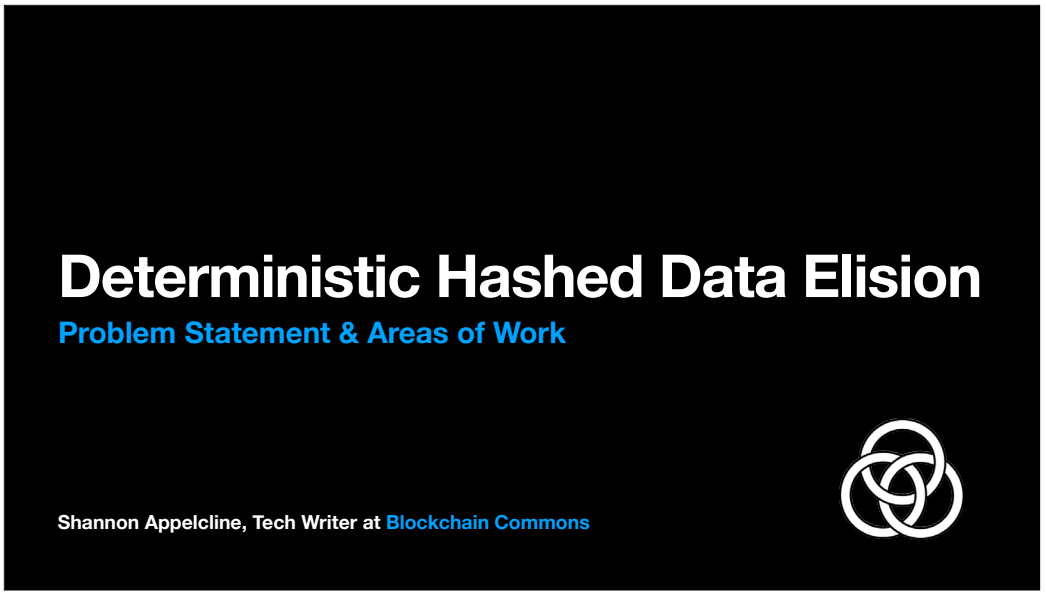


Hi, all. My name is Shannon Appelcline and I'm a technical writer for Blockchain Commons. We advocate for the creation of open, interoperable, secure & compassionate digital infrastructure to enable people to control their own digital destiny and to maintain their human dignity online.

Data is
The Problem

- Poor Data Control
- Poor Data Privacy
- No Human Rights

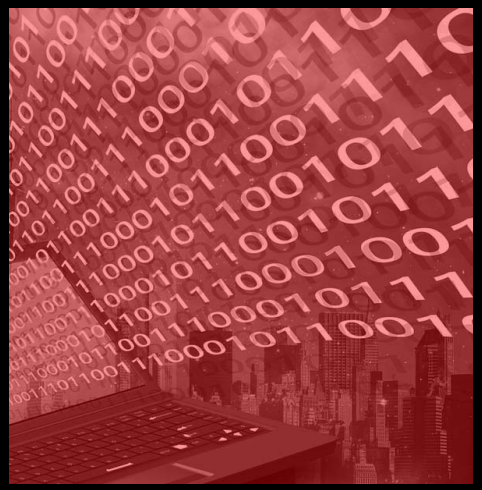
Data is the heart of the internet, but it's still the Wild West.
There are few controls on what data is shared or reshared.
Data privacy is an all-or-nothing thing, and once it's breached, it's out there.
Data is being shared with little concern for human rights.
We need to make the internet more humane.

The first issue with data is that it’s not minimized. Most interactions involve more data than is needed, even in the face of new regulations such as GDPR.

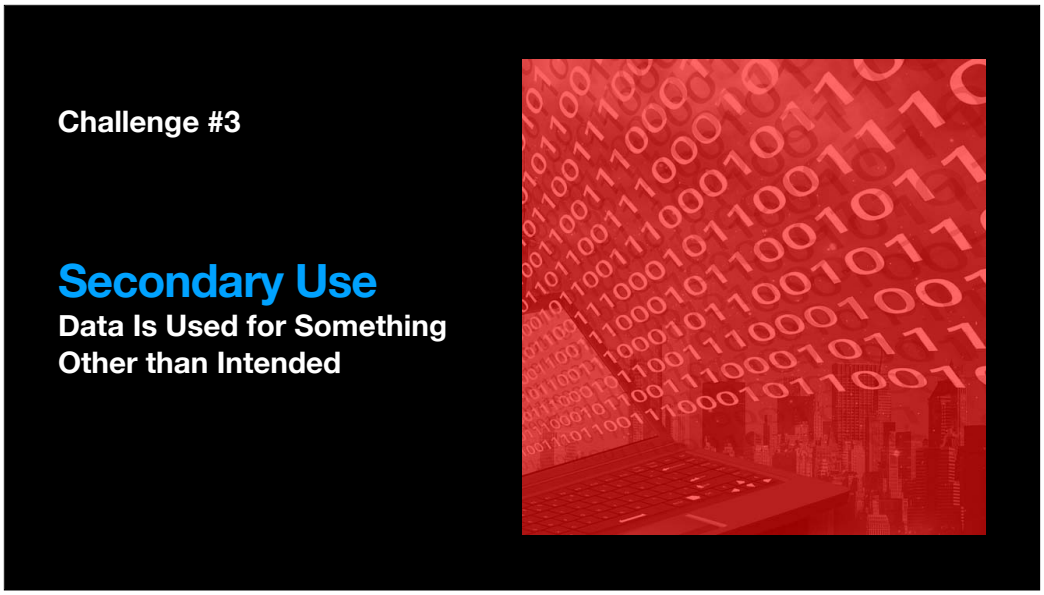
Challenge #2

Correlation

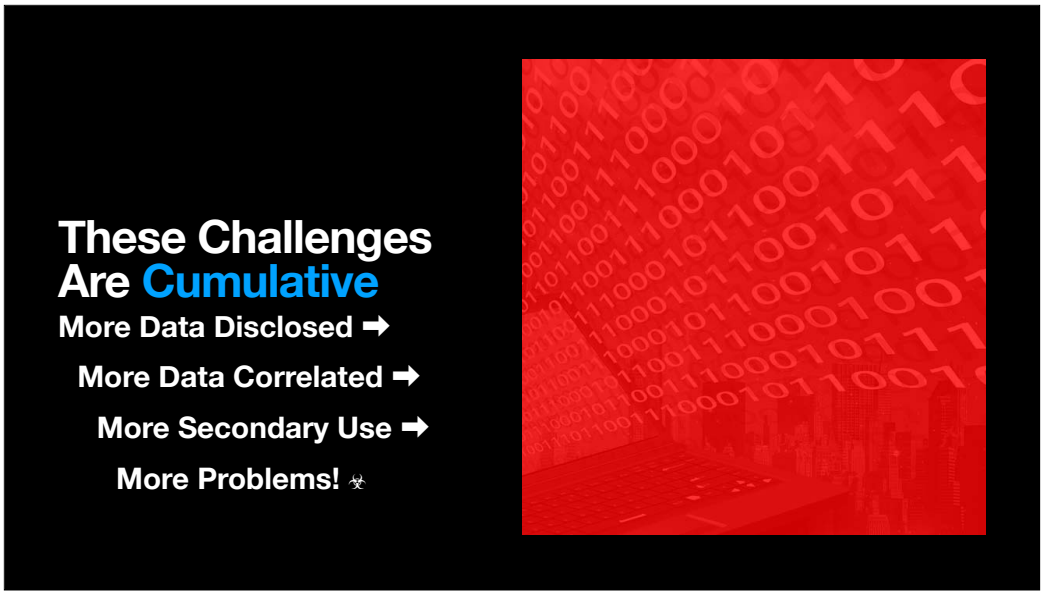
Discrete Data Can Be Aggregated



That becomes even more problematic when data gets out there. You can combine one batch of excessive data with another and suddenly you know a LOT about the data's subject.



And worse, when you have lots of data, you can use it in ways totally orthogonal to what was originally intended. You might have given your address to receive a shipment of vinyl records, but when that gets correlated with the financial data you provided to your bank as KYC, suddenly the burglars are knocking at your door (or rather pry barring your window).



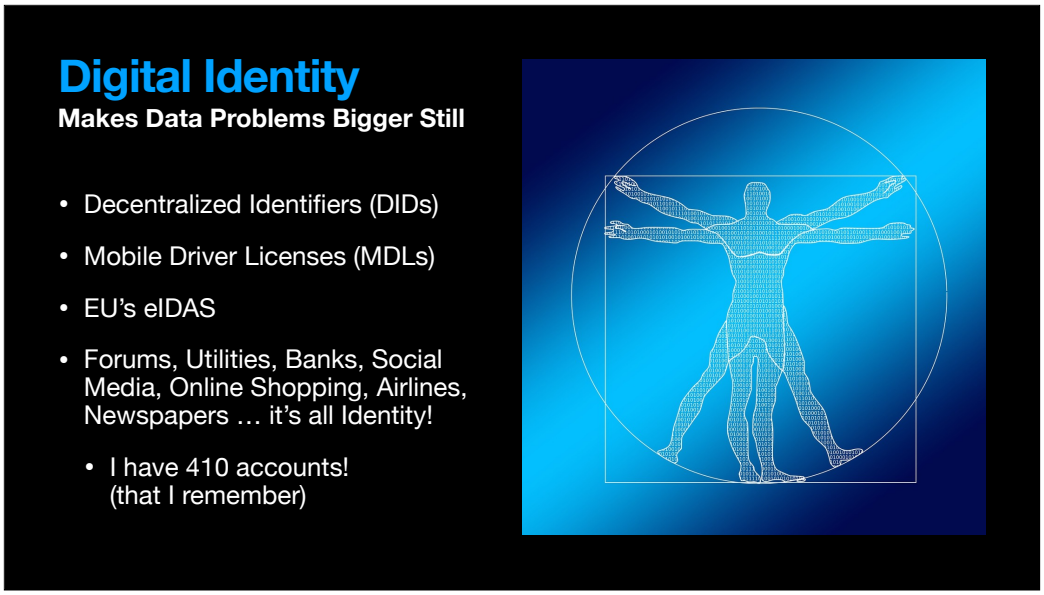
These challenges are cumulative. More data disclosed MEANS more data correlated MEANS more secondary use MEANS more problems.

The Data Problem is
Growing Larger

- More Data is Being Collected
- More Data is Online
- More Data is Sensitive

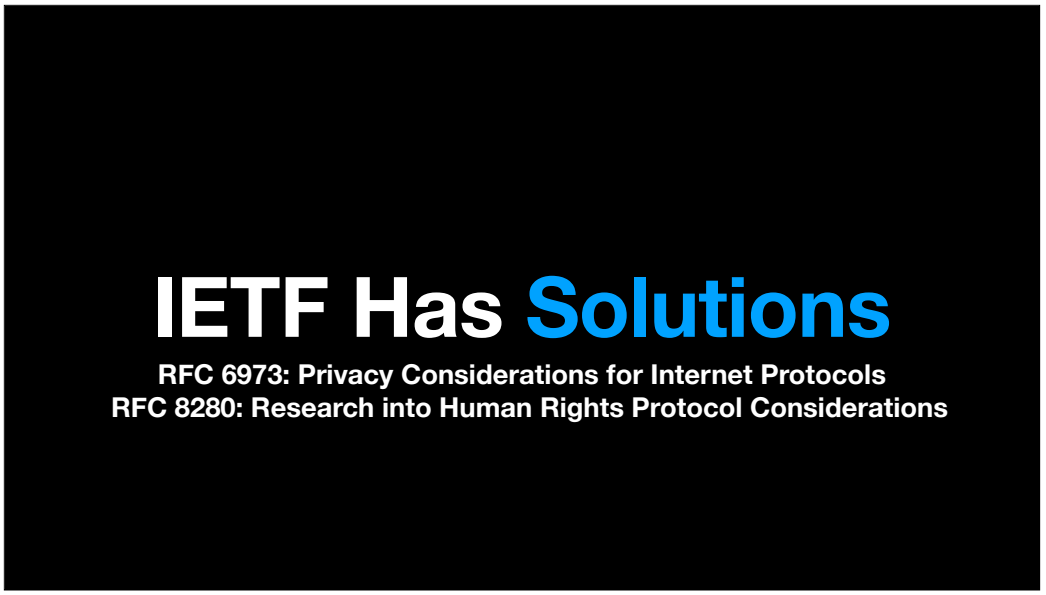
Worse, the amount of data being collected is growing every year, it’s more sensitive, and it’s more often placed online.

The activity trackers displayed here are a great example. They can record where you are and to a certain extent what you’re doing. How can we protect data like that?



Digital identity is the next big frontier. It's really coming of age right now with DIDs, MDLs, and eIDAS, but it's been around for a while. My shockingly huge count of 410 online accounts contain a lot of personally identifiable information, particularly if they're correlated and used in unexpected ways.

But honestly, identity, credentials, and healthcare are just the tip of the digital iceberg. How much information could a competitor gain if they accessed shipping records? How much espionage could a foreign country commit if they tracked the executive office's FitBits? How much trouble could hackers cause if they broke the chain of identity in software releases? WE NEED TO GET IN FRONT OF THIS PROBLEM.



IETF has some general solutions for data privacy & human rights: RFCs 6973 & 8280

RFC 6973 talks about how to introduce privacy into the design of Internet protocols.

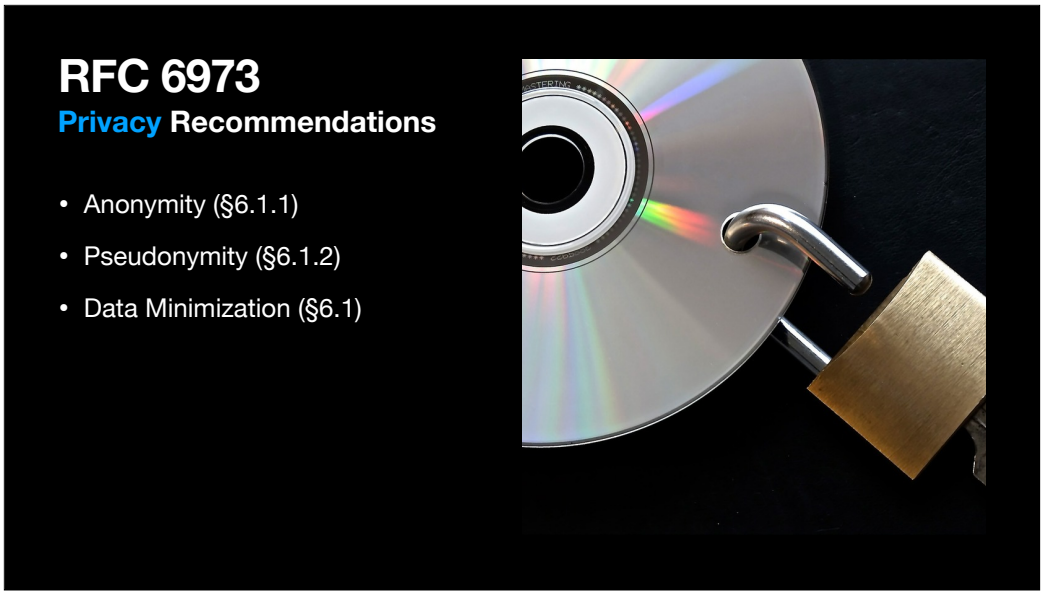
RFC 8280 expands that with a look at human rights considerations such as open, secure, and reliable connectivity.

RFCs 6973 & 8280

Aren't Securing Data

- They're Somewhat Dated
- They're Not Concrete
- They're Not Required
- They're Not Used

Unfortunately, these RFCs are not enough. Because they're not concrete and they're not required, they're not being used. Even if they were, new privacy innovations and requirements have appeared in the last decade. The RFCs are dated.



For example, let’s take a look at RFC 6973. Its first three recommendations for privacy are to incorporate anonymity, pseudonymity, and data minimization.

Anonymity / Pseudonymity

Privacy Problems

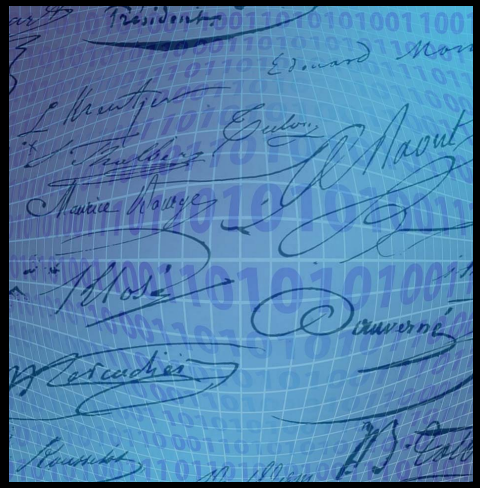
- They're Insufficient
- Can Still Have Too Much Disclosure
- Can Still Have Correlation
- Can Still Have Secondary Use

Bitcoin offers a great example of why pseudonymity isn't enough. With careful, forensic work you can often correlate data in ways that reveal someone's identity. If you have one data block that isn't anonymous or several pseudonymous [SUE-DON-E-MOUS] yet correlatable data blocks, then the anonymity all falls apart. And per slide 7: the amount of data is GROWING.

Data Minimization

Human Rights Problems

- Classic Data Minimization Violates Human Rights
- No Authenticity
  - *RFC 8280 §6.2.17*
- No Integrity
  - *RFC 8280 §6.2.16*
- No Decentralization
  - *RFC 8280 §6.2.13*



Meanwhile, classic data minimization as suggested in RFC 6973 actually breaks some of the guidelines of RFC 8280 such as authenticity, integrity, and decentralization. You lose validating signatures and data integrity, you require a central issuer to do the data minimization, or both.

There are cutting-edge technologies like zero-knowledge proofs, for instance BBS Proofs, but …

We need privacy tech that is simple, well understood, and in production … but still more advanced than 2013.

We need a middle ground.

Deterministic hashed data elision is that middle ground.

Deterministic means that data is always stored in the same way.

Hashed means a cryptographic hash is created for each element of data.

Elision means that data can be removed, and in particular we want data to be removable by any holder of the data, not just the subject or the issuer.

The format we've demonstrated with working code is a Merkel Tree, but there are other options.

In a Merkel Tree, leaves hash the data of their branch, nodes hash the hashes beneath them, and a root hash, which is what you're seeing here, verifies the entire structure.

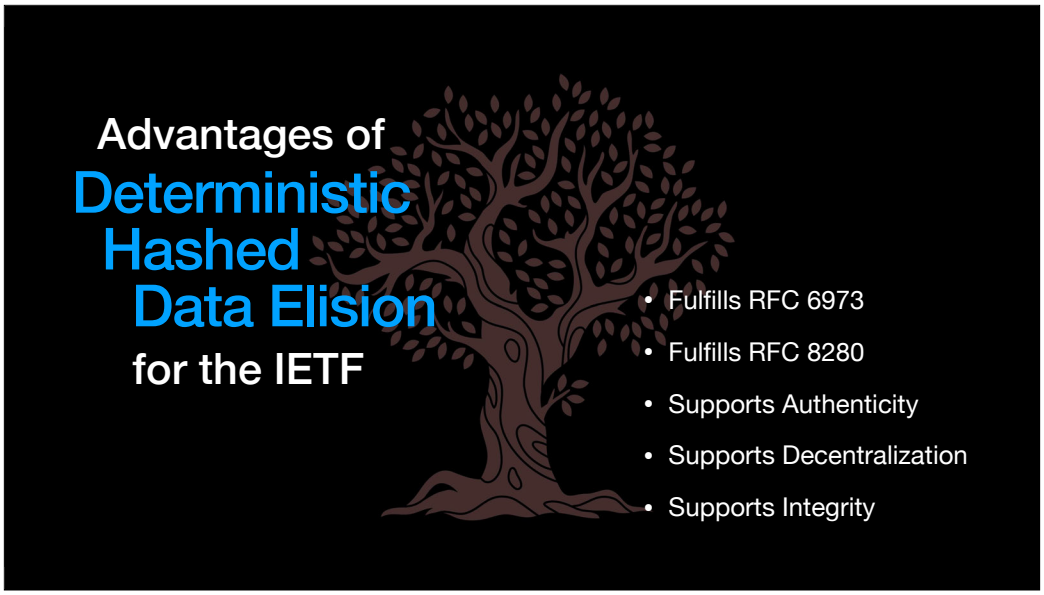
When you elide data, you remove one or more branches, but the hashes remain, validating the Integrity of the data.

Sign the root hash, and you'll also ensure Authenticity, even if everything else is elided!

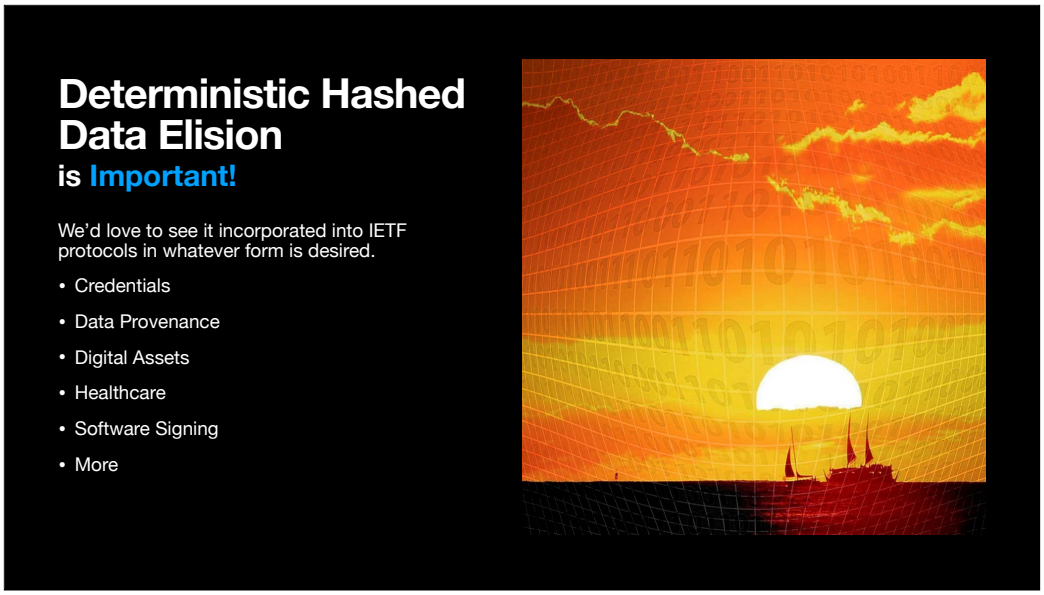
Here's some of the advantages of deterministic hashed data elision. Any holder of the data can choose to elide [E-LIED] data at any time. That supports data minimization, because it's suddenly easy to exclude information. Meanwhile, signatures remain valid even as data is minimized.

But, a hashed data elision system can go much further. Inclusion proofs mean that you can elide parts of a tree and then give proofs leading to elided hashes, which allow for the verification of data even when it's not there! Herd privacy takes the next step. You can publish only a root hash and give out inclusion proofs to data blocks, allowing individuals to reveal that data or not as they see fit.

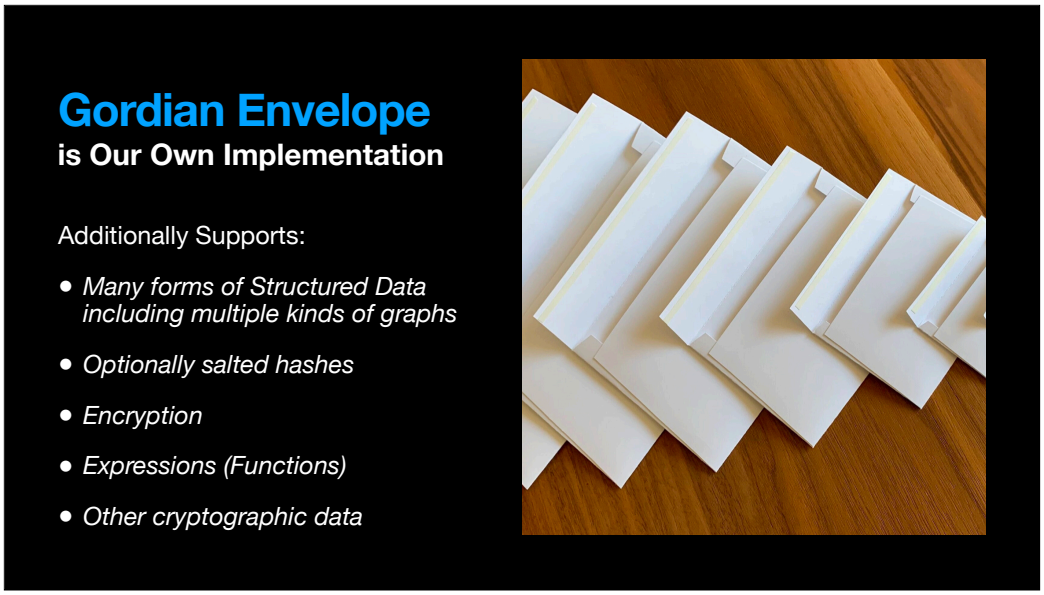
I spoke earlier of the dangers of correlation, but it's actually not all black & white. Sometimes you want to create correlation, sometimes not. Another advantage of Deterministic Hashed Data Elision is that it allows you to match the requirements of your dataset by choosing a hashing method that supports or hinders correlation, as you prefer.



For the IETF, we feel that deterministic hashed data elision fulfills the main thrust of both the privacy and humans-rights RFCs; we feel that it does so in a simple way using mature technology; and we feel that it simultaneously avoids gotchas that could go against the very protections we're trying to create.

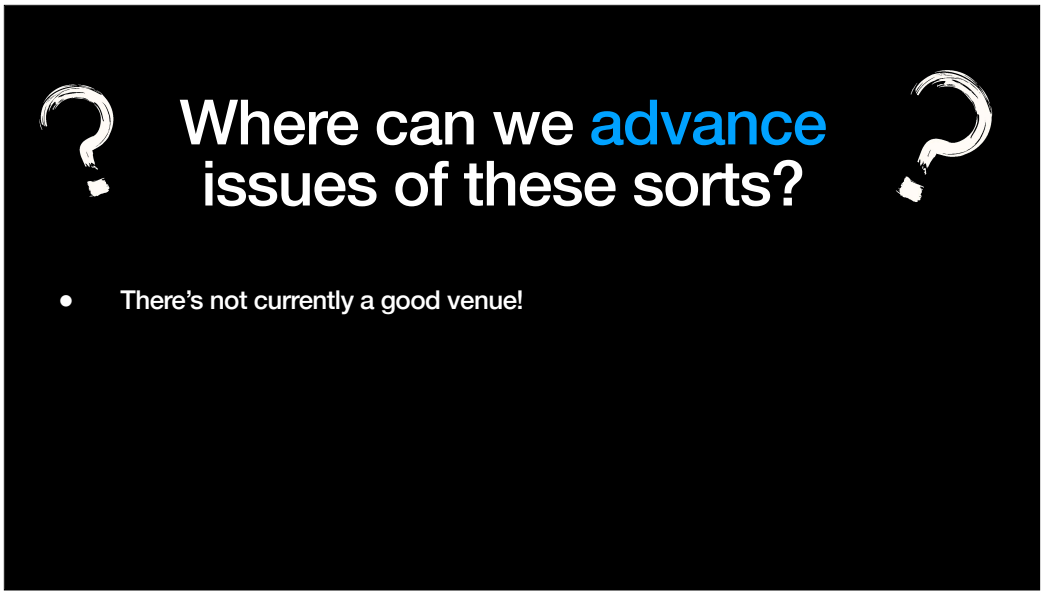


We are tired of privacy being relegated to "considerations" rather than “requirements". It’s much more important than that! As for Deterministic Hashed Data Elision, it can be used by MANY standards. Some people are considering it for credentials, but our own use cases focused on healthcare information and digital assets. We’ve also written use cases for data provenance and software signing. We want to see deterministic hashed data elision widely incorporated.

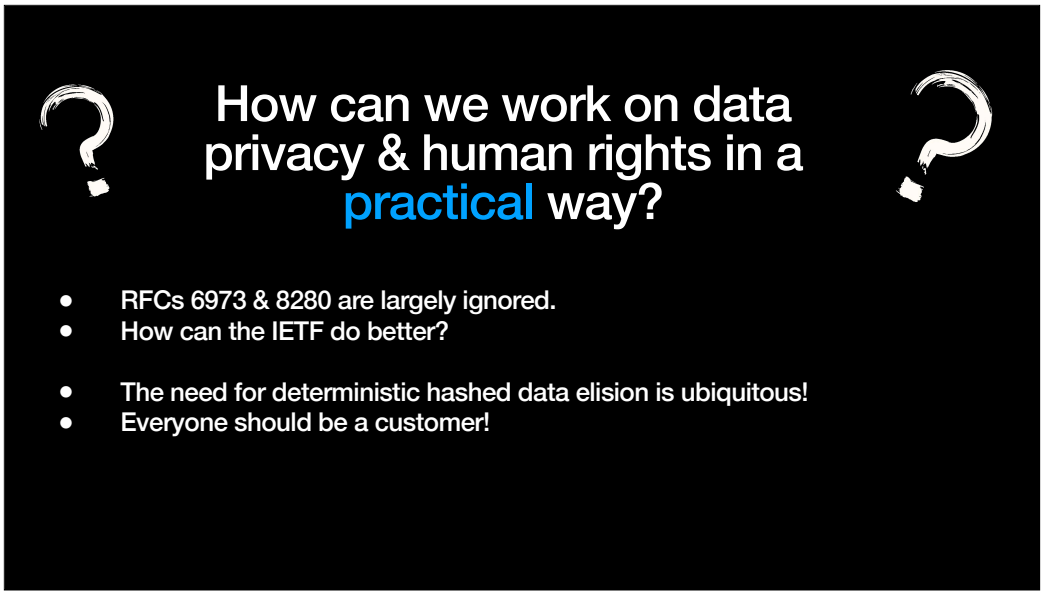


But we'd also love to see support for our own version of Deterministic Hashed Data Elision called Gordian Envelope. It includes all of the fundamentals discussed here, is built on a Merkel Tree, and also includes additional features such as encryption, operational functions, other cryptographic data, and lots more. A working prototype of this system is already deployed in a reference CLI.

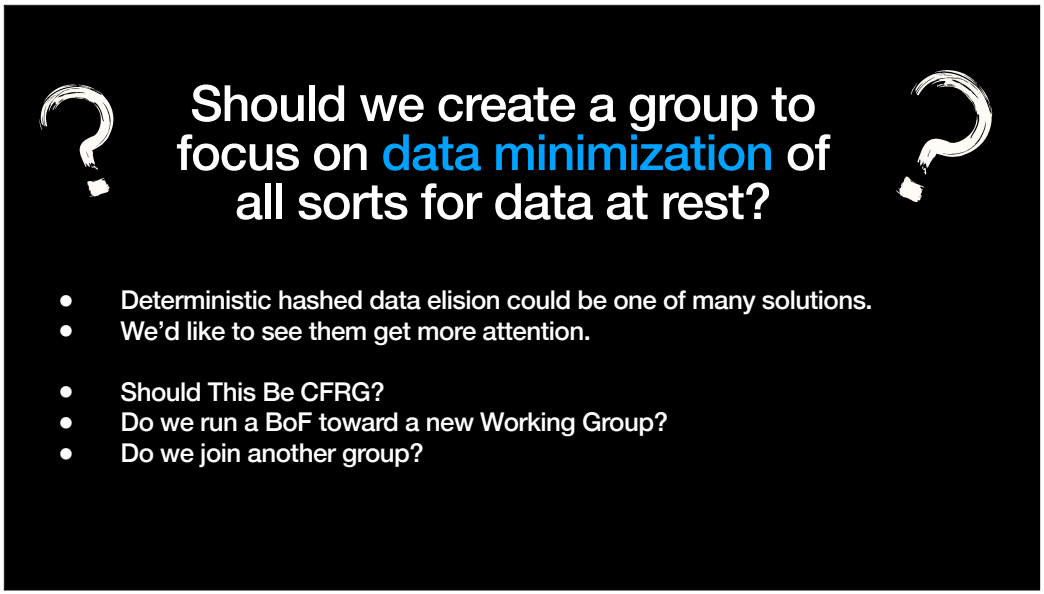
So, here’s where we’d like to get feedback from Dispatch.



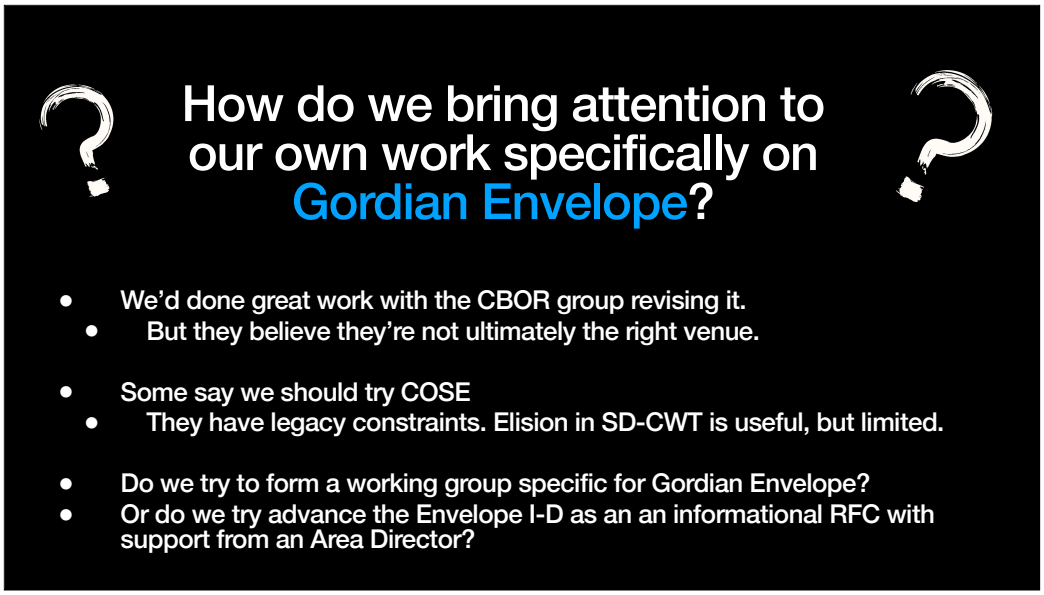
Generally, there doesn't seem to be a good venue to advance any of these ideas right now. But we have three more specific questions.



First, RFCs 6973 & 8280 have been largely ignored. For example, a recent request for changes to a new working-group charter asked that it review its architecture through the lens of these RFCs. The response was: “There appears to be extremely limited support for this”. How can we improve respect for these core IETF values?



Second, should we create a group that focuses on data minimization for data at rest? Do we start with a BoF? Or do we join an existing group?



And third, how do we bring attention to our own work on Gordian Envelope? Do we work with an existing group? Do we form our own working group? Do we advance the Envelope Draft with support from an Area Director?

For More Info

- **draft-appelcline-hashed-elision**
- **draft-mcnally-envelope**

That’s it! More info is available in our two drafts. Hashed-elision is this full problem statement, envelope is the specification of our model for deterministic hashed data elision.

Thank you!

Here's some more info on us. I'm Shannon Appelcline. Christopher Allen is the Executive Director & Principal Architect of Blockchain Commons. That QR links to some more info on Gordian Envelope.

Any questions?