Redactable Distributed Ledger (REd Ledger): A Clinical Research Use case

Rick Kuhn and Joanna DeFranco

National Institute of Standards and Technology

NIST INFORMATION TECHNOLOGY LABORATORY

DISCLAIMER: The views and opinions expressed in this presentation are solely those of the author/presenter and do not necessarily represent any policy or position of HIMSS.

Key Points

- Blockchain has valuable properties, but conflicts with privacy and exception management – “immutable” - deletion impossible

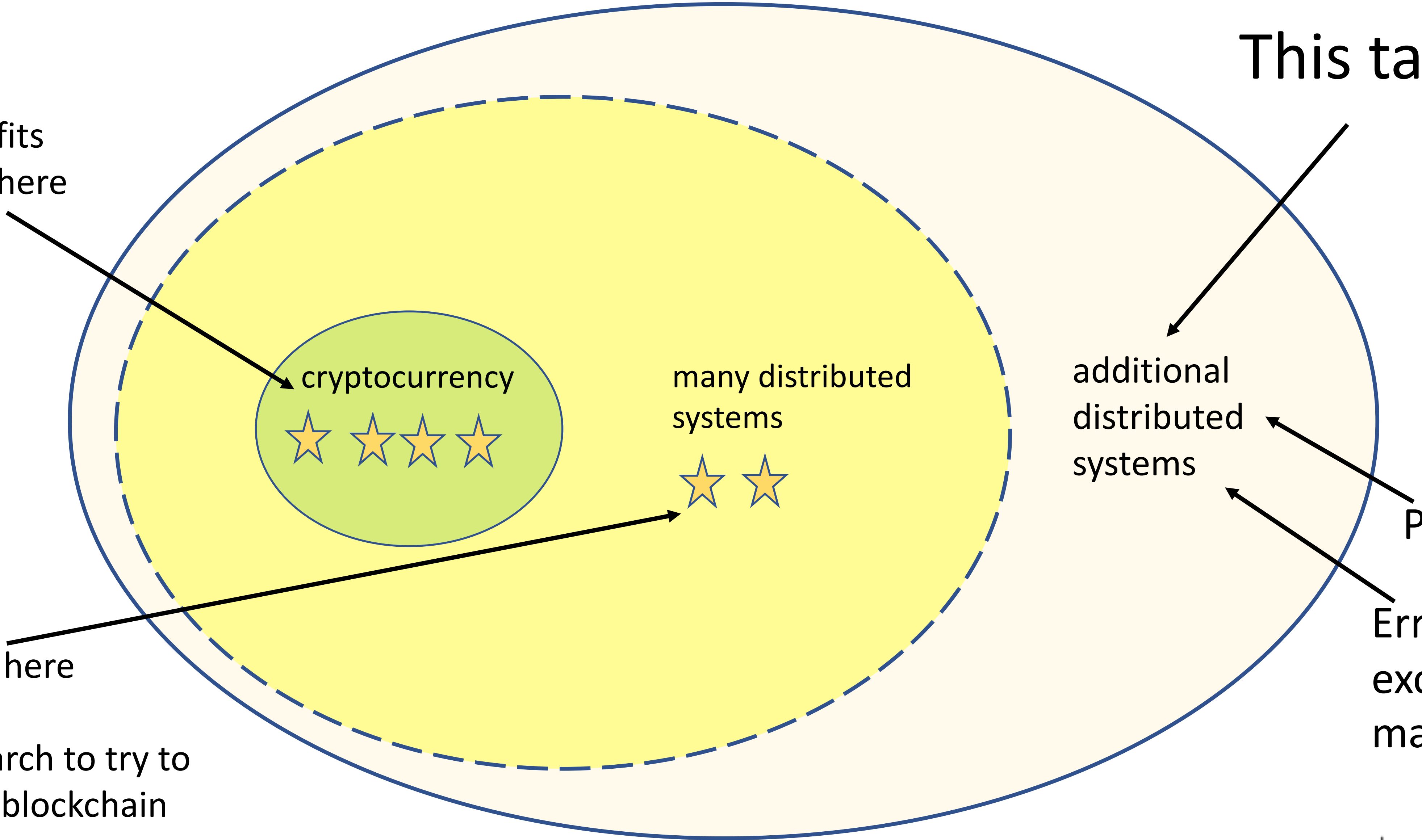
  ➡ Sometimes we don’t need blockchain,
  just some blockchain features

- Data structure called *blockmatrix* provides
  distributed trust, integrity protection of blockchain,
  but allows controlled edits for privacy or corrections

- Drop-in compatibility for Hyperledger Fabric applications

  ➡ Released and available

Market, range of applications for DLT



This talk here

Blockchain fits pretty well here

cryptocurrency ☆☆☆☆

many distributed systems ☆☆

additional distributed systems

Privacy

Error/ exception management

Not as well here

Much research to try to get around blockchain properties

Why use redactable DLT for privacy?

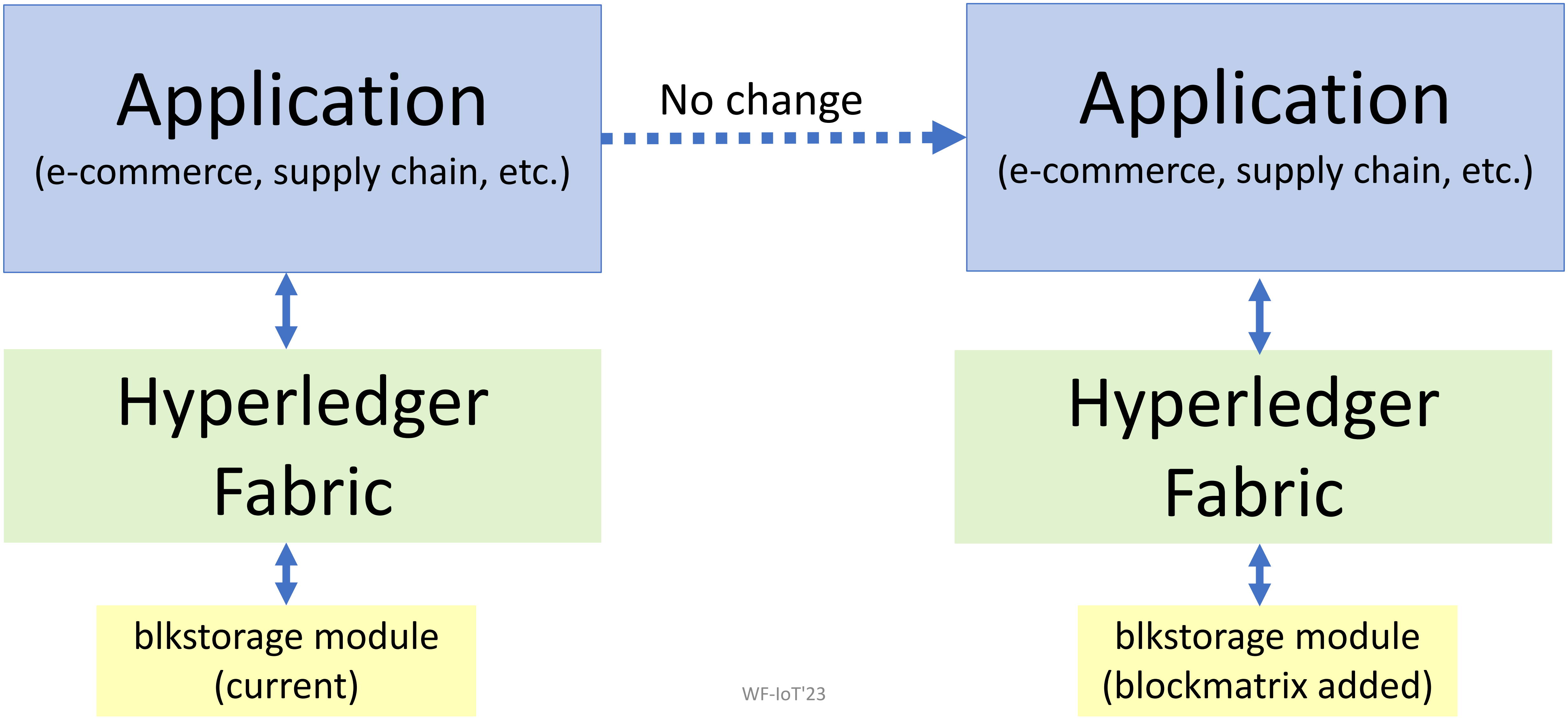
- Permanence/immutability conflicts with ‘right to erasure’ privacy regulations

- Privacy rules such as European Union General Data Protection Regulation (GDPR) require that all information related to a particular person can be deleted at that person's request
  - *any personal* data "concerning an identified or identifiable natural person"
  - includes pseudo-anonymized data linkable to person
  - US states adopting similar privacy rules, including California and Virginia

What's been tried to solve blockchain/privacy conflict?

- Don't put personal data on blockchain – but pseudo-anonymized data are still considered personal; Financial transactions are obviously personal data
- Encrypt data and destroy key to delete – but data must be secure for decades (e.g., DES replaced in only 17 years)
- Chameleon hash function – non-standard cryptography
- Off-chain storage of sensitive data – what if on-chain index to off-chain data is also sensitive?

Compatible with Hyperledger applications



Application
(e-commerce, supply chain, etc.)

No change

Application
(e-commerce, supply chain, etc.)

Hyperledger
Fabric

Hyperledger
Fabric

blkstorage module
(current)

blkstorage module
(blockmatrix added)

How does this work?

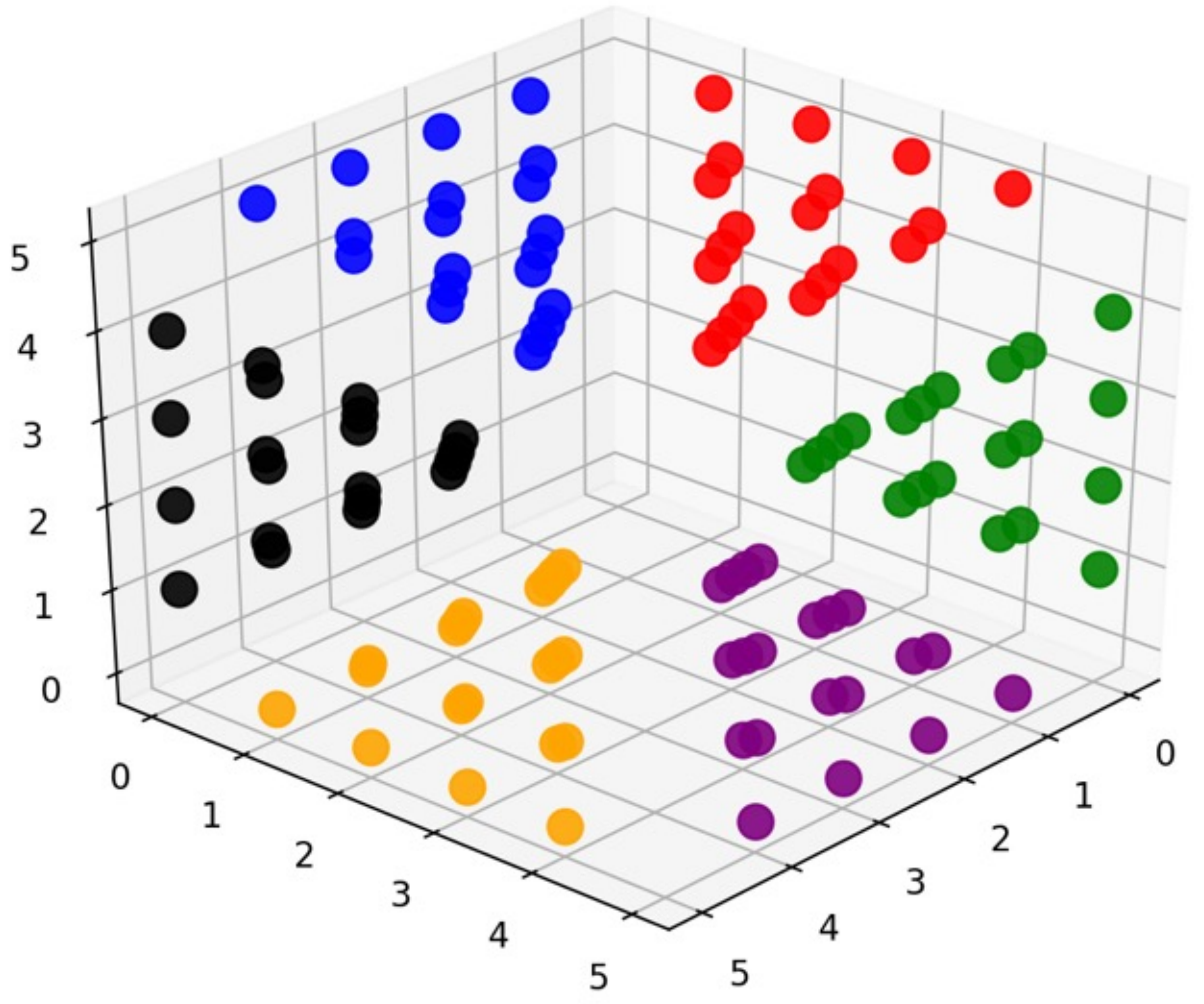
- Suppose we want to delete block 12

- disrupts the hash values of $H_{3,-}$ for row 3 and $H_{-,2}$ and column 2
- blocks of row 3 are included in the hashes for columns 0, 1, 3, and 4
- blocks of column 2 are included in the hashes for rows 0, 1, 2, and 4

	0	1	2	3	4	
0	□	1	3	7	13	$H_{0,-}$
1	2	□	5	9	15	$H_{1,-}$
2	4	6	□	11	17	$H_{2,-}$
3	8	10	12	□	19	$H_{3,-}$
4	14	16	18	20	□	$H_{4,-}$
	$H_{-,0}$	$H_{-,1}$	$H_{-,2}$	$H_{-,3}$	$H_{-,4}$	etc.

Structure can be extended to multiple dimensions

- Block dispersal for 3 dimensions
- Location in sectors 0..5 according to *b* mod 6 for block *b*



DBM Summary - why use this data structure?

Again, many blockchain applications <u>don’t need blockchain</u>, <u>just some features</u>

Blockchain: distributed trust, integrity protection, <u>immutability</u>
Data block matrix: distributed trust, integrity protection, <u>editable for privacy, error correction</u>

Enlarge the market for distributed ledger

- Solve the conflict between blockchain and privacy regulations
- Allow for exception management

Replace network communication with local data

- You can obviously do this with conventional database functions, but
- New data structure adds integrity checks as in blockchain

Easy-to-use component for distributed database design

Clinical Research Use Case

Overview: Secure Federated Data Sharing System (SFDS)

SFDS value in performing clinical research

Problem

How can organizations securely share data?

- The ability to share database resources among collaborating organizations is highly desirable – this is especially true in the performance of clinical research.

- However, challenges persist regarding **interoperability** in the exchange of resources among organizations and **preservation of organization's distinct protection policies**.

- Hard for users in **different organizations** to share DBMS data. Because the data is
  - from different systems,
  - in different formats,
  - organized under different schemas
  - protected under the host's access control policies.

Solution

Exchange attributes not data

- A standard means of providing **policy-preserving access** to the data where it currently **resides**, rather than its exchange, or centralized storage.

- **Transparent** to the otherwise normal business operations of participating organizations.

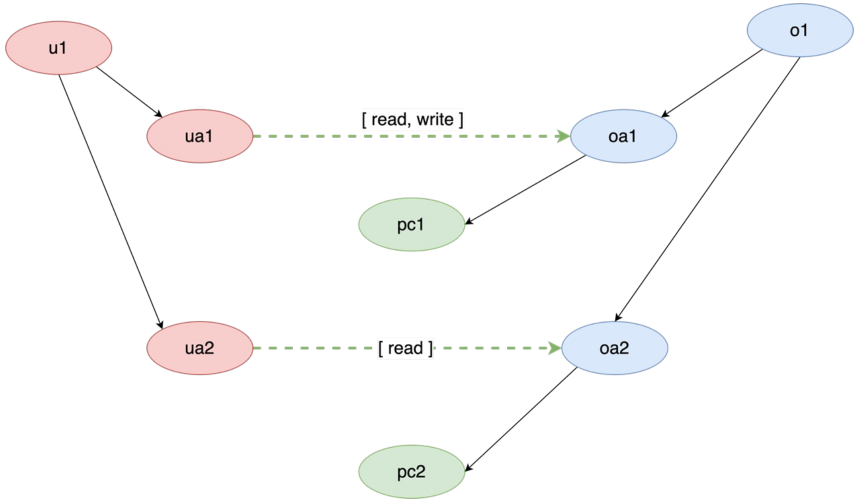
- Accomplished using two NIST developed technologies:

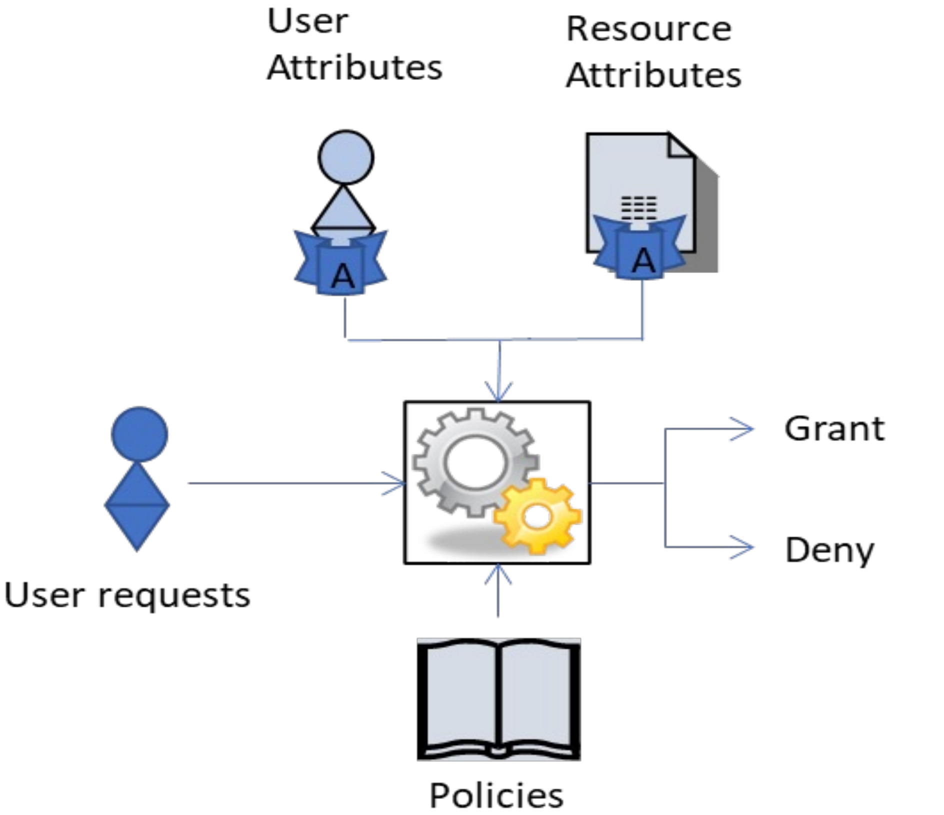
  - **Data Block Matrix (DBM) –** Verify user's attributes across a federation of organizations
  - **Next generation Database Access Control (NDAC) –** Control access to SQL databases with cell level access control

- Through consent, previously unknown users are onboarded into local **NDAC** systems using their **DBM** validated attributes, allowing them policy preserving access to local database resources.

Attribute-based Access Control (ABAC)

- Fundamental to our data sharing strategy





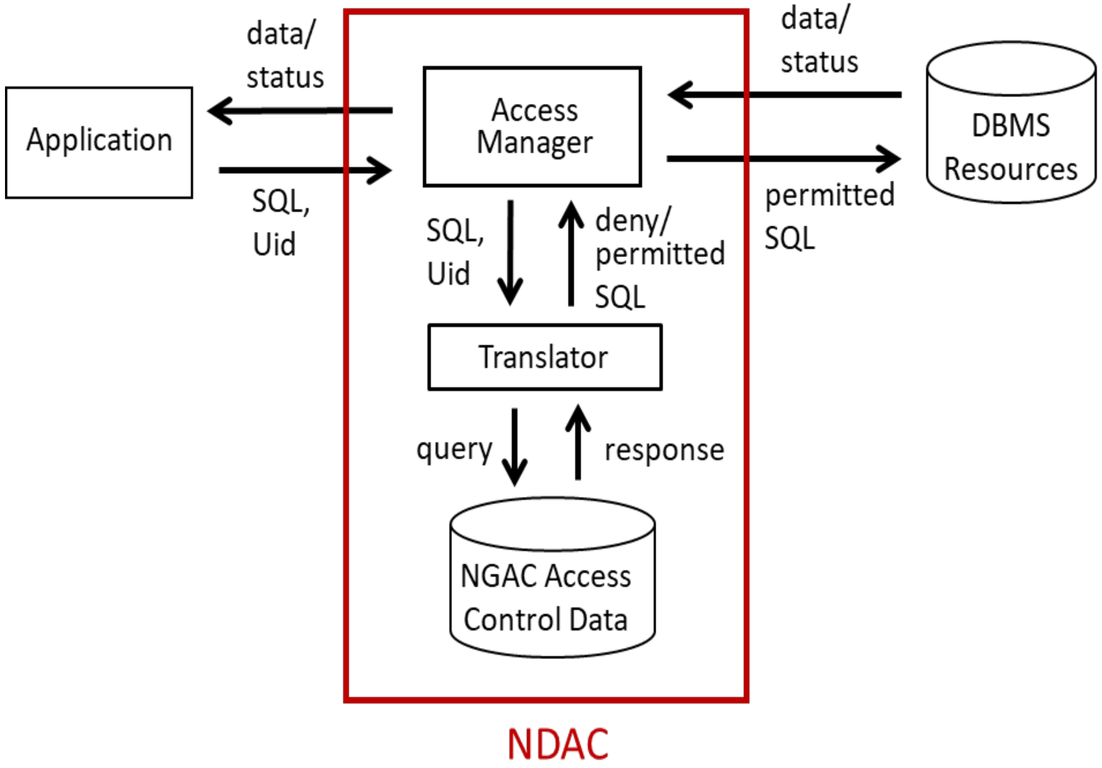
Vincent Hu et al., NIST special publication 800-162, (2014) ***Guide to attribute-based access control (abac) definition and considerations.***

Next Generation Database Access Control (NDAC)

- Provides a **universal access control layer** between applications and DBMSs, following a standardized ABAC model (NGAC*) to translate SQL statements into permitted SQL.

- DBMS and application agnostic.

- Enforce policies at a granularity not typically available to DBMSs.



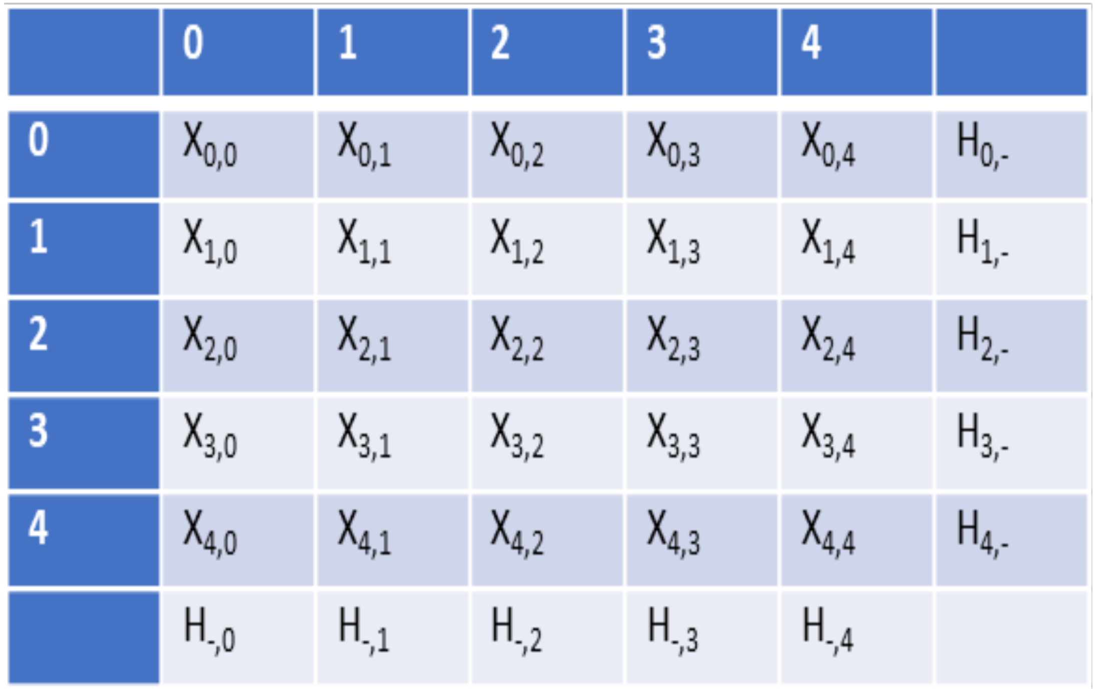
* An ANSI/INCITS Standard

D. Ferraiolo et al., "Imposing Fine-grain Next Generation Access Control over Database Queries," ABAC'17, Scottsdale, AZ, 2017

**Next Generation Access Control (NGAC) is an ANSI/INCITS standard*

Data Block Matrix (DBM): globally manage attributes

- A new type of distributed ledger, with the hashed data integrity protection of a blockchain, but with the additional ability to edit or delete data.
- Provides a means for storing, managing, and sharing attributes of users in the federation.
- Ideal for SFDS
    - user-attribute assignments need to be altered over time, and
    - accommodate privacy regulations: "deletion of PII when no longer needed".



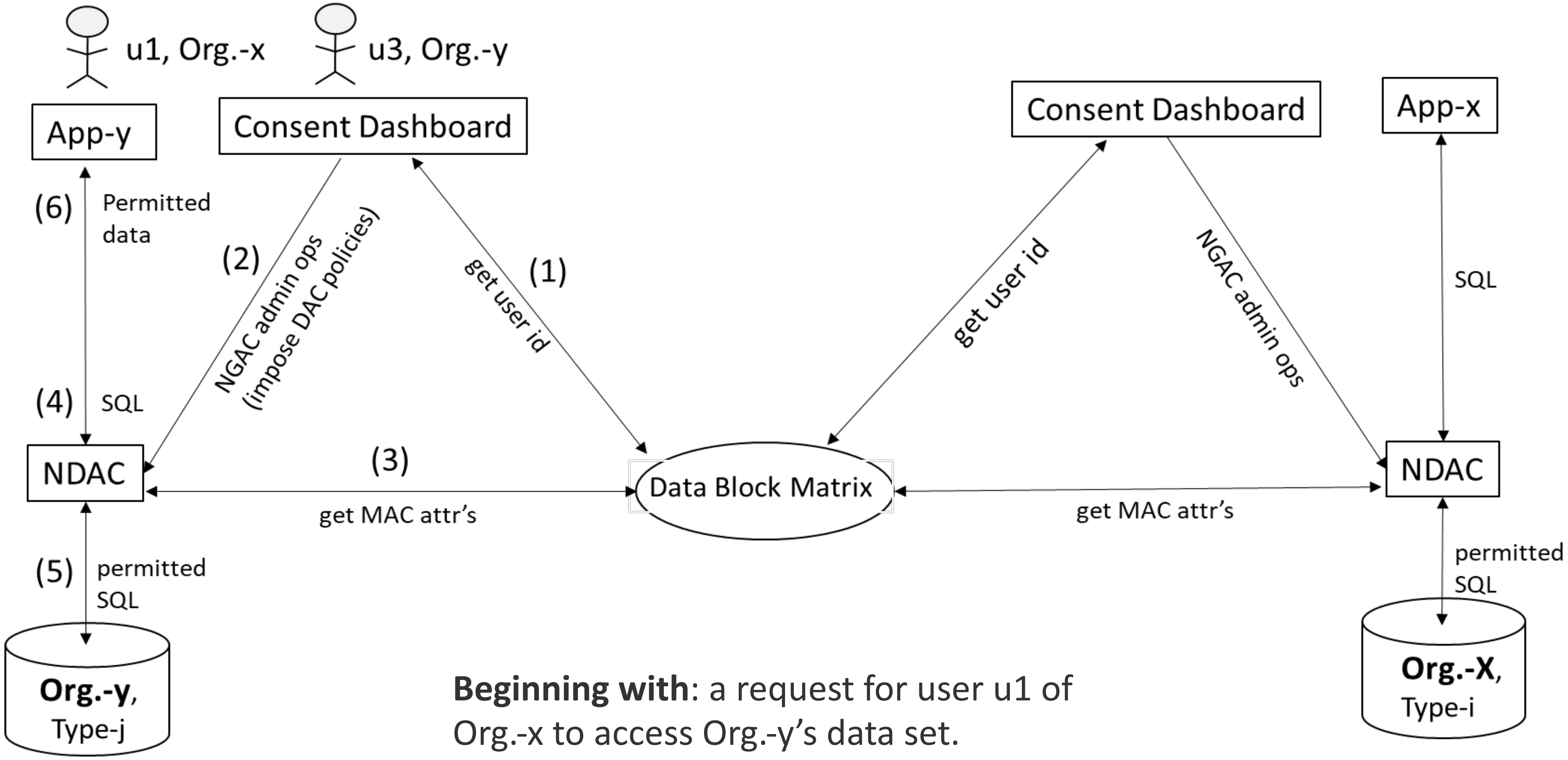
	0	1	2	3	4	
0	$X_{0,0}$	$X_{0,1}$	$X_{0,2}$	$X_{0,3}$	$X_{0,4}$	$H_{0,-}$
1	$X_{1,0}$	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$H_{1,-}$
2	$X_{2,0}$	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	$H_{2,-}$
3	$X_{3,0}$	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$	$H_{3,-}$
4	$X_{4,0}$	$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	$X_{4,4}$	$H_{4,-}$
	$H_{-,0}$	$H_{-,1}$	$H_{-,2}$	$H_{-,3}$	$H_{-,4}$	

R. Kuhn et al., "Rethinking Distributed Ledger Technology," Computer, vol. 52, no. 2, 2019

DAC and MAC

- Two types of policies come into play: Discretionary and Mandatory Access Control (DAC and MAC).
- DAC provide users with capabilities to grant or prohibit other users' access to resources that are placed under their control.
    - Access is based on the accessing user's name or id (one-to-one)
- MAC policies impose non-discretionary rules on users when accessing resources.
    - Enforcement of mandated policies or regulations
    - Access is based on a user roles or other types of attributes (many-to-one)
- NGAC can impose policy combinations. e.g.,
    - DAC and MAC policies must hold to access data resources

Operational Sharing of DBMS Resources



Beginning with: a request for user u1 of Org.-x to access Org.-y's data set.

Important!

1

Cybersecurity and Computer Science Experts

Our group is focused on the security of systems and applications, and we explore the implementation of security technology in **other domains**, such as health.

2

Flexibility and Confidence to Provide Access

The technology we are demonstrating is to **facilitate access to** information without moving large volumes of data and with sufficient confidence that the information shared is what is intended.

3

Variety of Situations, Technologies and Complexities

We can demonstrate enough of a scenario to understand what is possible given the problem, but it is **not a comprehensive representation of all possibilities and capabilities.**

Clinical Data Sharing Challenge

Institution A

Has a very large dataset with **very little restriction** on sharing with B.

Institution C

This institution has a **more restrictive posture** with sharing due to the nature of some of the population served.

Institution B

Has significant **research expertise** and resources.

The Participants

- Technologists
- Administrators
- Researchers
- Clinicians
- Subject Matter Experts
- Governance Bodies
- Assistants/Support
- Students

✓ *Likely to Vary Across Institutions*

✓ *Some May Serve Multiple Roles*

✓ *Participants May Change Over Time*

Barriers to Data Sharing

- Credit and recognition
- Potential misuse or misinterpretation of data
- Lack of resources
- Loss of control
- Socio-cultural factors and ethical and legal barriers

Devriendt T, Borry P, Shabani M (2021)
***Factors that influence data sharing through data sharing platforms: A qualitative study on the views and experiences of cohort holders and platform developers**.*

Establishing Policies

Institution A

- Planned Inpatient Visits
- Log Access for Audit
- Limit to 1 year

```
SELECT * FROM
ENCOUNTERS;
```

Institution B

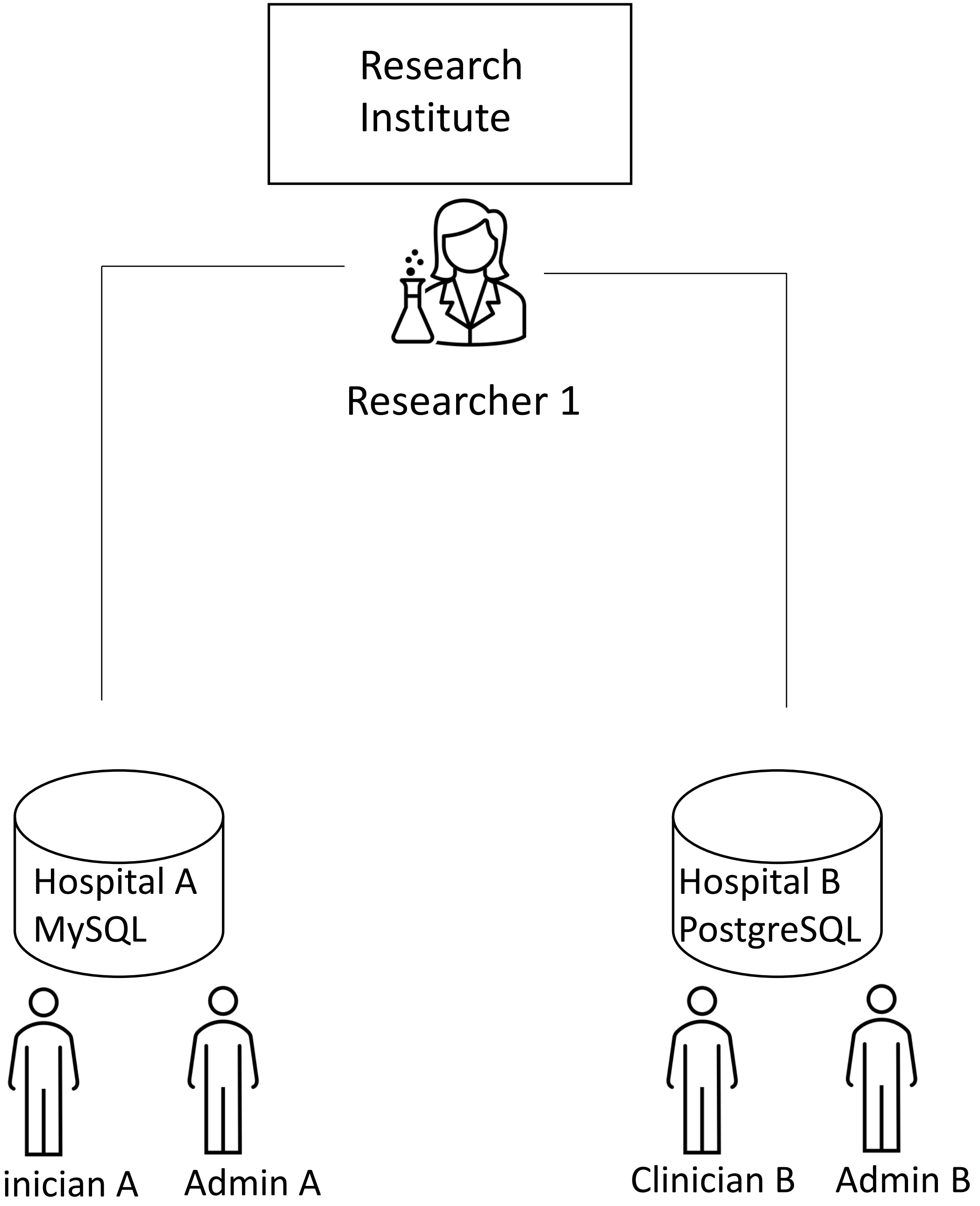
- Planned Inpatient Visits
- Log Access for Audit
- Limit to 1 year
- Withhold Mental Health Codes
- Redact Clinician Information

Use Case Scenario

- Principal Investigator at an academic research institution is studying the variables and indicators of patients at risk of sepsis. This research will support more accurate prediction of the risk of sepsis and encourage early intervention in vulnerable populations.
- Require data access from several different hospitals to effectively evaluate patient variables and indicators of sepsis.
- Patient data protection concerns need to be addressed.

Su, Y., Guo, C., Zhou, S. et al.
Eur J Med Res 27, 294 (2022)
Early predicting 30-day mortality in sepsis in MIMIC-III by an artificial neural networks model.

- **Researcher 1** wants access to clinical data at **Hospital A** and **Hospital B** to perform an analysis of the variables and indicators of patients at risk of **sepsis**.
- Hospital A and Hospital B restrict access to patient **PII** and data related to **mental health** (MAC policy).
- Hospital B considers substance abuse related to mental health, while Hospital A does not.
- Both hospitals will create similar DAC policies which grant Researcher 1 access to **patient identity** and **financial information** for only sepsis patients.



Metrics and Benchmarks

- NDAC
  - **Query rewrite:** ~3 seconds per 1 million rows
- DBM
  - **Write (excluding delete):** Comparable to regular fabric. The only difference is writing the bytes to a key value database instead of a file.
  - **Delete:** Linear to the number of blocks updated by the delete operation.
  - **Read:** Comparable to regular fabric. The only difference is reading the bytes from a key value database instead of a file.

Conclusion and Summary

- The **ability to share database resources** among collaborating organizations is highly desirable.
- However, challenges persist regarding **interoperability** in the exchange of resources between organizations and the preservation of local access policies.
- SFDS provides a **generic data sharing infrastructure** that effectively and securely achieves data sharing objectives.
- It is completely **transparent** to the otherwise normal business operations of participating organizations. It requires **no changes** to DBMSs, or existing methods of authenticating and authorizing local user access to local resources.
- This ease of deployment, granularity of control and its efficiency make this new infrastructure solution practical for **meeting the data sharing and protection objectives** of the clinical research community.

Future Work

- The current focus of SFDS is to allow sharing of database resources. Included in our plans is to extend the infrastructure to allow controlled access to non-structured data such as **files**.

- We would like to organize a **full-scale pilot** study involving institutions that house medical information and/or conduct clinical research.

- Our goal is to transition SFDS from a research project to operational use.

More Information

Foundation:

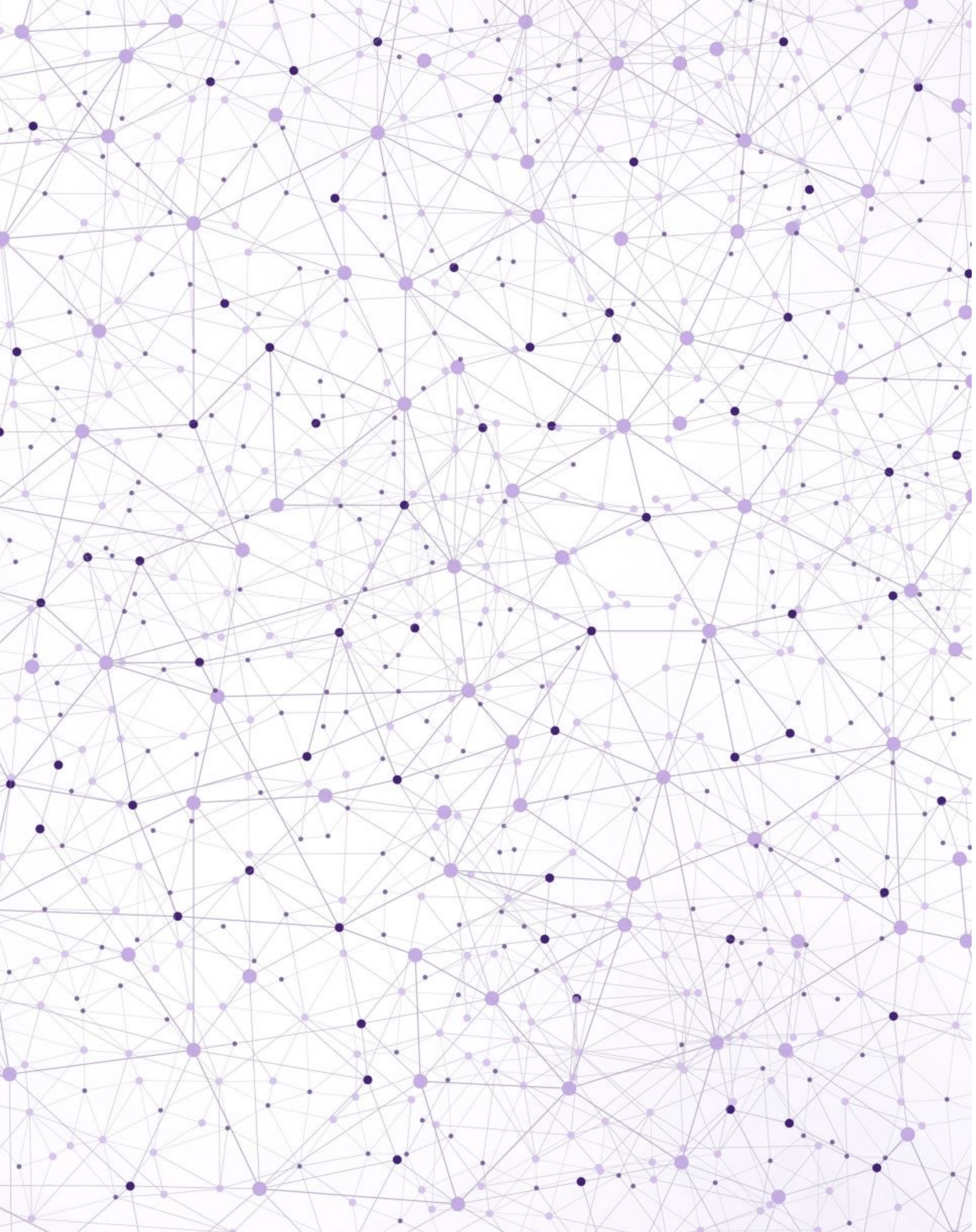
- Kuhn, R., Yaga, D. and Voas, J., 2019. Rethinking Distributed Ledger Technology. *Computer*, *52*(2), pp.68-72.
- Kuhn, D. R. (2018). A Data Structure for Integrity Protection with Erasure Capability. https://csrc.nist.gov/publications/detail/white-paper/2022/05/20/data-structure-for-integrity-protection-with-erasure-capability/final

Applications:

- Roberts, J. D., Defranco, J. F., & Kuhn, D. R. (2023). Data Block Matrix and Hyperledger Implementation: Extending Distributed Ledger Technology for Privacy Requirements. *ACM Distributed Ledger Technologies: Research and Practice*, 2(2), 1-11.
- D. R. Kuhn, J. D. Roberts, D. Ferraiolo and J. DeFranco, "A Distributed Ledger Technology Design using Hyperledger Fabric and a Clinical Trial Use Case," *2022 IEEE 29th Annual Software Technology Conference (STC)*, Gaithersburg, MD, USA, 2022, pp. 168-173, doi: 10.1109/STC55697.2022.00031.

Project sites with links to source code and publications

- https://csrc.nist.gov/Projects/enhanced-distributed-ledger-technology
- https://csrc.nist.gov/projects/redactable-distributed-ledger



Contacts

Joanna DeFranco -	joanna.defranco@nist.gov jfd104@psu.edu
Rick Kuhn -	kuhn@nist.gov

Acknowledgment

- David Ferraiolo, Supervisory Computer Scientist
- Joshua Roberts, Computer Scientist
- D. Chris Compton, Cybersecurity Specialist and Clinical Informatics Advisor and