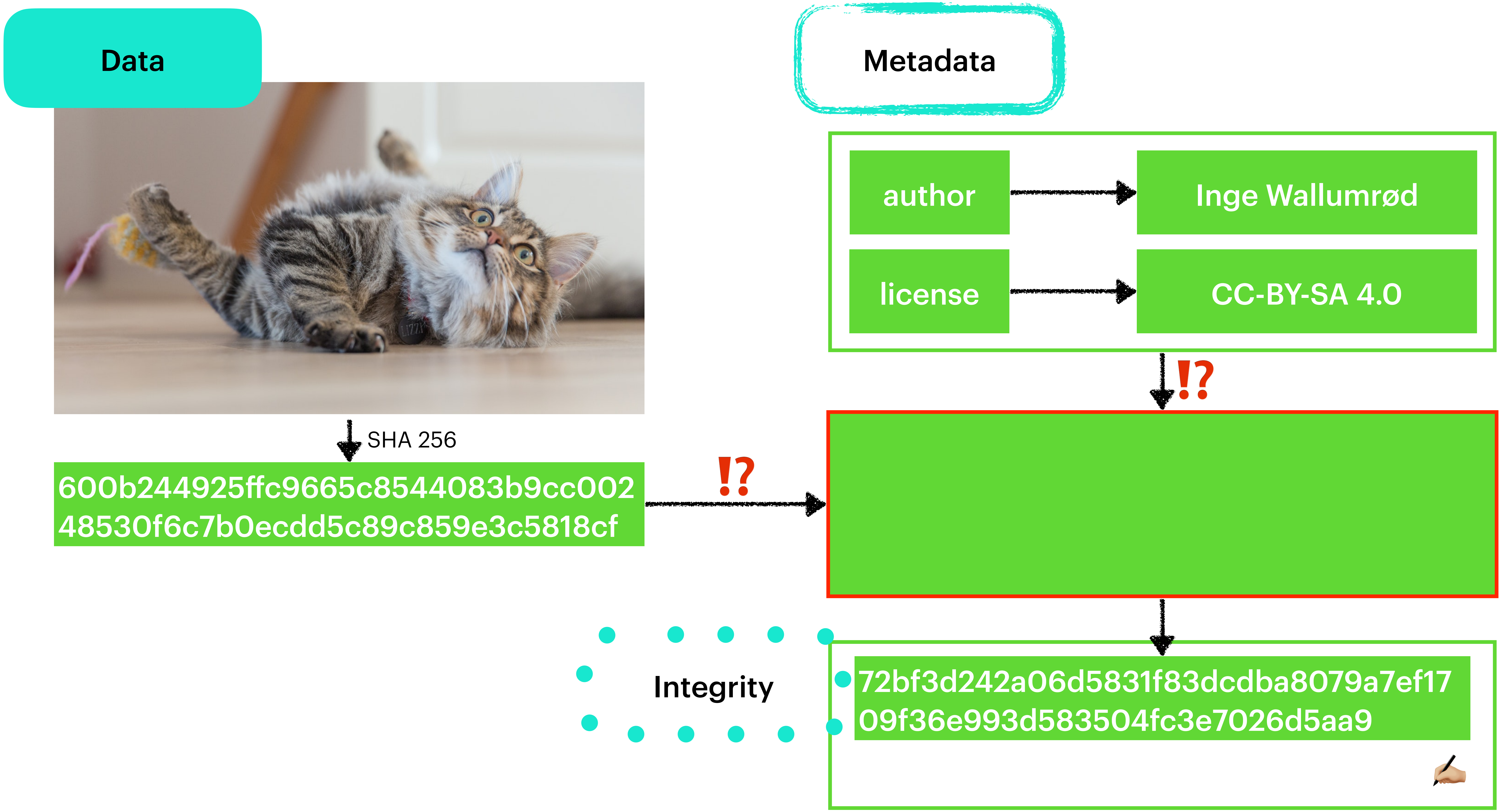


Data

Metadata

Integrity

- The verifier also needs:
  - The data hashing algorithm



Data

Metadata

author → Inge Wallumrød

license → CC-BY-SA 4.0

⁉

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

⁉

Integrity

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

✍🏻

```
hash,600b244925ffc9665c8544083b9cc00248530f6c7b0ecdd5c89c859e3c5818cf
license,CC-BY-SA 4.0
author,Inge Wallumrød
```

```
{
   “data”:”600b244925ffc9665c8544083b9cc00248530f6c7b0ecdd5c89c859e3c5818cf”,
   “license:CC-BY-SA 4.0”,
   “author”:"Inge Wallumrød”
}
```

```
{
   “data”:”600b244925ffc9665c8544083b9cc00248530f6c7b0ecdd5c89c859e3c5818cf”,
   “author”:"Inge Wallumrød”,
   “license:CC-BY-SA 4.0”
}
```

Let’s try verifying

- A verifier is given:

Data

Metadata

Integrity

- The verifier also needs:
  - The data hashing algorithm
  - The metadata packaging system

Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author → Inge Wallumrød

license → CC-BY-SA 4.0

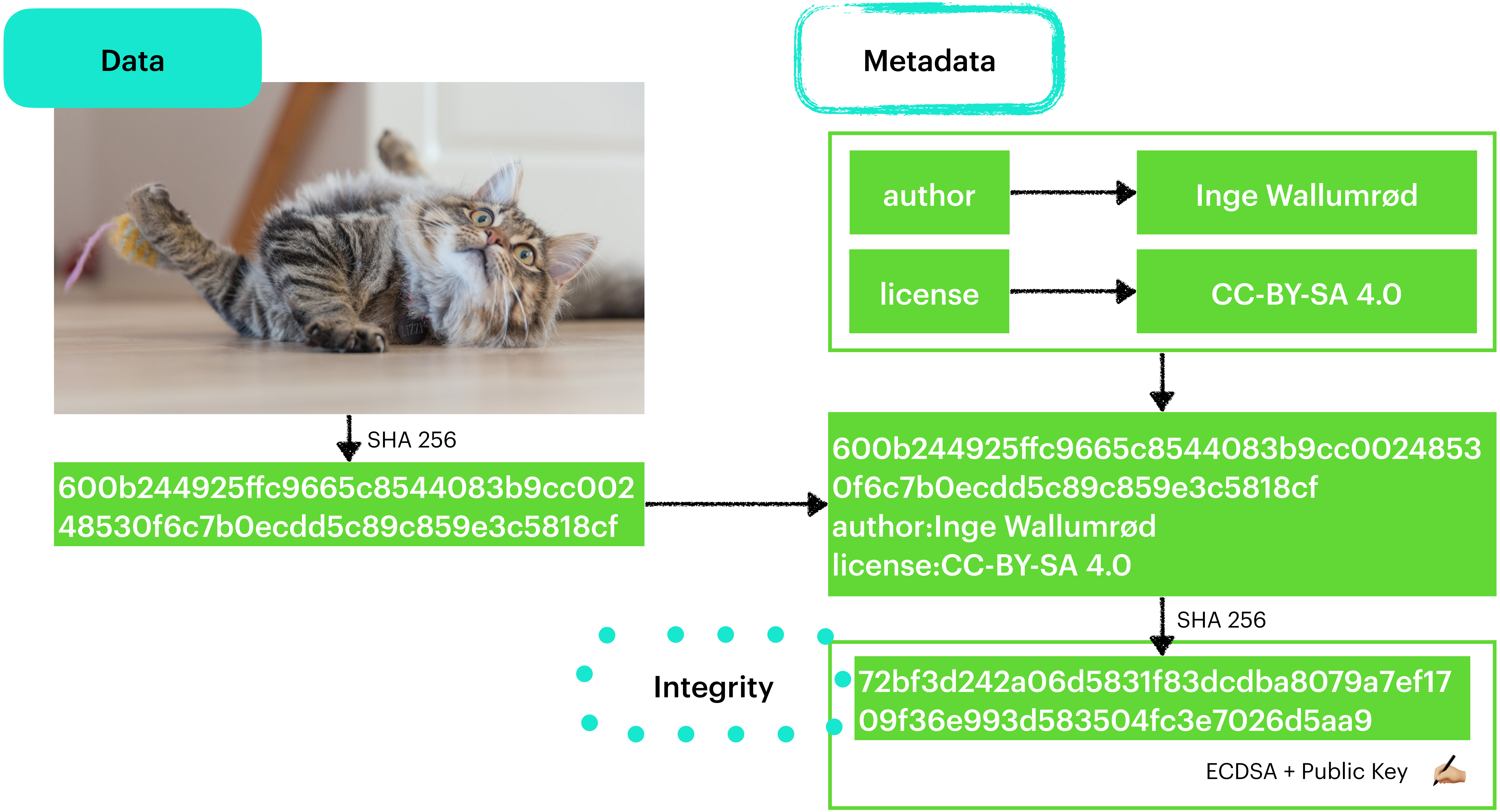
600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

⁉

Integrity

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

⁉✍🏼



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author → Inge Wallumrød

license → CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

ECDSA + Public Key ✍

Integrity

Let’s try verifying

- A verifier is given:

Data

Metadata

Integrity

- The verifier also needs:
  - The data hashing algorithm
  - The metadata packaging system
  - The signature verification method
  - The public key to verify against

Some standards we've implemented

- A verifier is given:

Data

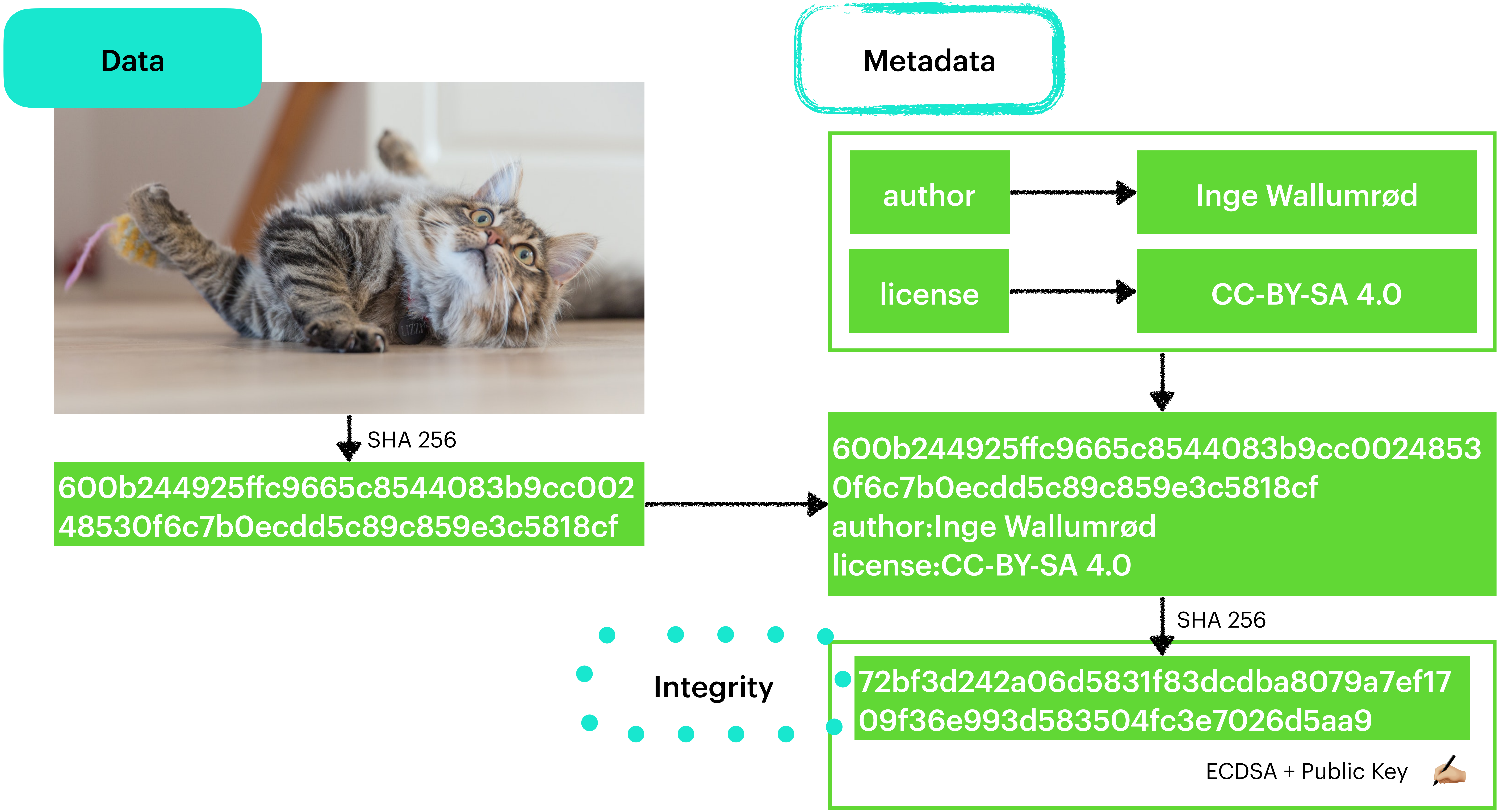
Metadata

Integrity

- The verifier also needs:
  - The data hashing algorithm **(CID, Blake3)**
  - The metadata packaging system **(C2PA, Numbers Protocol, ISCN, JCS, JSON-LD, RDFC, Authenticated Attributes)**
  - The signature verification method **(Verifiable Credentials, ZK proofs)**
  - The public key to verify against **(GPG, DID, BBS+)**

Back to our original design goal ...

I want to ensure my data, and its associated metadata, is not tampered.



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author	Inge Wallumrød
license	CC-BY-SA 4.0

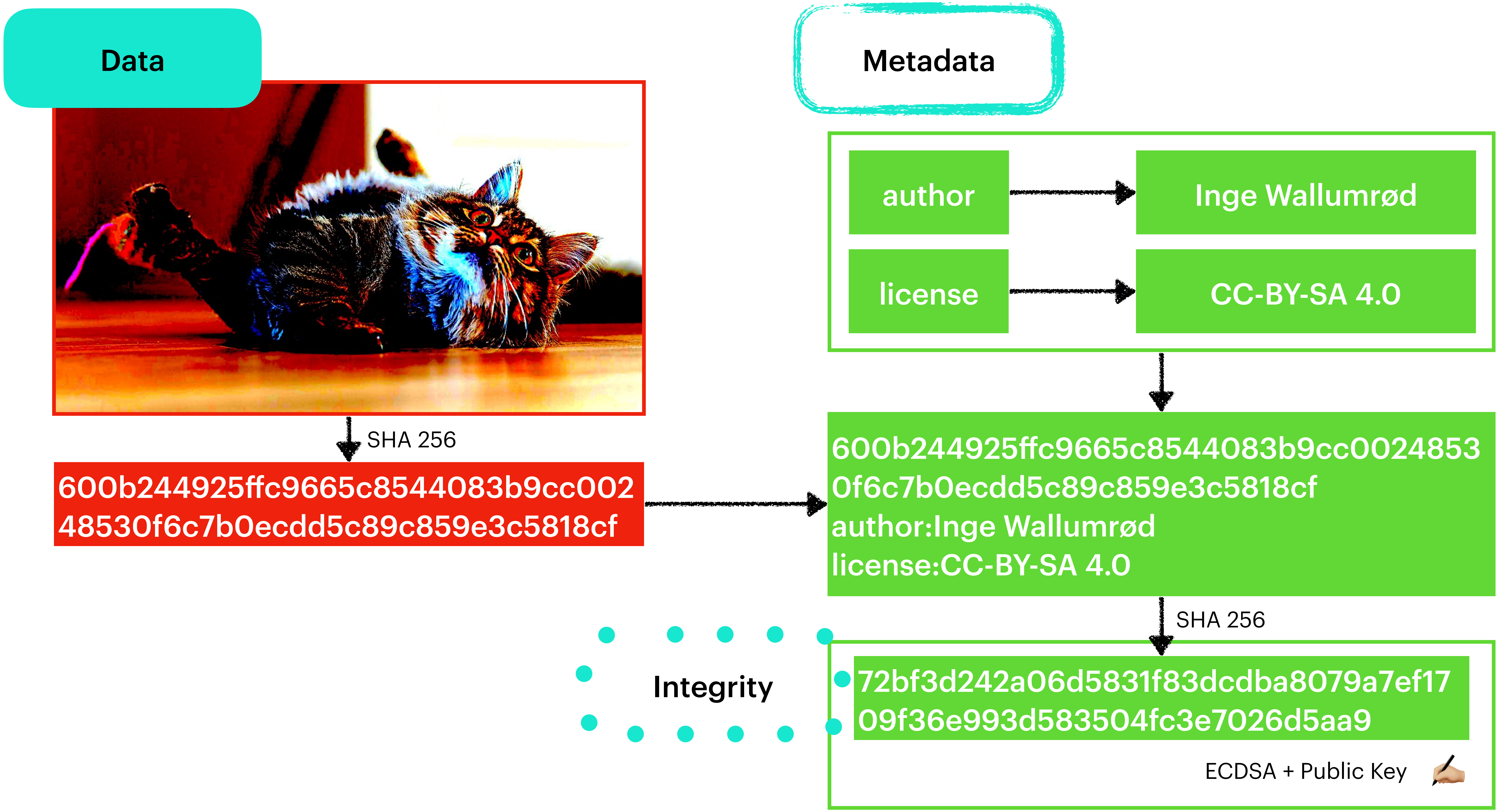
600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

ECDSA + Public Key ✍

Integrity



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author	Inge Wallumrød
license	CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

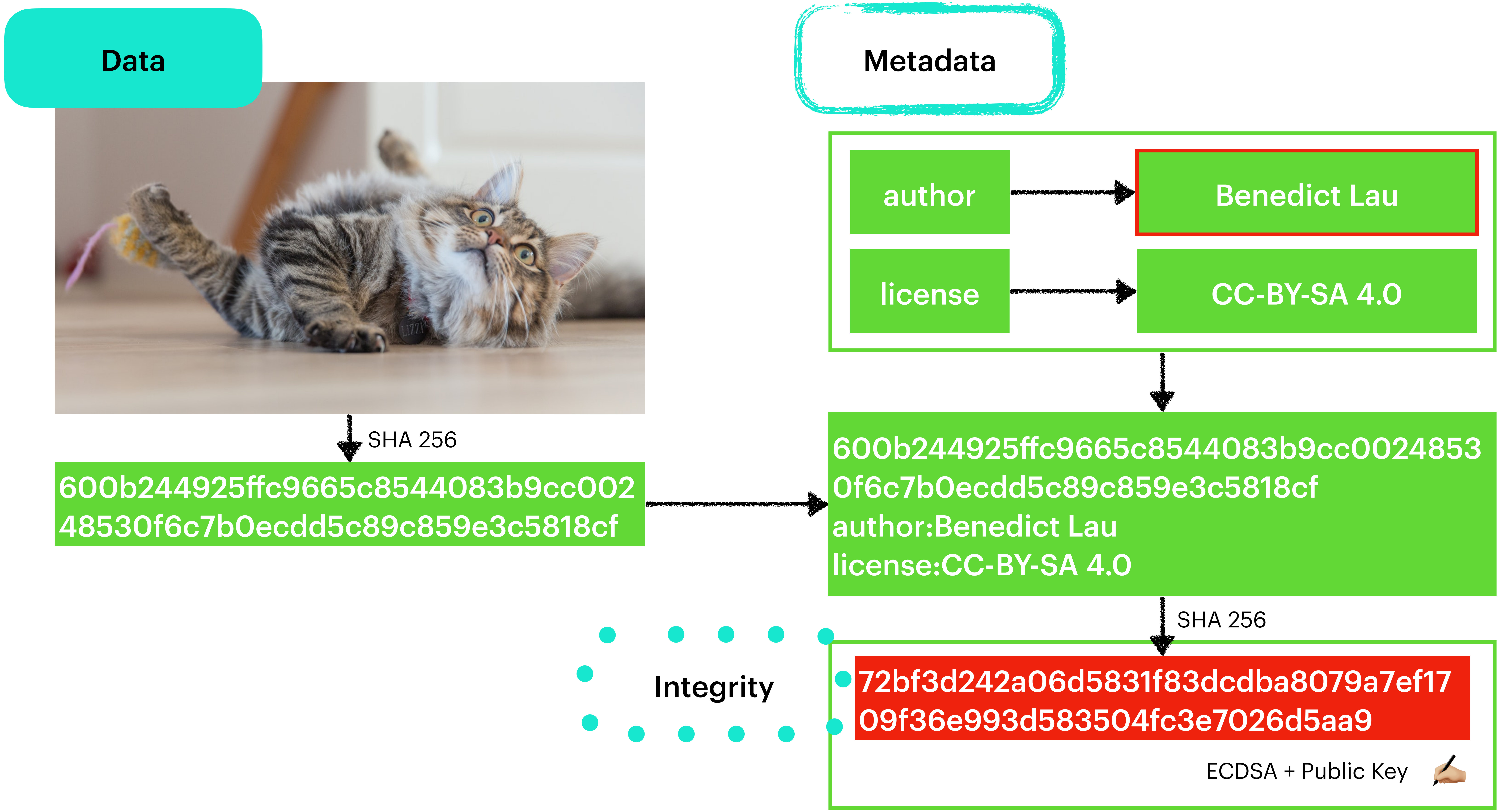
SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

ECDSA + Public Key ✍

Integrity

Data Tampering



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author	Benedict Lau
license	CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Benedict Lau
license:CC-BY-SA 4.0

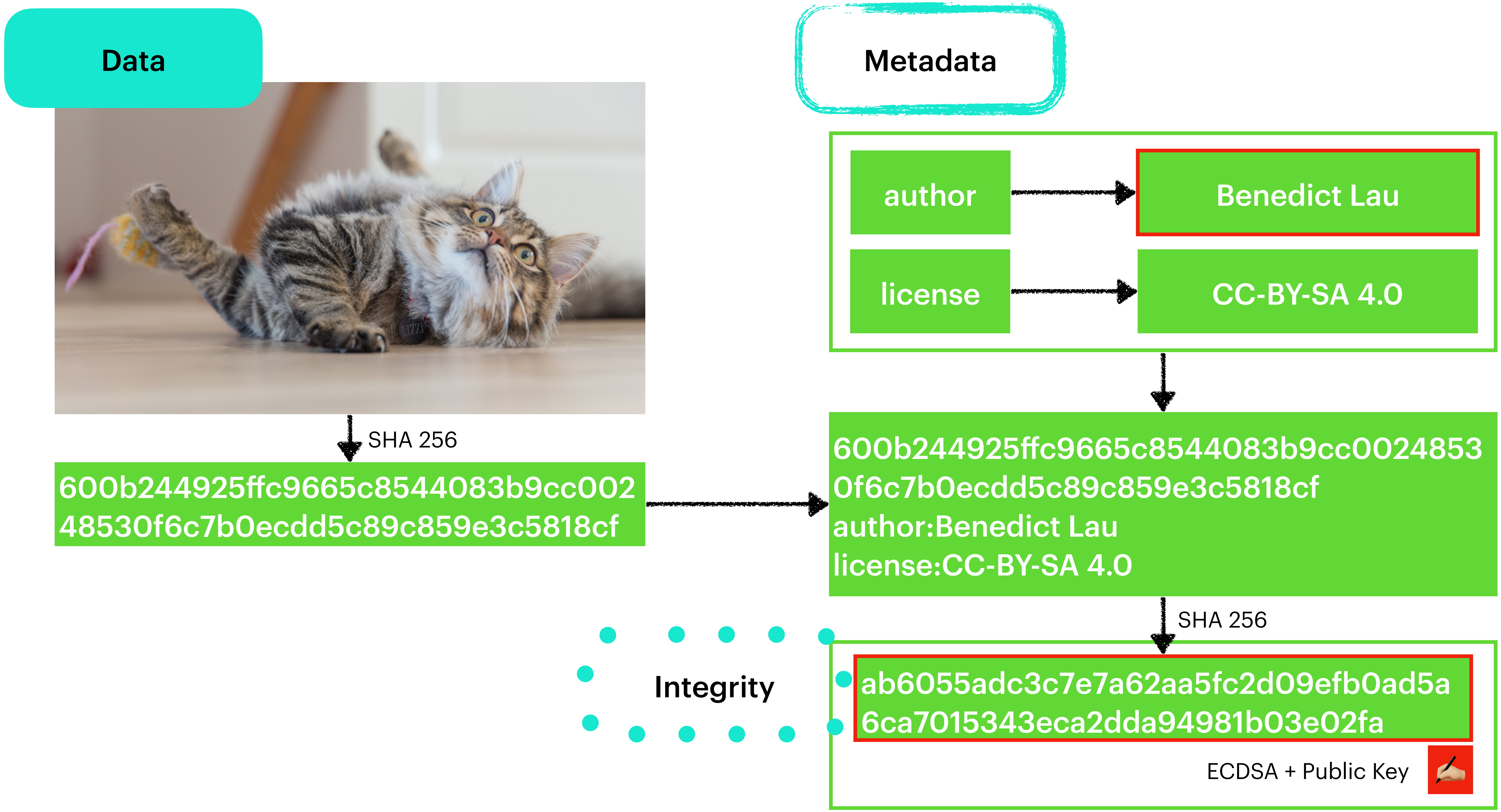
SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

ECDSA + Public Key ✍

Integrity

Metadata Tampering



Data

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

Metadata

author	Benedict Lau
license	CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Benedict Lau
license:CC-BY-SA 4.0

SHA 256

ab6055adc3c7e7a62aa5fc2d09efb0ad5a
6ca7015343eca2dda94981b03e02fa

ECDSA + Public Key ✍

Integrity

Signature Tampering

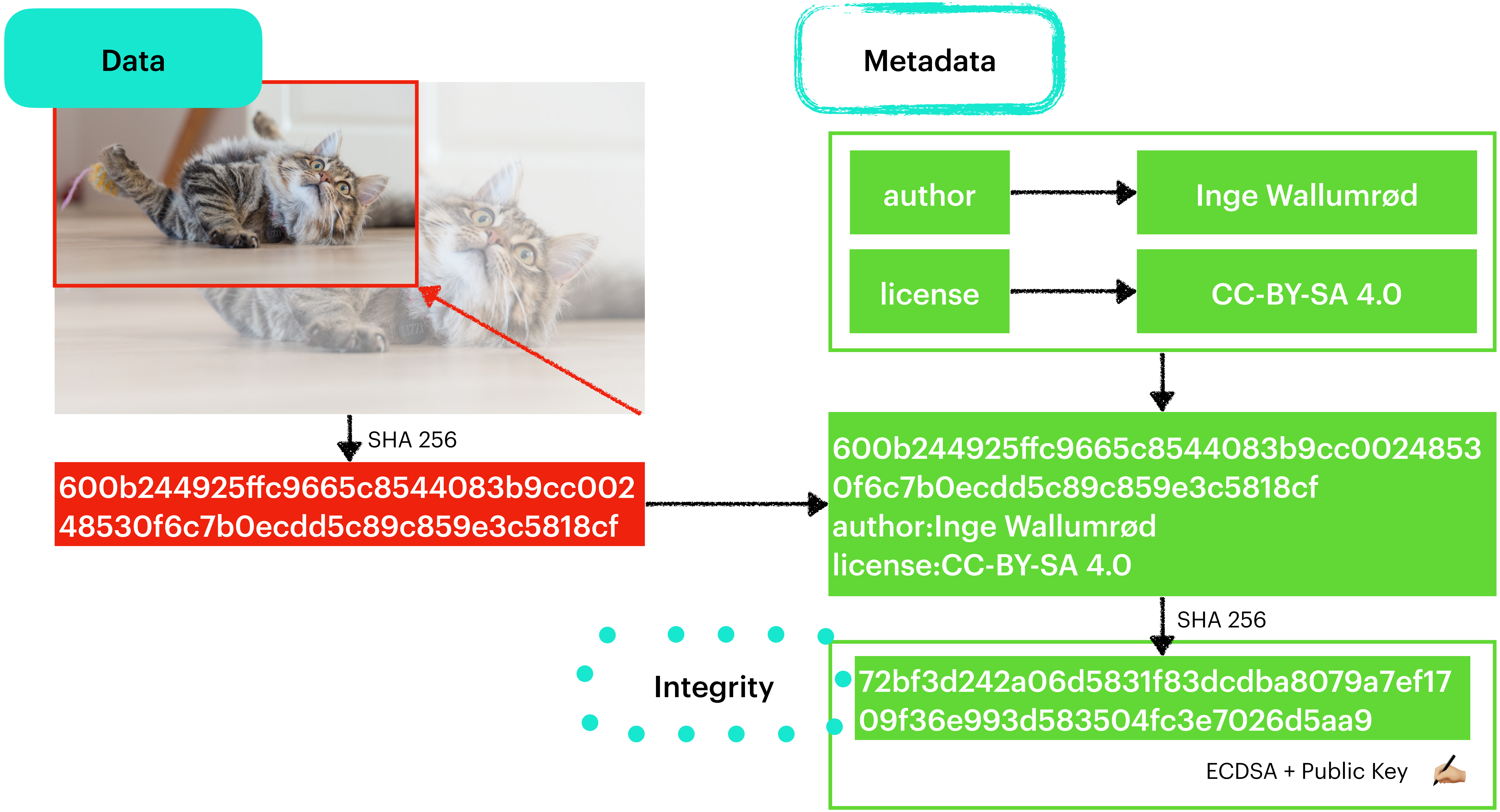
More accurately ...

~~I want to ensure my data, and its associated metadata, is not tampered.~~

Tampered versions of my data, and its associated metadata, cannot be attributed to me.

Scenarios

What if I resize the picture?



Data

Metadata

author → Inge Wallumrød

license → CC-BY-SA 4.0

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød
license:CC-BY-SA 4.0

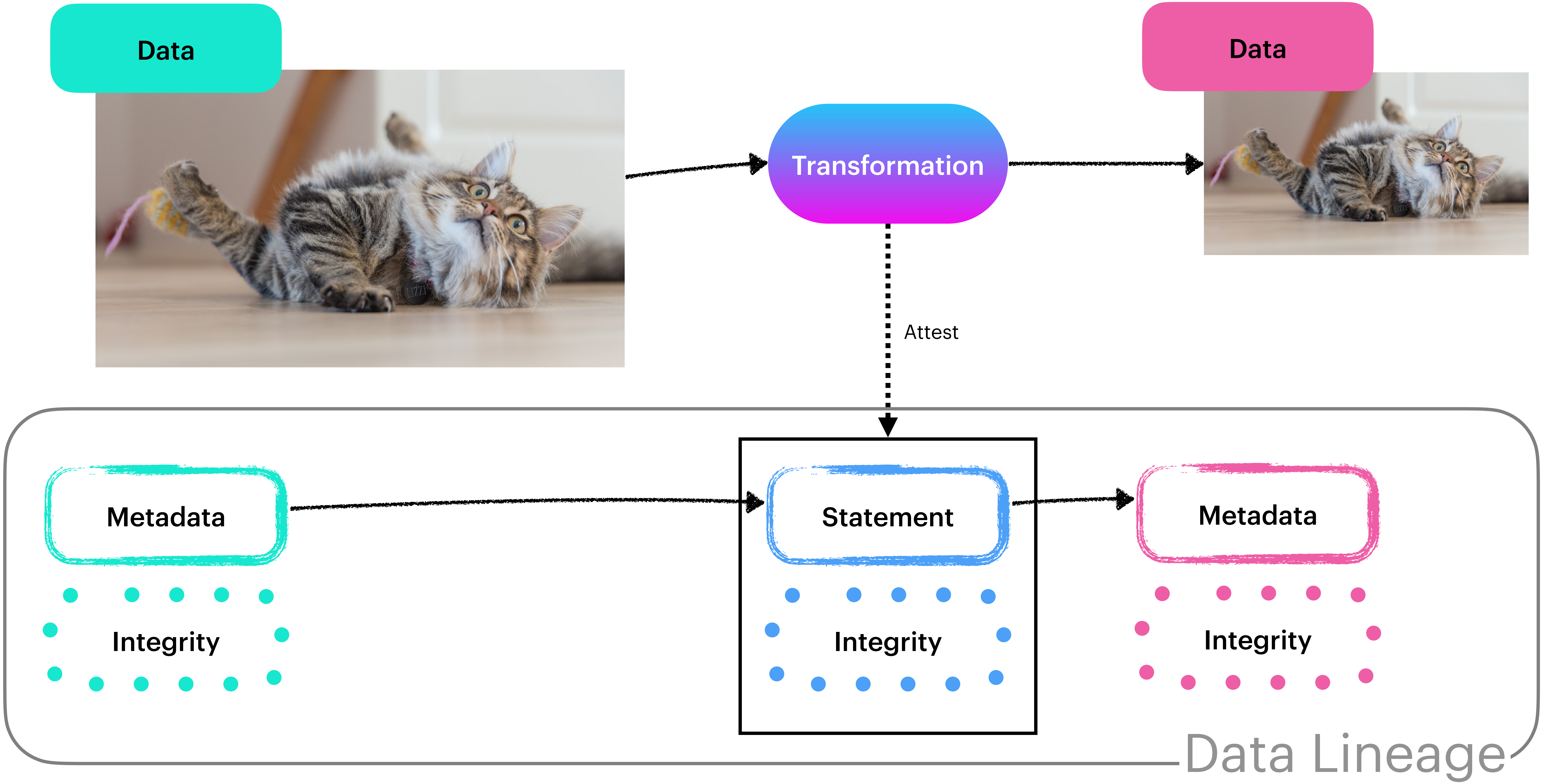
SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

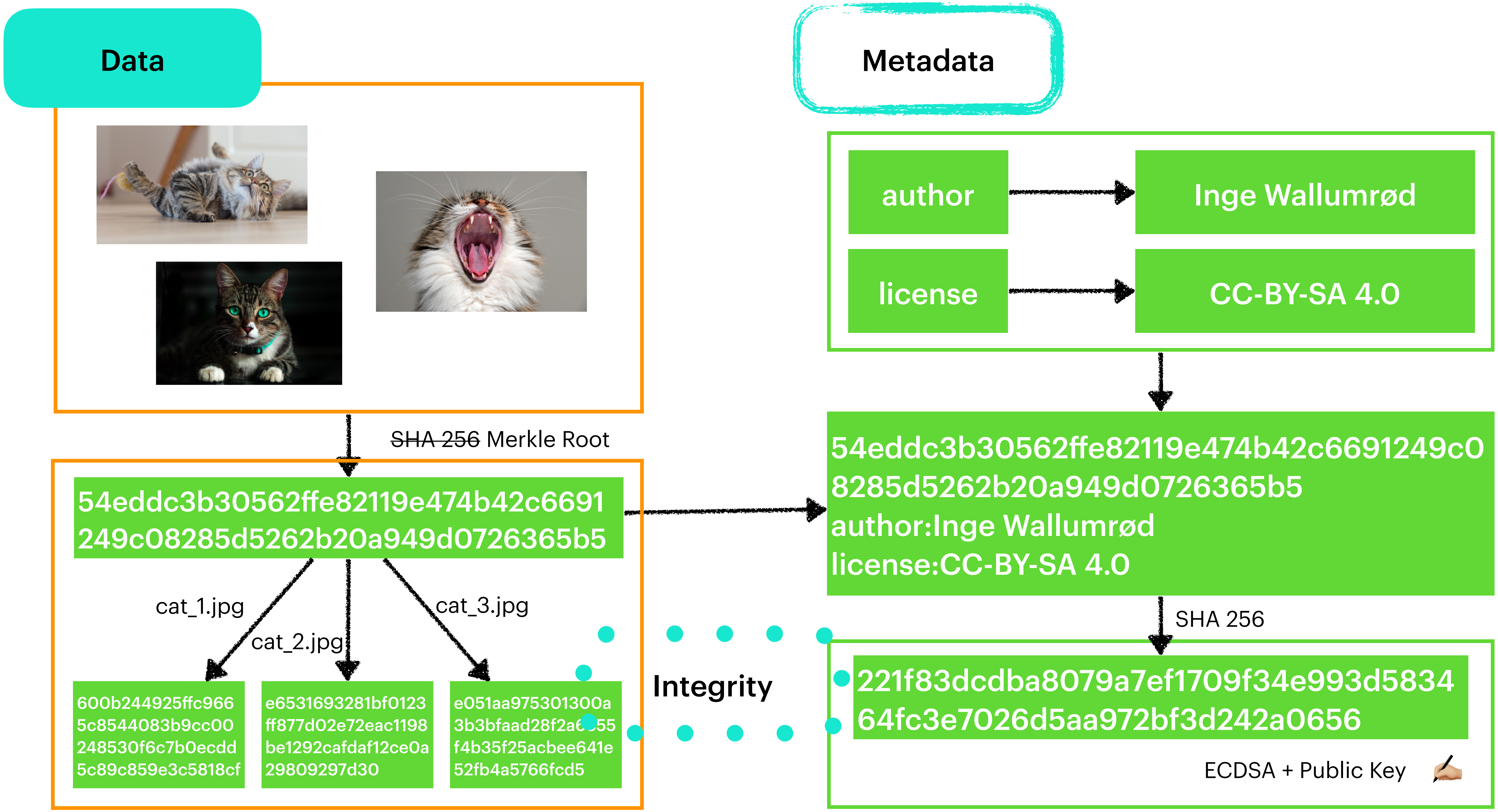
ECDSA + Public Key ✍🏼

Integrity

… or do legitimate edits, or necessary redactions, or transcoding, or when web optimization kicks in …


Data
Transformation
Data
Attest
Metadata
Integrity
Statement
Integrity
Metadata
Integrity
Data Lineage

What if my data is a collection of files?



Data

~~SHA 256~~ Merkle Root

54eddc3b30562ffe82119e474b42c6691
249c08285d5262b20a949d0726365b5

cat_1.jpg

cat_2.jpg

cat_3.jpg

600b244925ffc966
5c8544083b9cc00
248530f6c7b0ecdd
5c89c859e3c5818cf

e6531693281bf0123
ff877d02e72eac1198
be1292cafdaf12ce0a
29809297d30

e051aa975301300a
3b3bfaad28f2a6[illegible]55
f4b35f25acbee641e
52fb4a5766fcd5

Integrity

Metadata

author	Inge Wallumrød
license	CC-BY-SA 4.0

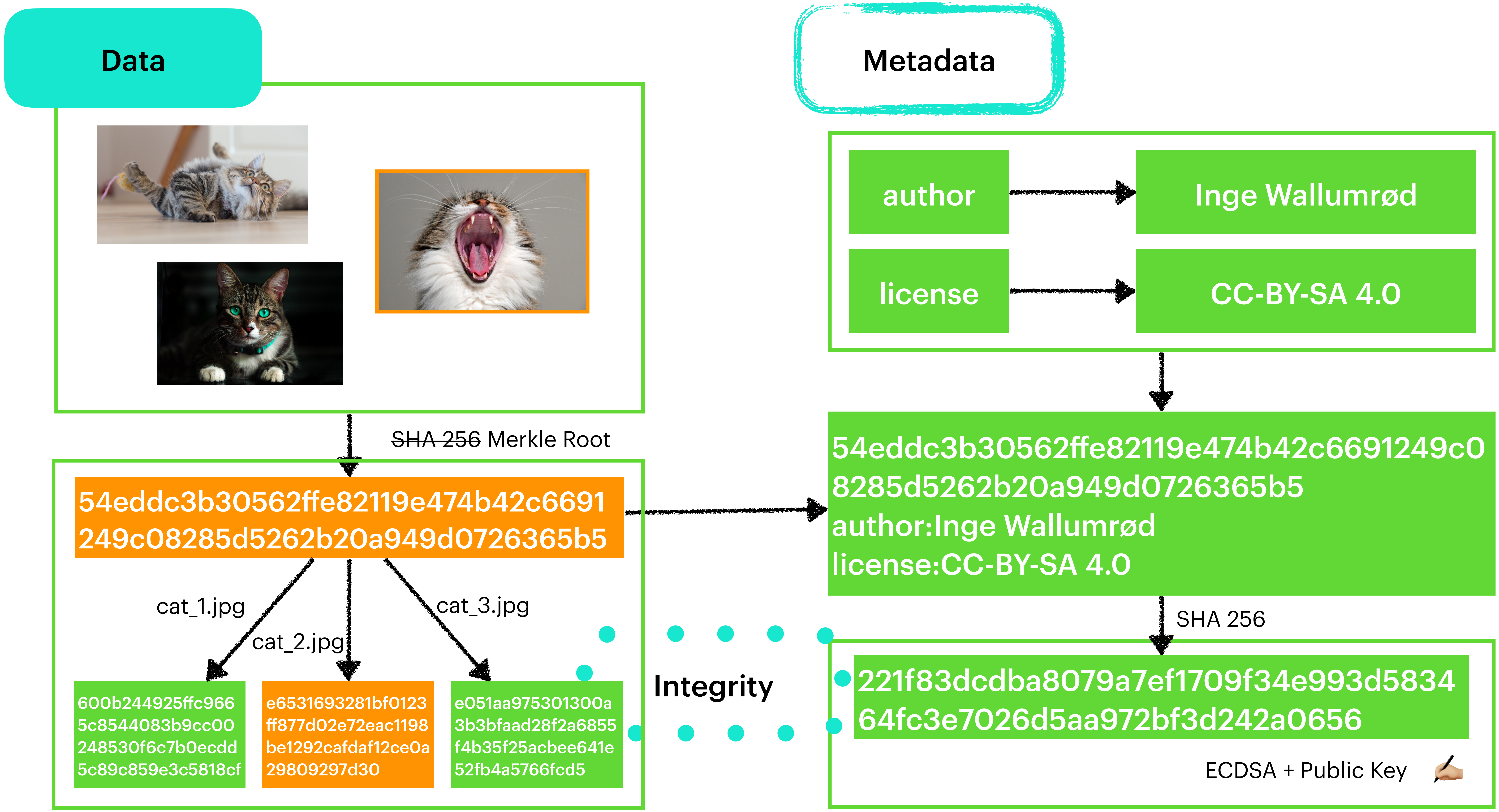
54eddc3b30562ffe82119e474b42c6691249c0
8285d5262b20a949d0726365b5
author:Inge Wallumrød
license:CC-BY-SA 4.0

SHA 256

221f83dcdba8079a7ef1709f34e993d5834
64fc3e7026d5aa972bf3d242a0656

ECDSA + Public Key ✍

… and if I need to reference a subset of files in my collection, or a data range within a single file …



Data

Metadata

author → Inge Wallumrød

license → CC-BY-SA 4.0

~~SHA 256~~ Merkle Root

54eddc3b30562ffe82119e474b42c6691
249c08285d5262b20a949d0726365b5

cat_1.jpg

cat_2.jpg

cat_3.jpg

600b244925ffc966
5c8544083b9cc00
248530f6c7b0ecdd
5c89c859e3c5818cf

e6531693281bf0123
ff877d02e72eac1198
be1292cafdaf12ce0a
29809297d30

e051aa975301300a
3b3bfaad28f2a6855
f4b35f25acbee641e
52fb4a5766fcd5

54eddc3b30562ffe82119e474b42c6691249c0
8285d5262b20a949d0726365b5
author:Inge Wallumrød
license:CC-BY-SA 4.0

SHA 256

221f83dcdba8079a7ef1709f34e993d5834
64fc3e7026d5aa972bf3d242a0656

ECDSA + Public Key ✍

Integrity

Let’s try verifying

- A verifier is given:

Data

Metadata

Integrity

221f83dcdba8079a7ef1709f34e993d5834
64fc3e7026d5aa972bf3d242a0656

ECDSA + Public Key ✍

54eddc3b30562ffe82119e474b42c6691
249c08285d5262b20a949d0726365b5

e6531693281bf0123
ff877d02e72eac1198
be1292cafdaf12ce0a
29809297d30

Merkle Path

Inclusion Proof

Let’s try verifying

Inclusion proof fails verification if the data is tampered

- A verifier is given:

Data

Metadata

Integrity

221f83dcdba8079a7ef1709f34e993d5834
64fc3e7026d5aa972bf3d242a0656

ECDSA + Public Key ✍

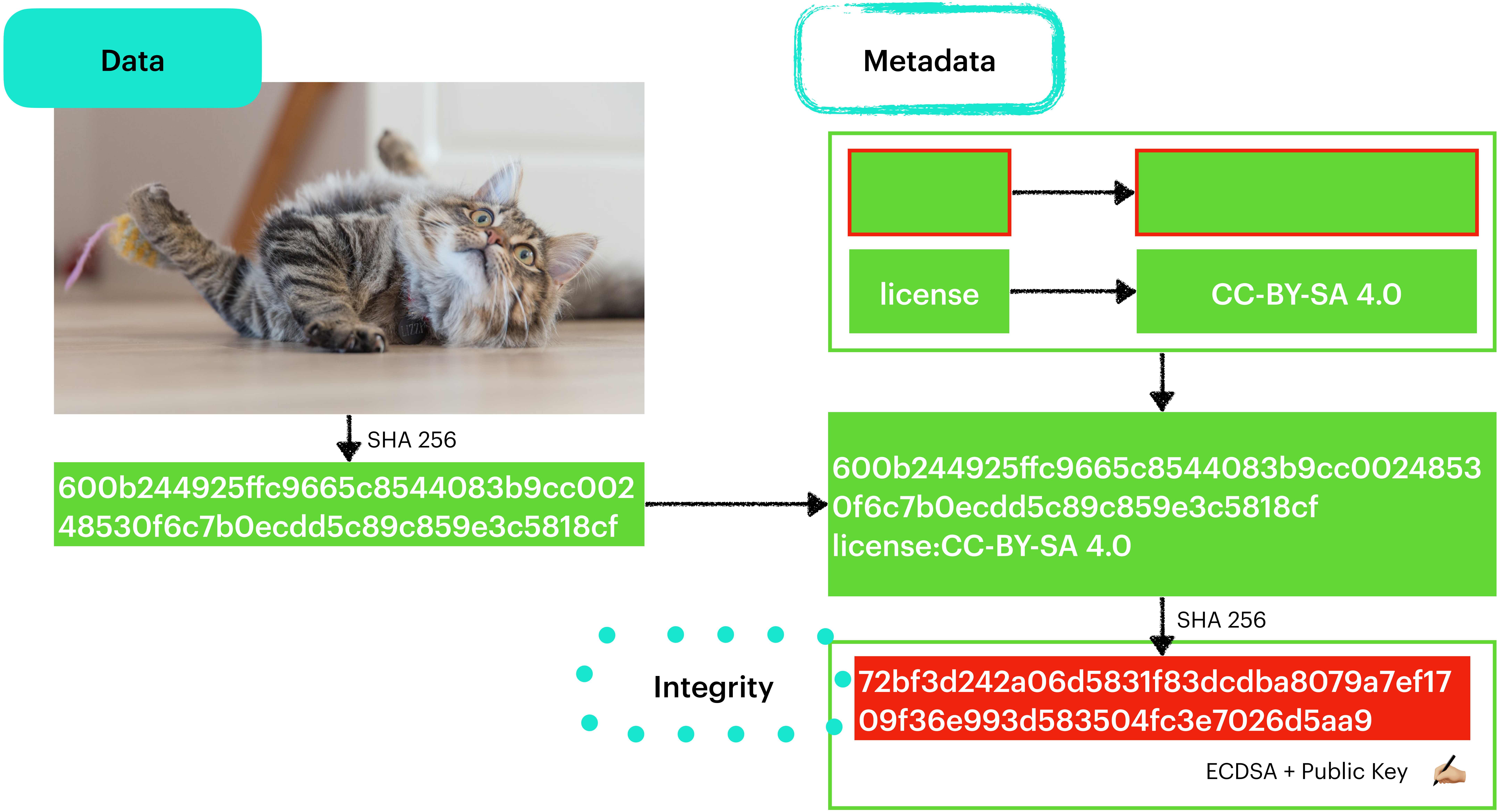
54eddc3b30562ffe82119e474b42c6691
249c08285d5262b20a949d0726365b5

e6531693281bf0123
ff877d02e72eac1198
be1292cafdaf12ce0a
29809297d30

Merkle Path

Inclusion Proof

What if I need to redact part of the original *metadata*?



Data

Metadata

license

CC-BY-SA 4.0

SHA 256

600b244925ffc9665c8544083b9cc002
48530f6c7b0ecdd5c89c859e3c5818cf

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
license:CC-BY-SA 4.0

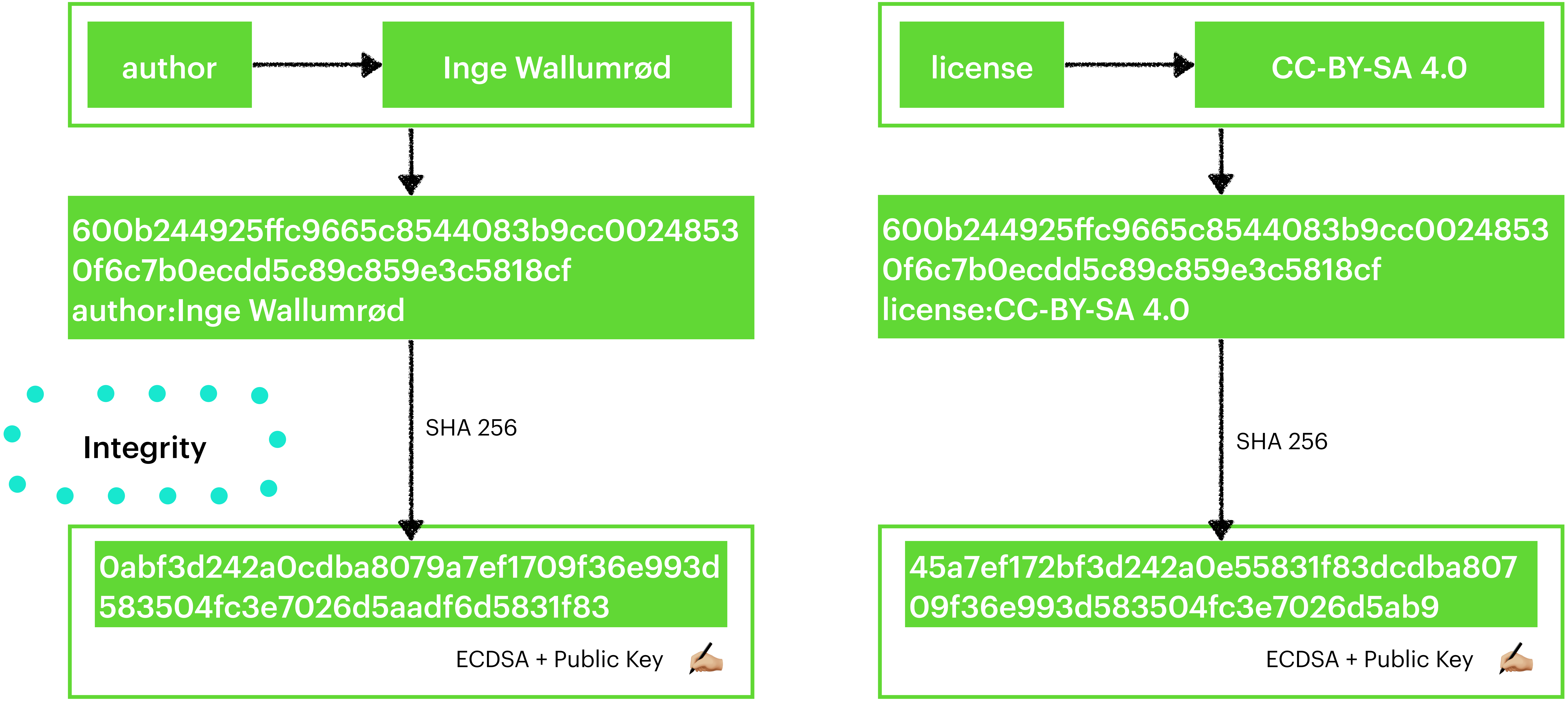
SHA 256

72bf3d242a06d5831f83dcdba8079a7ef17
09f36e993d583504fc3e7026d5aa9

ECDSA + Public Key ✍

Integrity

Metadata



Integrity

author → Inge Wallumrød

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
author:Inge Wallumrød

SHA 256

0abf3d242a0cdba8079a7ef1709f36e993d
583504fc3e7026d5aadf6d5831f83

ECDSA + Public Key ✍

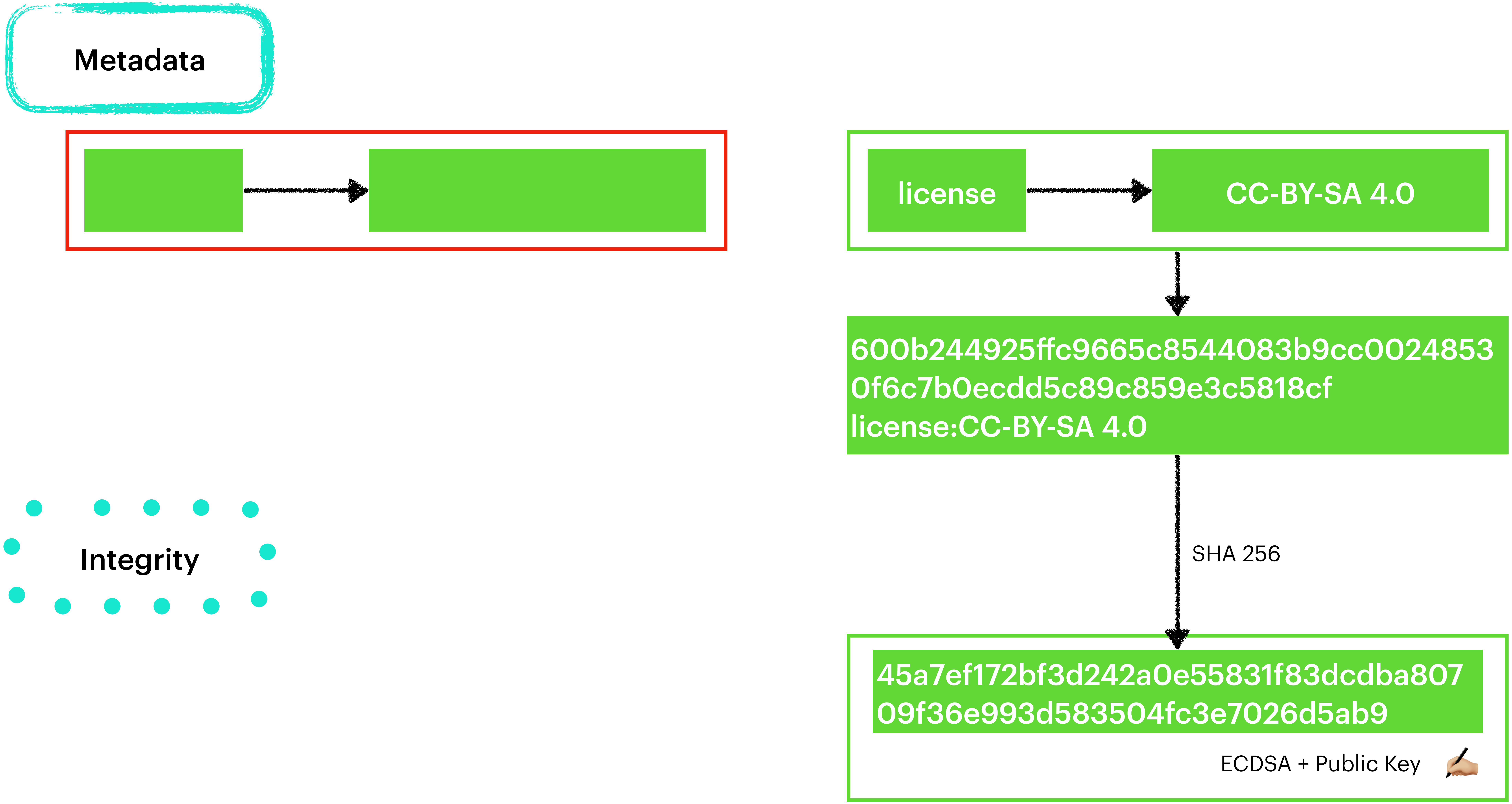
license → CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
license:CC-BY-SA 4.0

SHA 256

45a7ef172bf3d242a0e55831f83dcdba807
09f36e993d583504fc3e7026d5ab9

ECDSA + Public Key ✍



Metadata

license → CC-BY-SA 4.0

600b244925ffc9665c8544083b9cc0024853
0f6c7b0ecdd5c89c859e3c5818cf
license:CC-BY-SA 4.0

SHA 256

45a7ef172bf3d242a0e55831f83dcdba807
09f36e993d583504fc3e7026d5ab9

ECDSA + Public Key ✍🏼

Integrity

I want to know *when* something is attested.

These scenarios are just the beginning ...

What if two cats are talking about the *same* data, but one used SHA 256 and another Blake3 to fingerprint them?

How do I know no one is *withholding* a subset of attributes from me?

Benedict Lau
benedict@hypha.coop