Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 19: Secure Rule Mining, Blockchain, and Electronic Informed Consent

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)

Assistant Professor of Biomedical Engineering

ShanghaiTech University

May 23, 2025

Learning Objectives of This Lecture

After this lecture, students should be able to:

- Know the concept of secure rule mining
- Know the concept of Blockchain
- Know the concept of eConsent

Introduction to Secure Rule Mining

- How can we learn information from multiple locations if they don’t want to – or can’t - share their information?

- Computation over disparate locations’ *private* datasets

- We’ll focus on *ring* computations

Outline

- Association Rule Mining
- Distributed Rule Mining
- Secure Rule Mining - Semi-Honest Setting
- Secure Rule Mining – Malicious (almost)

Items & Associations

- *Items* $I = \{i_1, i_2, \ldots, i_n\}$
- Database of transactions DB
- Transaction $T \subseteq I$
- Itemset $X \subseteq I$
  - *k-itemset*: Itemset with *k* items
- Transaction includes itemset, when $X \subseteq T$
- Association Rule: Implication of the form

  $X \Rightarrow Y$, such that $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varnothing$

T	Items
1	Soy milk, lettuce
2	Lettuce, diaper, wine, beet
3	Soy milk, diaper, wine, orange
4	Lettuce, soy milk, diaper, wine,
5	Lettuce, soy milk, diaper, orange

Example

- Items: {*ICD-9 codes*}
- Database of transactions T: subset of ICD-9 codes
- Locations: {*Set of Hospitals*}
- Itemset: Subset of codes
- Association Rule: Conditional probabilities of rule implications
    - *Warfarin* $\Rightarrow$ *Stroke*

- How can we learn the statistics of $X \Rightarrow Y$ from all locations without revealing what any of them had?

Statistical Foundations

- Rule $X \Rightarrow Y$ has <u>support</u> s, when $X \cap Y$ in s% of transactions

- Rule $X \Rightarrow Y$ has <u>confidence</u> c, when c% of transactions that contain X also contain Y

- <u>Strong rules</u> satisfy both:
  - a minimum support *minsup*
  - a minimum confidence *minconf*

The Apriori Algorithm

- Uses a *level-wise* search to find all strong rules
- Uses *k*-itemsets to probe for (*k*+1)-itemsets
  - First, find frequent 1-itemsets, or L_1
  - Use $L_1 \cup L_1$ to find L_2
  - Use $L_2 \cup L_2$ to find L_3
  - …

R. Agrawal and R. Srikant. Fast algorithms for mining association rules. Proc VLDB. 1994: 487-499.

Apriori

- C_k = candidate k-itemsets
- L_k = frequent k-itemsets
- L_1 = 1-itemsets satisfy *minsup*
- For (k = 1, L_k != ∅; k++)
  - C_{k+1} = candidates generated from L_k (i.e., L_k join L_k)
  - For each transaction in database
    - Increment the count of all candidates in C_{k+1} in the transaction
    - Prune k-itemsets that fail to satisfy *minsup*
- For each itemset
  - generate all subsets and test if rules fail to satisfy *minconf*

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	
B	
C	
D	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	
C	
D	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Itemset	Sup
A, B	
A, C	
A, D	
B, C	
B, D	
C, D	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Itemset	Sup
A, B	10%
A, C	
A, D	
B, C	
B, D	
C, D	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Itemset	Sup
A, B	10%
A, C	40%
A, D	50%
B, C	20%
B, D	10%
C, D	50%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Itemset	Sup
A, B	10%
A, C	40%
A, D	50%
B, C	20%
B, D	10%
C, D	50%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Itemset	Sup
A, C	40%
A, D	50%
C, D	50%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup
$A \Rightarrow C$	
$A \Rightarrow D$	
$C \Rightarrow D$	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup
$A \Rightarrow C$	40%
$A \Rightarrow D$	50%
$C \Rightarrow D$	50%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup
A ⇒ C	40%
C ⇒ A	
A ⇒ D	50%
D ⇒ A	
C ⇒ D	50%
D ⇒ C	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup
$A \Rightarrow C$	40%
$C \Rightarrow A$	40%
$A \Rightarrow D$	50%
$D \Rightarrow A$	50%
$C \Rightarrow D$	50%
$D \Rightarrow C$	50%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
A ⇒ C	40%	
C ⇒ A	40%	
A ⇒ D	50%	
D ⇒ A	50%	
C ⇒ D	50%	
D ⇒ C	50%	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
A $\Rightarrow$ C	40%	66%
C $\Rightarrow$ A	40%	
A $\Rightarrow$ D	50%	
D $\Rightarrow$ A	50%	
C $\Rightarrow$ D	50%	
D $\Rightarrow$ C	50%	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
A ⇒ C	40%	66%
C ⇒ A	40%	57%
A ⇒ D	50%	
D ⇒ A	50%	
C ⇒ D	50%	
D ⇒ C	50%	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
A $\Rightarrow$ C	40%	66%
C $\Rightarrow$ A	40%	57%
A $\Rightarrow$ D	50%	83%
D $\Rightarrow$ A	50%	
C $\Rightarrow$ D	50%	
D $\Rightarrow$ C	50%	

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
$A \Rightarrow C$	40%	66%
$C \Rightarrow A$	40%	57%
$A \Rightarrow D$	50%	83%
$D \Rightarrow A$	50%	71%
$C \Rightarrow D$	50%	71%
$D \Rightarrow C$	50%	71%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
$A \Rightarrow C$	40%	66%
$C \Rightarrow A$	40%	57%
$A \Rightarrow D$	50%	83%
$D \Rightarrow A$	50%	71%
$C \Rightarrow D$	50%	71%
$D \Rightarrow C$	50%	71%

Minimum Support = 30%
Minimum Confidence = 60%

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
$A \Rightarrow C$	40%	66%
$A \Rightarrow D$	50%	83%
$D \Rightarrow A$	50%	71%
$C \Rightarrow D$	50%	71%
$D \Rightarrow C$	50%	71%

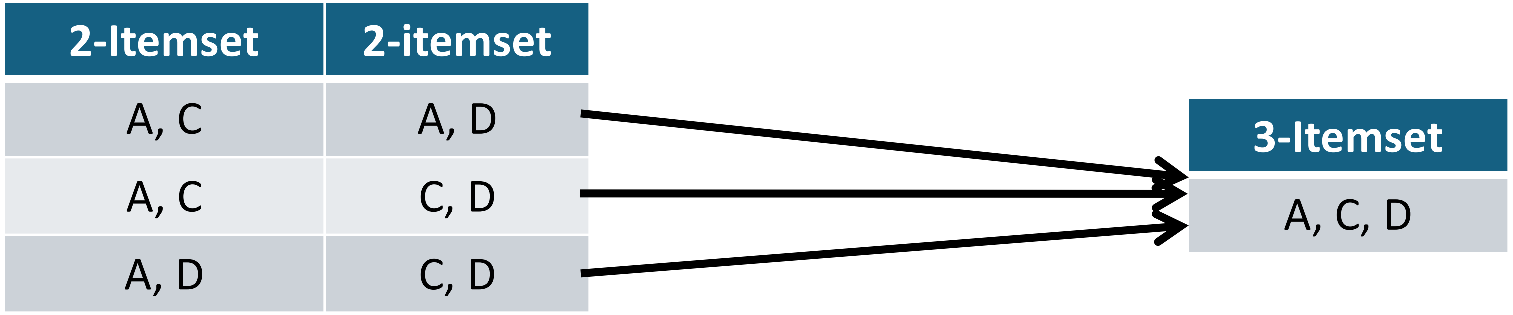
So, what is C_3?

Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Itemset	Sup
A	60%
B	30%
C	70%
D	70%

Rule	Sup	Conf
$A \Rightarrow C$	40%	66%
$A \Rightarrow D$	50%	83%
$D \Rightarrow A$	50%	71%
$C \Rightarrow D$	50%	71%
$D \Rightarrow C$	50%	71%

So, what is C_3?



2-Itemset	2-itemset
A, C	A, D
A, C	C, D
A, D	C, D

3-Itemset
A, C, D

This is why some rule mining algorithms (beyond Apriori), look forward several levels, for planning purposes.

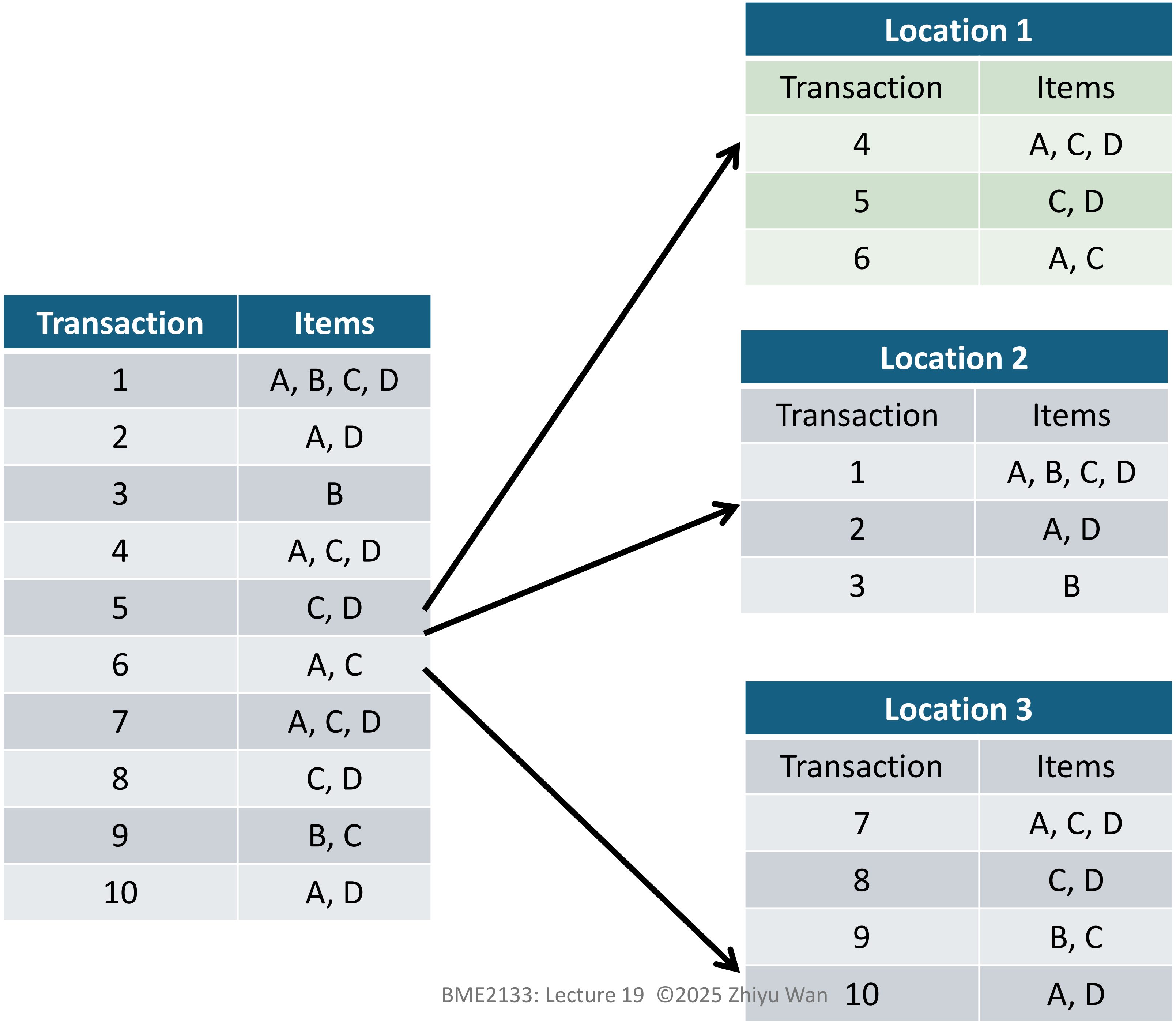
Outline

- Association Rule Mining
- Distributed Rule Mining
- Secure Rule Mining - Semi-Honest Setting
- Secure Rule Mining - Malicious (almost)

Distributed

- *DB* is “horizontally” partitioned among *n* sites

$$DB = DB_1 \cup DB_2 \cup \ldots \cup DB_n$$



Transaction	Items
1	A, B, C, D
2	A, D
3	B
4	A, C, D
5	C, D
6	A, C
7	A, C, D
8	C, D
9	B, C
10	A, D

Location 1

Transaction	Items
4	A, C, D
5	C, D
6	A, C

Location 2

Transaction	Items
1	A, B, C, D
2	A, D
3	B

Location 3

Transaction	Items
7	A, C, D
8	C, D
9	B, C
10	A, D

Distributed

- *DB* is "horizontally" partitioned among *n* sites

$$DB = DB_1 \cup DB_2 \cup \ldots \cup DB_n$$

- *Sup*$_i$ (*X*): Local support of Rule *X* at location L_i
- *Global support* $Sup(X) = \sum_{i=1}^{n} Sup_i(X)$
- Given *minsup = s%*, *X* is globally supported if

$$Sup(X) \geq s\left(\sum_{i=1}^{n} |DB_i|\right)$$

- Global confidence of a rule X $\Rightarrow$Y: $Sup(X \cap Y) \big/ Sup(X)$

Useful Tidbit of Knowledge

- Notice, if Sup(*itemset*) > *minsup* globally, then
  - Sup(*itemset*) > *minsup*
    must be true for <u>at least one</u> of the locations

New Goal!

- *Large itemsets L^k*: Set of all *k*-itemsets that are globally supported
- *Locally large itemsets LL_i^k :* Set of all *k*-itemsets that are supported at location *i*
- *Globally-local itemsets GL_i^k*: Set of large itemsets that are local itemsets

$$GL_i^k = L^k \cap LL_i^k$$

- Distributed Mining Goal:
  - Discover the L^k, for all $k > 1$, and
  - The support counts for L^k , which leads to
  - The global association rules that satisfy *minsup* and *minconf*

Distributed Mining

- Fast Distributed Mining of Association Rules (FDM)*
- Four basic steps
    1. Candidate Set Generation
    2. Local Pruning
    3. Support Count Exchange
    4. Broadcast Mining Results

D. Cheung, et al. A Fast distributed algorithm for mining association rules. Proc PDIS. 1996: 31-42.

FDM: Candidate Set Generation

1. Generate candidate sets CG_i^k based on $GL_i^{k\text{-}1}$

   Globally-local itemsets

   Can find itemsets supported by location *i* in the (*k*-1)th iteration using the standard *Apriori* algorithm

FDM: Local Pruning

For each $X \in CG_i^k$ Candidate-local itemsets

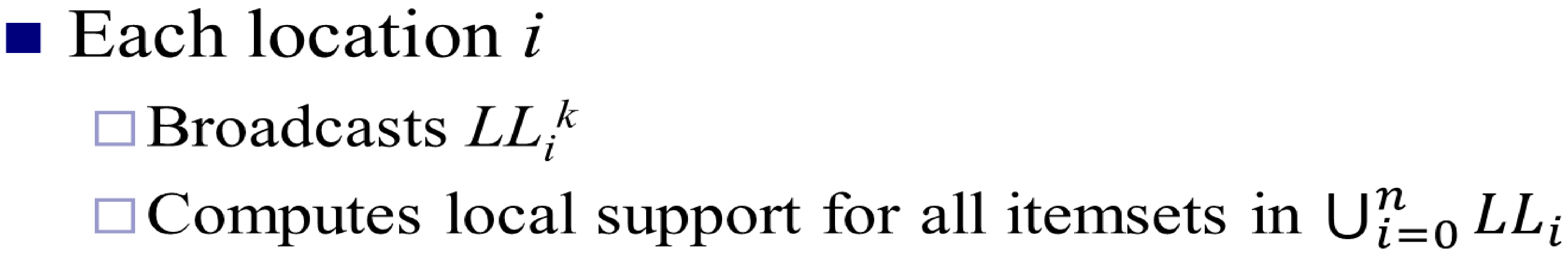
- Scan DB_i to generate $Sup_i(X)$
- If X is locally large

  then add X to LL_i^k

  Locally large itemsets

Recall: if X is supported globally, then it is supported by at least one location

FDM: Exchange



- Each location i
  - □ Broadcasts LL_i^k
  - □ Computes local support for all itemsets in $\bigcup_{i=0}^{n} LL_i$

FDM: Broadcast

- Each location i
  - □ Broadcasts support for each itemset in $\bigcup_{i=0}^{n} L_i$

- From broadcasts, we can easily compute the L^k

Large itemsets

Outline

- Association Rule Mining
- Distributed Rule Mining
- Secure Rule Mining - Semi-Honest Setting
- Secure Rule Mining - Malicious (almost)

A Naïve Solution

- Request each location submit rules with support > *s*
- For each rule
  - Ask each location to submit 2 pieces of information
    1. *Sup*(*rule*)
    2. Total number of transactions in local database
- Allows for the accurate construction of global association rule results

- Big Benefit: No location submits individual records; only aggregates!

Ahem…

- But it also reveals the exact results of each location
    - Potentially sensitive or proprietary knowledge
    - And this is certainly not what we want to do…

Two Phase Solution*

1. FIND ITEMSETS

   Use *commutative encryption* to share rules between locations and eliminate *duplicates*

2. TEST ITEMSETS

   Each locally supported itemset is tested if it holds globally via *random number masking*

*M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering. 2004; 16(9): 1026-1037.

A *View* on Security

- Semi-honest Models
  - Each *participant* (i.e., location) executes the protocol as specified, but
  - Retains what it observes to compromise security

- Computation is *secure* if
  - The "view" of each location during the protocol can be simulated by
    - the input and
    - output of the location

Some More Structure

- Participants
  - *Location:* data holder
  - *Central Authority*: third party, no data at stake

- Honesty
  - *Semi-Honest*: records observations (honest, but curious)
  - *Malicious*: anything goes, wreak havoc

Crypto Crash Course

- *Accumulator*: commutative hash that satisfies:

$$h(h(x, y_1), y_2) \quad = \quad h(h(x, y_2), y_1)$$

- If $A = B$, then $h(h(A,y_1), y_2) = h(h(B,y_2), y_1)$
  - Central authority can compare encrypted data correctly

- *Modular exponentiation* (common in public key encryption) is commutative!
- So you can make an *asymmetric keyed cryptosystem* by pairing keys $\langle y_i, z_i \rangle$

Encryption: $\mathbf{x} = h(\ldots h(\mathbf{v}, y_1), \ldots y_m)$
Decryption: $\mathbf{v} = h(\ldots h(\mathbf{x}, z_1), \ldots z_m)$

More Formally

- Encryption algorithm is commutative if the following hold for
  - Encryption keys $K_1, \ldots, K_n$
  - Message M
  - Any permutation of *i* and *j*

$$E_{Ki}(\ldots E_{Kj}(M) \ldots) = E_{Kj}(\ldots E_{Ki}(M) \ldots)$$

$$\Pr(E_{Ki}(\ldots E_{Kj}(A) \ldots) = E_{Ki}(\ldots E_{Kj}(B) \ldots)) \text{ when } A \neq B$$

$$< 1/2^b$$

[Think of b as message length in bits]

What’s Commutative?

- RSA (Rivest, Shamir, & Adleman)
- $e_n(x, y) = x^y \bmod n$, for appropriately chosen n
- n = product of two large integers p and q
- Computing x from $e_n(x, y)$ can not be accomplished in polynomial time
- Note: e_n can lead to hash collisions…so we further restrict the values of n

Restricted Values

- *n* is chosen from the set of *rigid integers*
  - The product of two *safe primes* *p* and *q*

- Prime is safe if

  $p = 2p' + 1$

  p' is an odd prime

- Further details (e.g., congruency, collision resistance, etc.) are in Benaloh & deMare

J. Benaloh & M. deMare. One-way accumulators: a decentralized alternative to digital signatures. In: EUROCRYPT. 1993: 274-285.

What’s Commutative?

- RSA
- Pohlig-Hellman
- ElGamal

Secure Union: General Idea

- Each location encrypts own
    - locally supported itemsets, and
    - “fake” itemsets (to hide the total number)
- Each location encrypts others itemsets
- Sets of “fully” encrypted itemsets are merged & duplicates are deleted
- Itemsets are decrypted

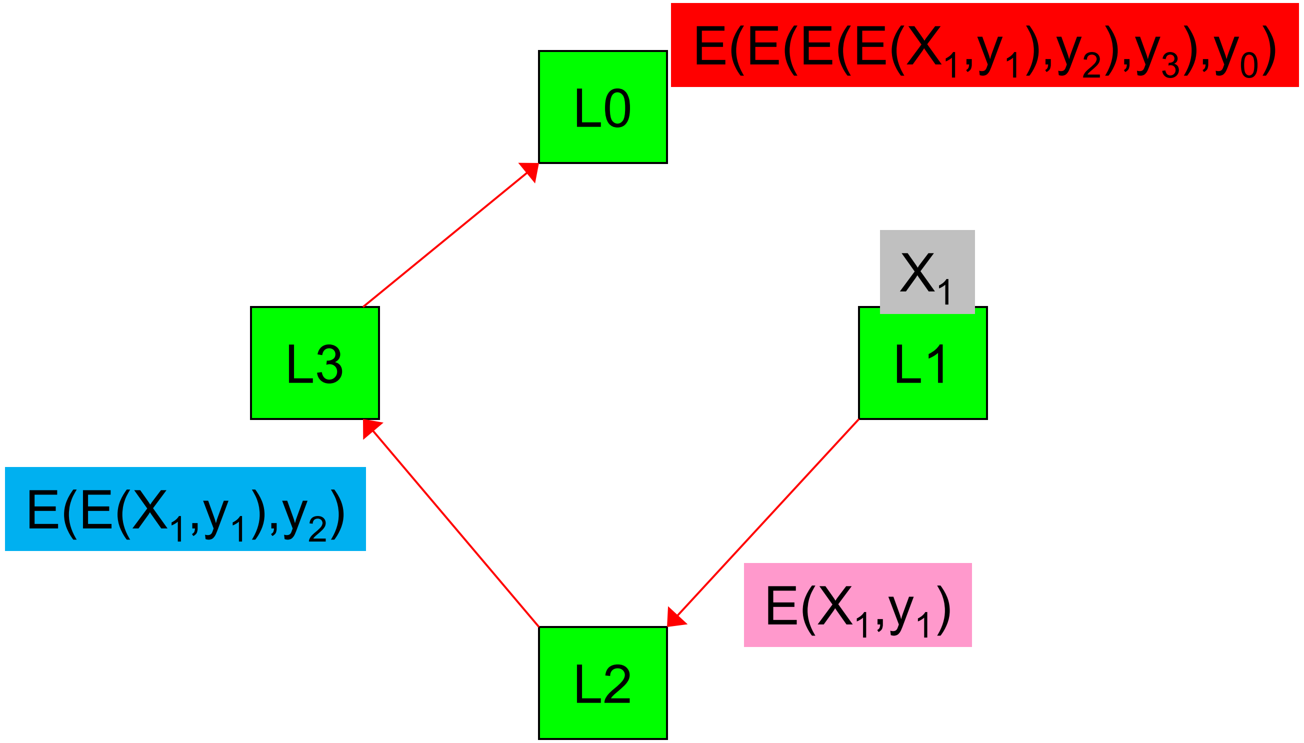
Phase 0: Generation / Encryption

- For each location *i*
    - Generate LL_i^k
    - $E_i(LL_i^k) \leftarrow \{E_\text{i}(X) \mid X \in LL_i^k\} \cup E_i(\textit{Fake Itemsets})$

Phase 1: Encryption (ring)

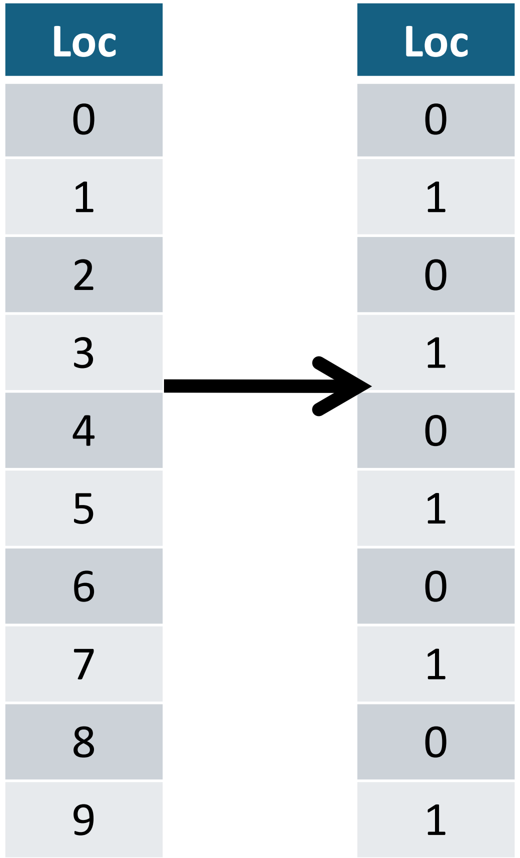
- For j = 0 to N-1 (rounds)
  - if j = 0
    - Each location i sends row shuffled $E_i(LL_i^k)$ to location (i+1) mod N
  - Otherwise
    - Each location i encrypts all items in $E_{i\text{-}j \bmod N}$ $(LL_{i\text{-}j \bmod N}^k)$ with E_i
    - Permutes the encrypted records
    - Sends the results to site (i+1)mod$N$
- Notice: Location i has the fully encrypted itemsets of location (i+1) mod N

Example: Encrypt and Pass



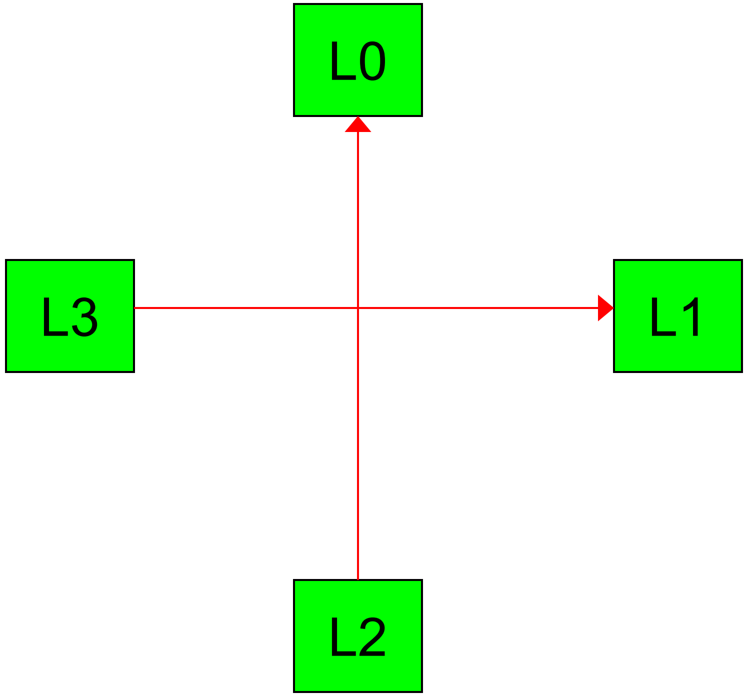
Phases 2 and 3: Merging

- Phase 2: Merge odd / even locations
  - Each location *i* sends fully encrypted records to location 1-(((i+1)mod *N*)mod 2)
  - Location 0 unions the received set of records
  - Location 1 unions the received set of records
- Phase 3: Merge all
  - Location 1 sends permuted result to location 0
  - Location 0 merges the records to form *All_Itemset*

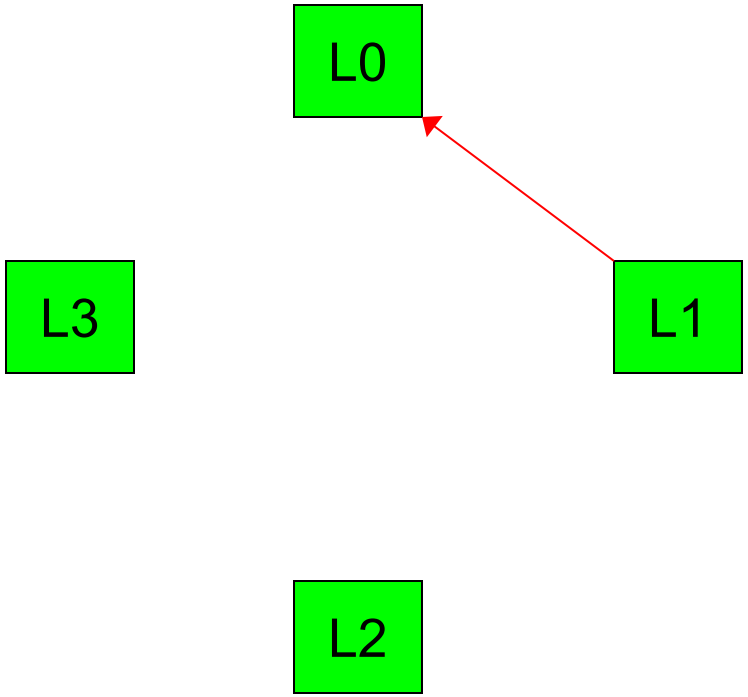


Loc	Loc
0	0
1	1
2	0
3	1
4	0
5	1
6	0
7	1
8	0
9	1

Merge Odd / Even



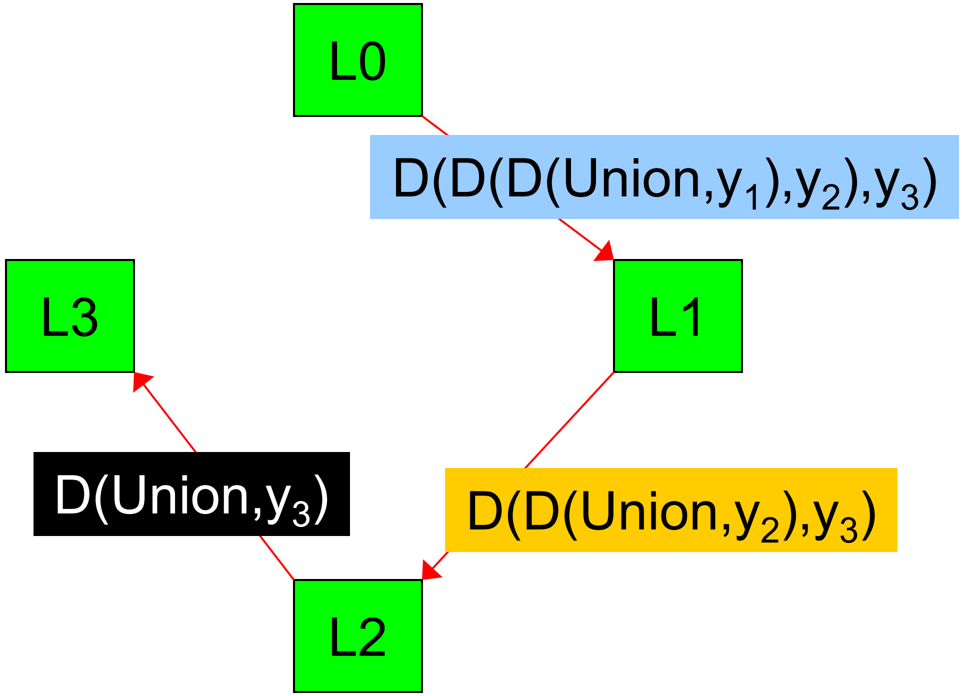
Merge All



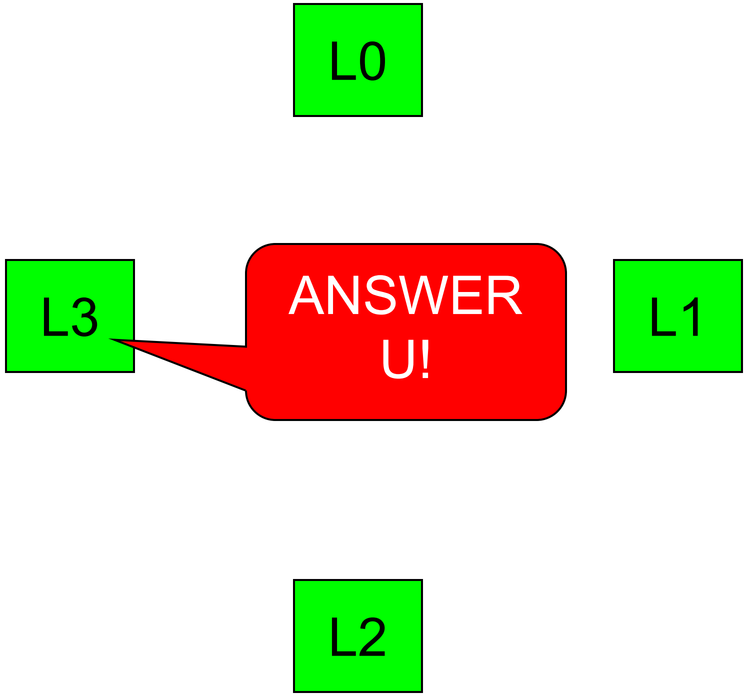
Phase 4: Decryption

- For i = 0 to N-1
  - Location i decryptes *All_Itemset* with D_i
  - Location i sends *All_Itemset* to location (i+1) mod N
- Site N-1 decrypts *All_Itemset* using $D_{N\text{-}1}$
- Site N-1 removes fake itemsets
- Site N-1 broadcasts *All_Itemset* to all sites

Decryption Down the Ring

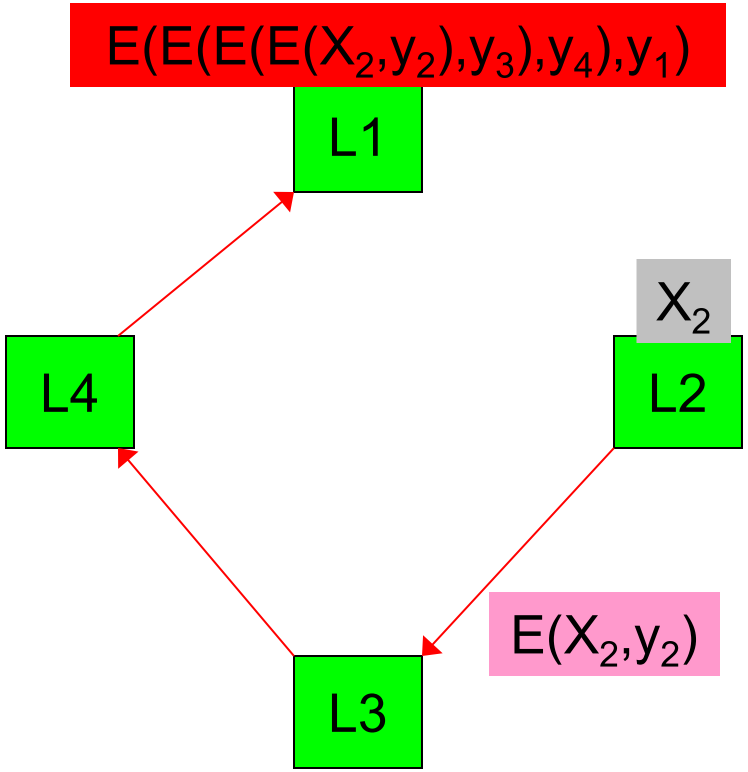


Broadcast!



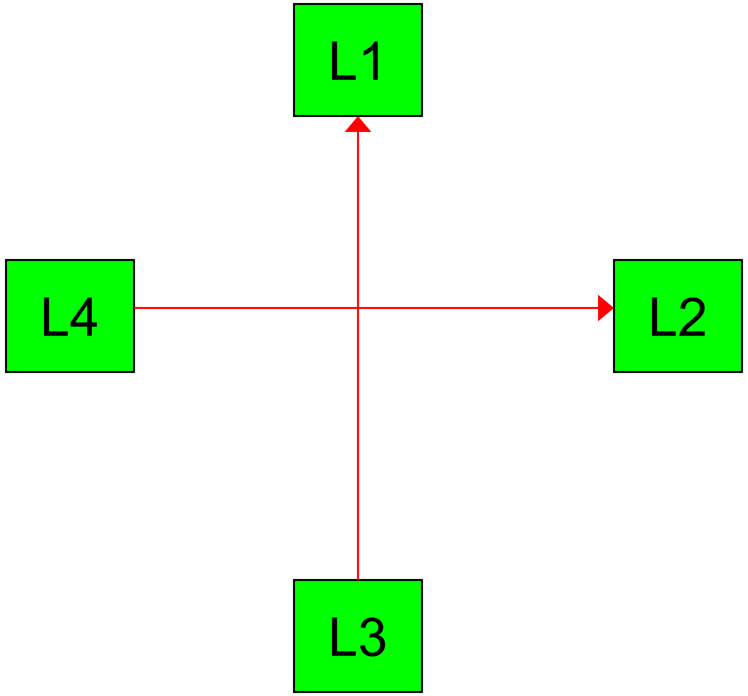
Example: Set Union

- Prior model
  - Encrypt and pass



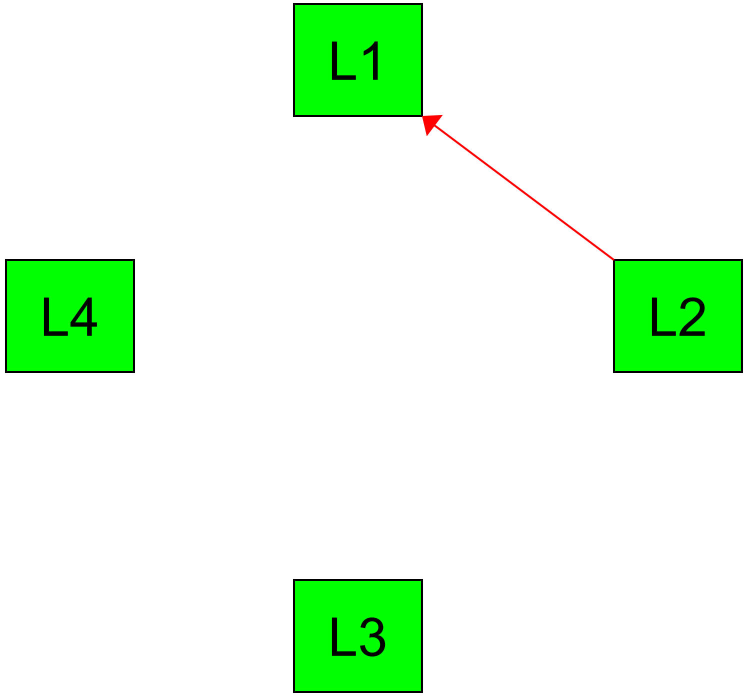
Example: Set Union

- Prior model
  - Encrypt and pass
  - “odd”/ “even” locations pass to location L1 / L2



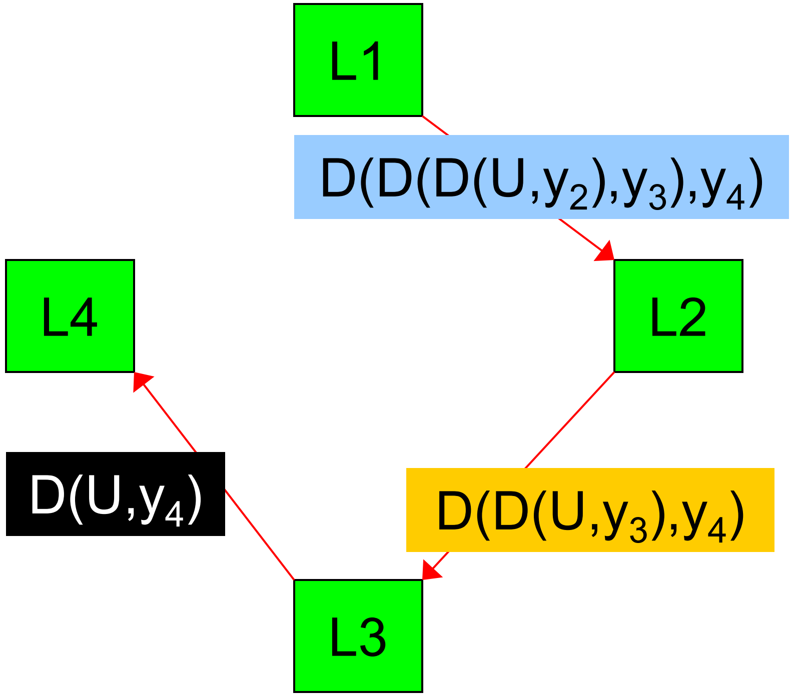
Example: Set Union

- Prior model
  - Encrypt and pass
  - “odd”/ “even” locations pass to location L1 / L2
  - L1 and L2 union data → U



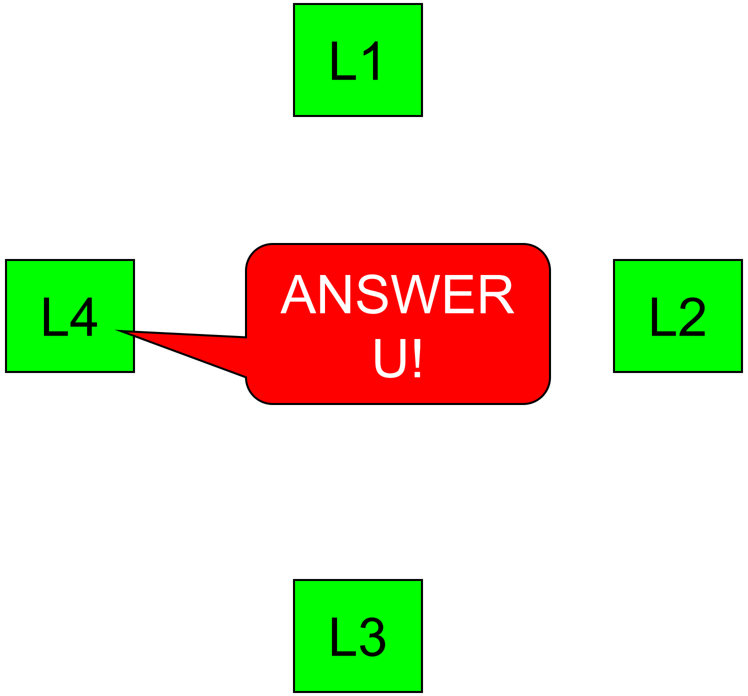
Example: Set Union

- Prior model
  - Encrypt and pass
  - “odd”/ “even” locations pass to location L1 / L2
  - L1 and L2 union data
  - Decryption down the ring



Example: Set Union

- Prior model
  - Encrypt and pass
  - “odd”/ “even” locations pass to location L1 / L2
  - L1 and L2 union data
  - Decryption down the ring
  - Last location broadcasts



Blockchain

- A blockchain is a specific type of distributed ledger.
- A reconceptualization of trust relationships
  - No need for third party intermediaries
  - Highly secure
- Distributed trust – many witnesses are better than one
- Blockchain features
  - Peer-to-peer (decentralized) network
    - Distributed ledger
  - Hashing
  - Consensus mechanism

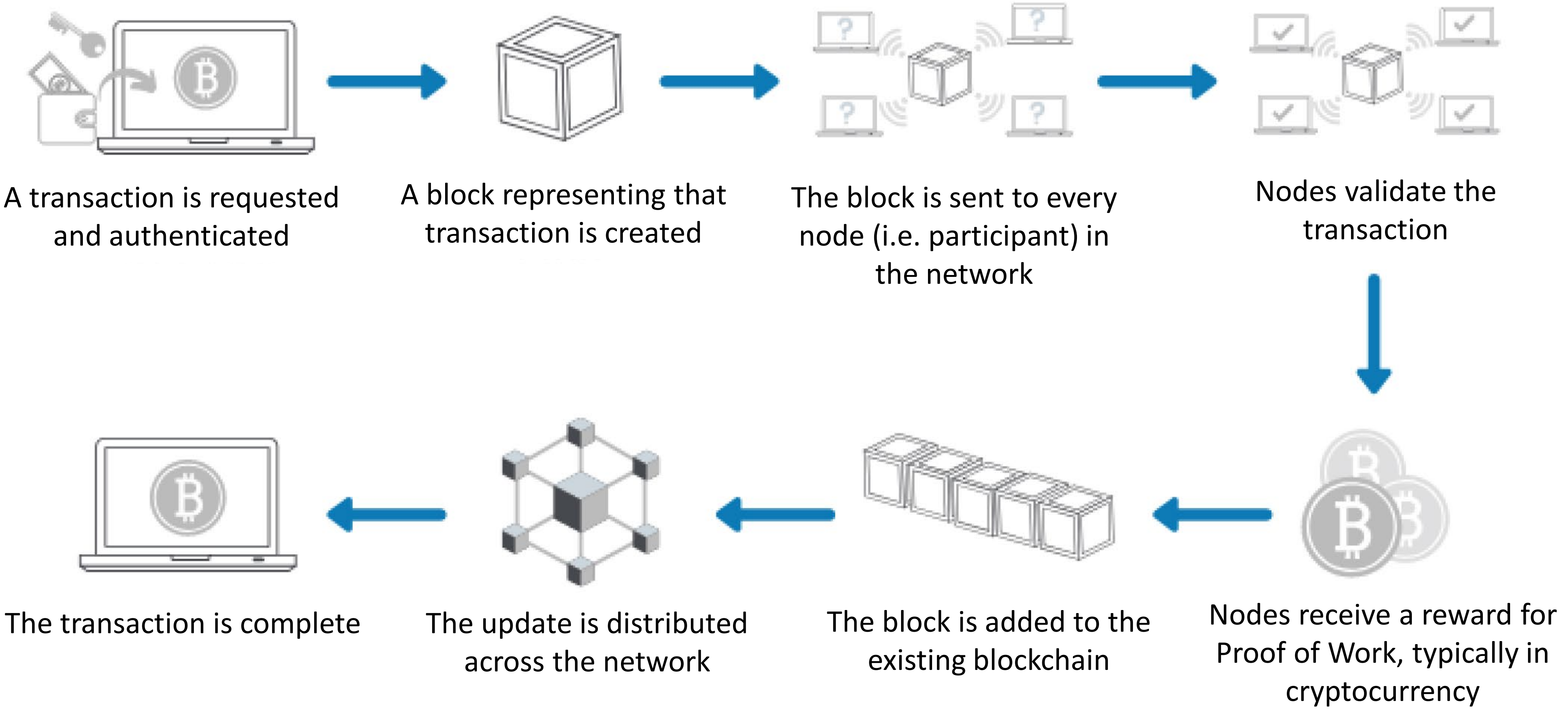
Concepts and Definitions

- **Ledger** – A record of transactions over time while still allowing for tracking and analysis. It documents the transfer of ownership and is ultimately a means for proving ownership.
- **Block** – A block is a unit of data (or record) that holds a collection of transactions which, together with many other blocks arranged in a specific order, form a blockchain.
- **Hash** – Digital equivalent of a fingerprint; unique and useful for detecting change in a file. This is one component that makes the blockchain secure.
- **Consensus mechanism** – A fault-tolerant process to achieve agreement about a set of data among many users or nodes. Proof of work is one of the most common consensus mechanisms.

Concepts and Definitions (Cont.)

- **Miner** – A blockchain user/nodes who participates in a competition with others to solve complex cryptographic problems, in order to validate a particular block, have that block added to the blockchain, and receive a reward for doing so.
- **Blockchain** can refer to:
  - A data structure which represents a series of immutable transaction records
  - An algorithm
  - A collection of technologies
  - A distributed, peer-to-peer network of systems
  - A system of recording information in a way that makes it difficult or impossible to change, hack, or cheat the system.

Blockchain Functionality



A transaction is requested and authenticated

A block representing that transaction is created

The block is sent to every node (i.e. participant) in the network

Nodes validate the transaction

Nodes receive a reward for Proof of Work, typically in cryptocurrency

The block is added to the existing blockchain

The update is distributed across the network

The transaction is complete

History of Blockchain

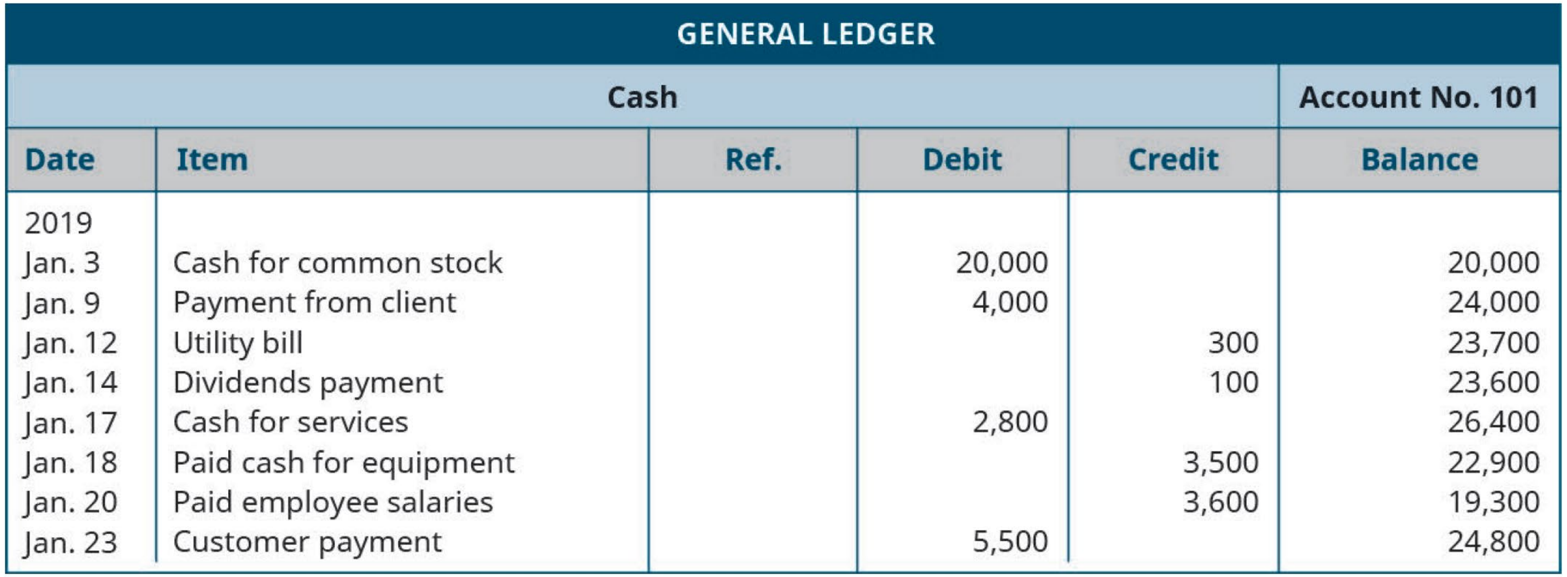
- 1982 – Cryptographer David Chaum proposes an early form of blockchain in his dissertation, "Computer Systems Established, Maintained, and Trusted by Mutually Suspicious Groups."
- 1991 – Stuart Haberand and Scott Stornetta improved on the idea in their Journal of Cryptology article, "How to time-stamp a digital document.“
- 1995 – David Chaum created the first digital currency, DigiCash, which used blind signatures for anonymous transactions. Digicash would go bankrupt in 1998.
- 1996 – E-Gold was started by Douglas Jackson and Barry Downey, which was a digital currency backed by gold; the company was eventually brought down due to facilitation of fraud.
- 1997 – Cryptographer Adam Back develops Hashcash, an email filter based on a proof-of-work system to prevent spam and denial of service (DoS) attacks. It appended a textual encoding of a hashcash stamp to the email header to prove that the sender utilized some CPU power in calculating the hashcash stamp.
- 1998 – Computer engineer Wei Dai published a paper called “B-money, an Anonymous, Distributed Electronic Cash System” which contained many of the concepts used by modern cryptocurrencies, such as anonymity and the lack of traceability, and the ability to enforce contracts within the network.
- 2007 to 2008 – Global financial crisis
- 2008 – Satoshi releases white paper implements Blockchain and Bitcoin

Global Impact of Blockchain

- Estimated 300M global cryptocurrency users
  - Bitcoin
    - Global market cap: $775B
    - Accepted by more than 15K businesses for payment globally
  - Thousands of altcoins in existence
  - 81 countries considering implementing central bank digital currencies; 9 have implemented pilot programs
- Defi (decentralized finance) as a branch of finance is growing aggressively (more than $20B)
- Cryptoeconomics (economics based on blockchain technologies) is now a recognized academic field
- Non-fungible tokens (NFTs) are selling for millions of dollars
- Decentralized Autonomous Organizations (DAOs) are becoming more common

What is a ledger?

- A record of accounts and transactions, usually including a beginning and ending balance.
  - Financial transactions related to a company
  - Three types: Creditors, debtors and general
  - General ledger: Assets, liabilities, income, expenses and capital



GENERAL LEDGER					
Cash					Account No. 101
Date	Item	Ref.	Debit	Credit	Balance
2019					
Jan. 3	Cash for common stock		20,000		20,000
Jan. 9	Payment from client		4,000		24,000
Jan. 12	Utility bill			300	23,700
Jan. 14	Dividends payment			100	23,600
Jan. 17	Cash for services		2,800		26,400
Jan. 18	Paid cash for equipment			3,500	22,900
Jan. 20	Paid employee salaries			3,600	19,300
Jan. 23	Customer payment		5,500		24,800

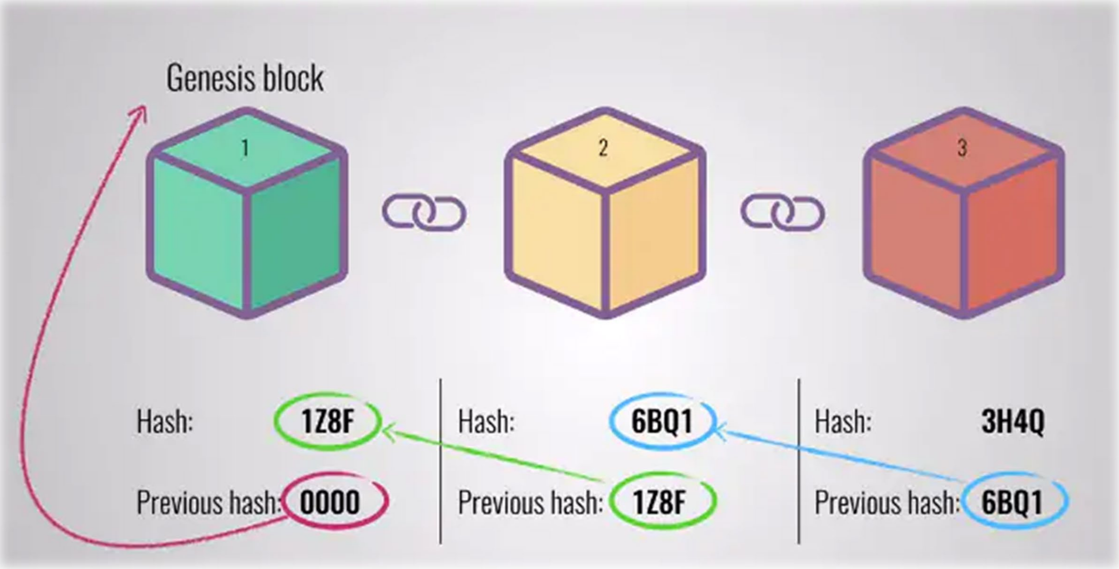
Block Composition

- Each block contains:
  - Data – Purpose of the blockchain will dictate type of data
  - Hash – Digital fingerprint, identifies block and all its contents uniquely
  - Previous hash – Links current block to previous block; key to security

Again, each of these blocks refers to the previous block due to the fact that it contains a hash of it. If a block is tampered with, the hash of it will change for it and all previous blocks, making detection trivial.

This is one of the key concepts of blockchain integrity, and is one of several mechanisms that ensures that records can't be deleted or modified.

Also worth noting: The first block in any blockchain is called the genesis block.



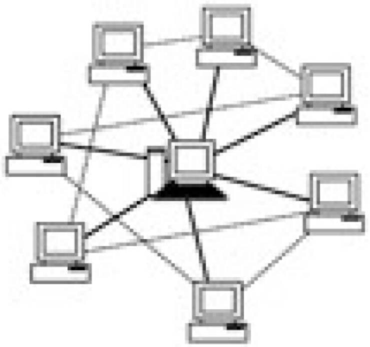
Consensus Mechanisms

- Hashes are necessary but not sufficient to prevent tampering. Why?
  - Hashing means tampering with a block makes all following blocks invalid. However, due to modern processing capabilities, it is possible to calculate a large number of hashes – those in all the following blocks – all over again, making that modified blockchain valid again. Therefore, we need something else.
- Blockchain has a consensus mechanism called proof of work. What is a consensus mechanism?
  - A consensus mechanism refers to any number of methodologies used to achieve agreement, trust, and security across a decentralized computer network. It slows down the creation of new blocks.

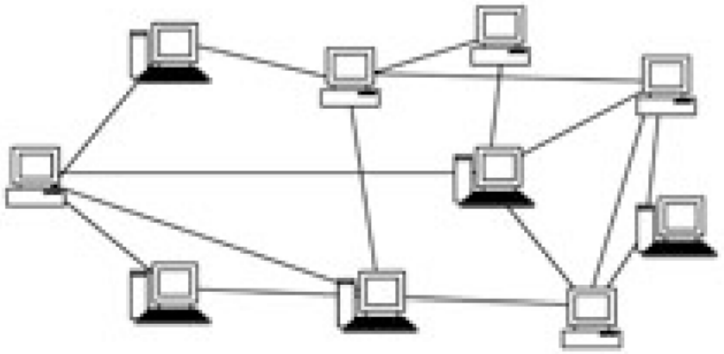
Proof of Work

- Proof of work is used to **validate transactions and broadcast new blocks to the blockchain**.
- Proof of work is a way for someone to prove that they have engaged in a significant amount of computational effort. That effort can be validated in a way that is easier and quicker than the original calculations.
- How does it work?
  - Miners on a network will compete against each other in solving complex computational puzzles.
    - “Miners” are actually working to guess a pseudorandom number.
    - When the solution is found by a miner, other miners will validate it.
    - Upon validation, the miner is rewarded with a block reward.

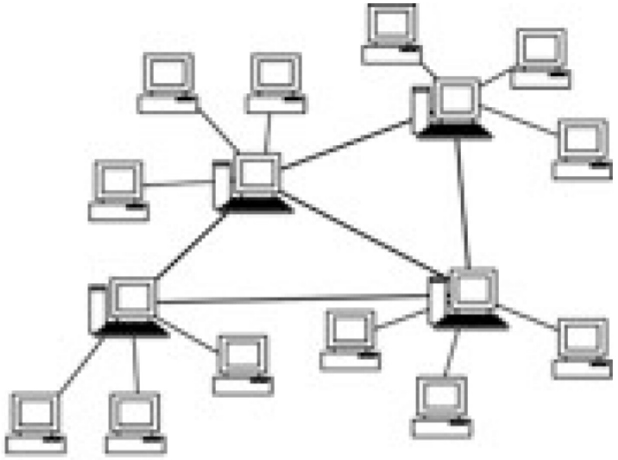
Decentralized Networks



Centralized overlay.
Central peer facilitates the interactions among the leaf-peers.



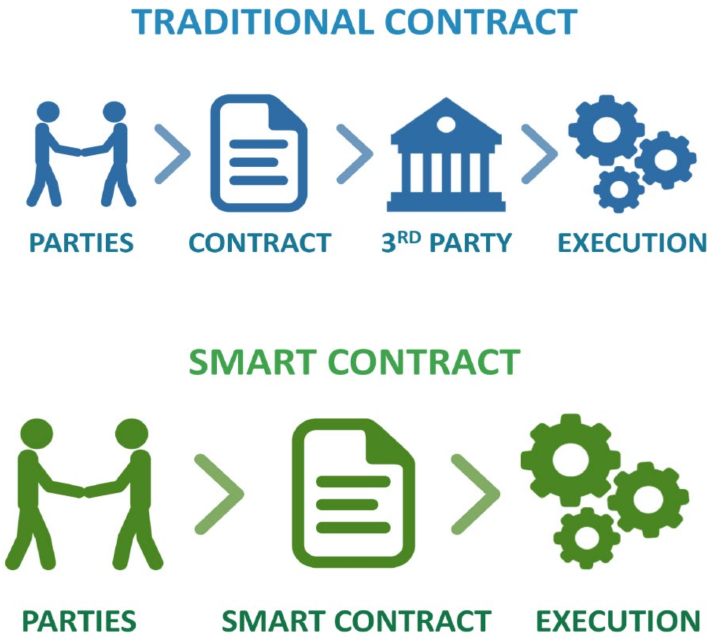
Decentralized overlay.
No central authority, all peers treated equally.



Hybrid overlay.
Hierarchical topology,
Interconnected super-peers locally serve the subsets of leaf-peers.

Blockchain Implementation: Smart Contracts

- What is a smart contract?
  - A self-executing contract has certain terms of the agreement, which are automatically initiated when specified conditions are met
  - Run on blockchain and use algorithms to create and measure execution conditions
- Properties:
  - Self-executable
  - Self-verifiable
  - Highly resistant to tampering
- Benefits
  - High level of trust
  - Minimization of errors
  - Resistant to fraud
  - Cost-efficient

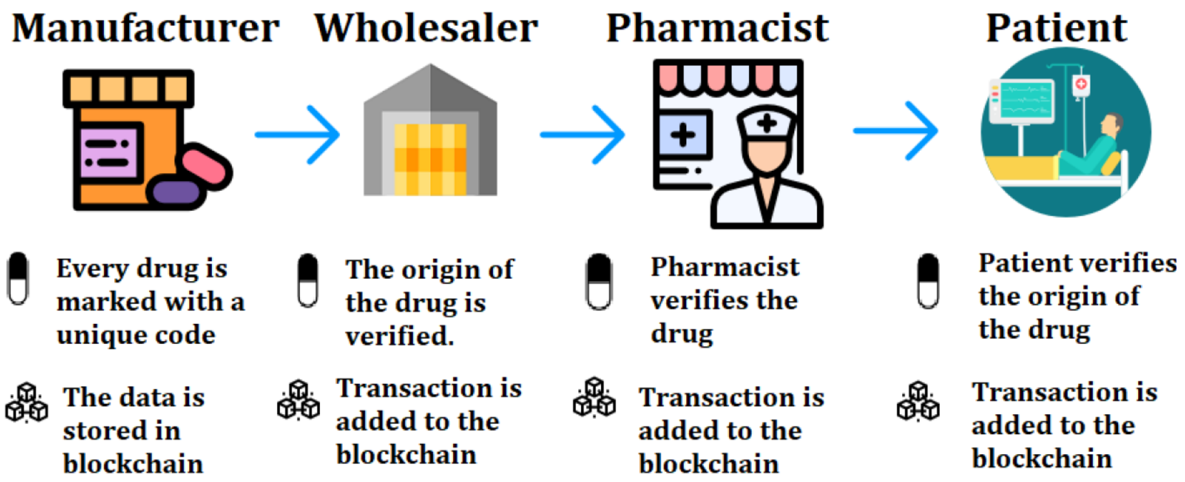


Blockchain Implementation: Non-fungible Tokens (NFTs)

- What is a non-fungible token?
  - It’s “non-fungible” – it’s unique and can’t be exchanged for another identical NFT.
  - It’s digital in its form – it doesn’t exist in the physical world.
  - It’s a unique digital/virtual asset with ownership certified by blockchain.
  - One example: Digital collector of unique art
    - Jack Dorsey sold a certified copy of his first tweet for $2.9M, and donated the proceeds to charity.
    - An artist sold a collage of his art in JPG –a digital graphics format –for $69M.
  - Markets for NFTs exist and are growing quickly.

Healthcare Blockchain Use Case: Supply Chain Transparency

- Challenge: Assuring the authenticity, origin and supply chain of medical products – easier said than done in a globalized world where international commerce can create complications
- Especially important in developing markets where counterfeit prescription medicines and medical devices can cause tens of thousands of deaths annually
- To solve this, companies and end consumers need to be able to track each package's end-to-end movement from the point of origin, including manufacturers, wholesale, transport, etc.
- Blockchain can enable companies throughout the prescription drug supply chain to verify the authenticity of medicines, expiry dates and other important information.



Manufacturer → Wholesaler → Pharmacist → Patient

Manufacturer: Every drug is marked with a unique code. The data is stored in blockchain

Wholesaler: The origin of the drug is verified. Transaction is added to the blockchain

Pharmacist: Pharmacist verifies the drug. Transaction is added to the blockchain

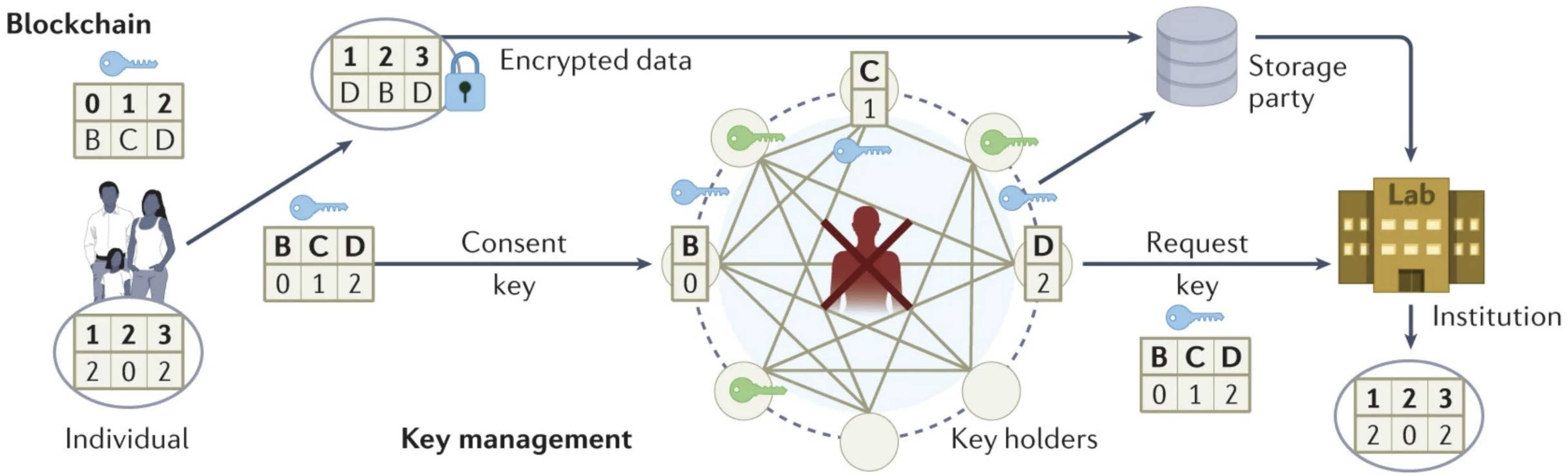
Patient: Patient verifies the origin of the drug. Transaction is added to the blockchain

Healthcare Blockchain Use Case: Electronic Health Records

- Immediate and secure access to health records by patients and their healthcare providers
    - Challenge: Ensuring patient access to all their heath/medical records across all service providers in order to have a complete view of medical histories, while ensuring their records are secure.
    - Johns Hopkins University published research in 2016 revealing that the third leading cause of death in the US was medical errors that resulted from poorly coordinated care, such as planned actions not completed as intended, or errors of omission in patient records.
    - Blockchain-based medical record systems can be linked into existing medical record software and act as an overarching, single view of a patient’s record without placing patient data on the blockchain.
    - Each new record can be appended to the blockchain in the form of a unique hash function, which can only be decoded if the person who owns the data – in this case, the patient – gives their consent.
    - Benefits:
        - A comprehensive, single-source for accurate medical records
        - Direct access by medical insurers of validated, confirmed of healthcare services directly from patients, not requiring time and cost of an intermediary
        - The development of further advances in analytics

Healthcare Blockchain Use Case: Genomic Data Privacy

- A blockchain enables encrypted immutable records stored on a decentralized network. Here, the individual manages the decryption key using a blockchain while sharing encrypted data with researchers



Remote Consent and eConsent

- Both remote consent and eConsent were in limited use pre-pandemic.

- Not surprisingly, there was an increase in their use during the Covid-19 pandemic.

- Will there be continued or increasing use of these modalities in the consent process moving forward?

Remote Consent

- Remote consent is a consent process that allows the person conducting the consent and the potential participant to engage in the informed consent process in a way that is similar to what would be conducted in-person under normal circumstances without being in the same physical location.

eConsent

- eConsent is an electronic media/format that can be used to supplement or replace paper-based informed consent forms to provide information to a potential research participant. It can also be used to obtain documentation of consent (e-signatures.)

- eConsent can be conducted both in person or remotely.

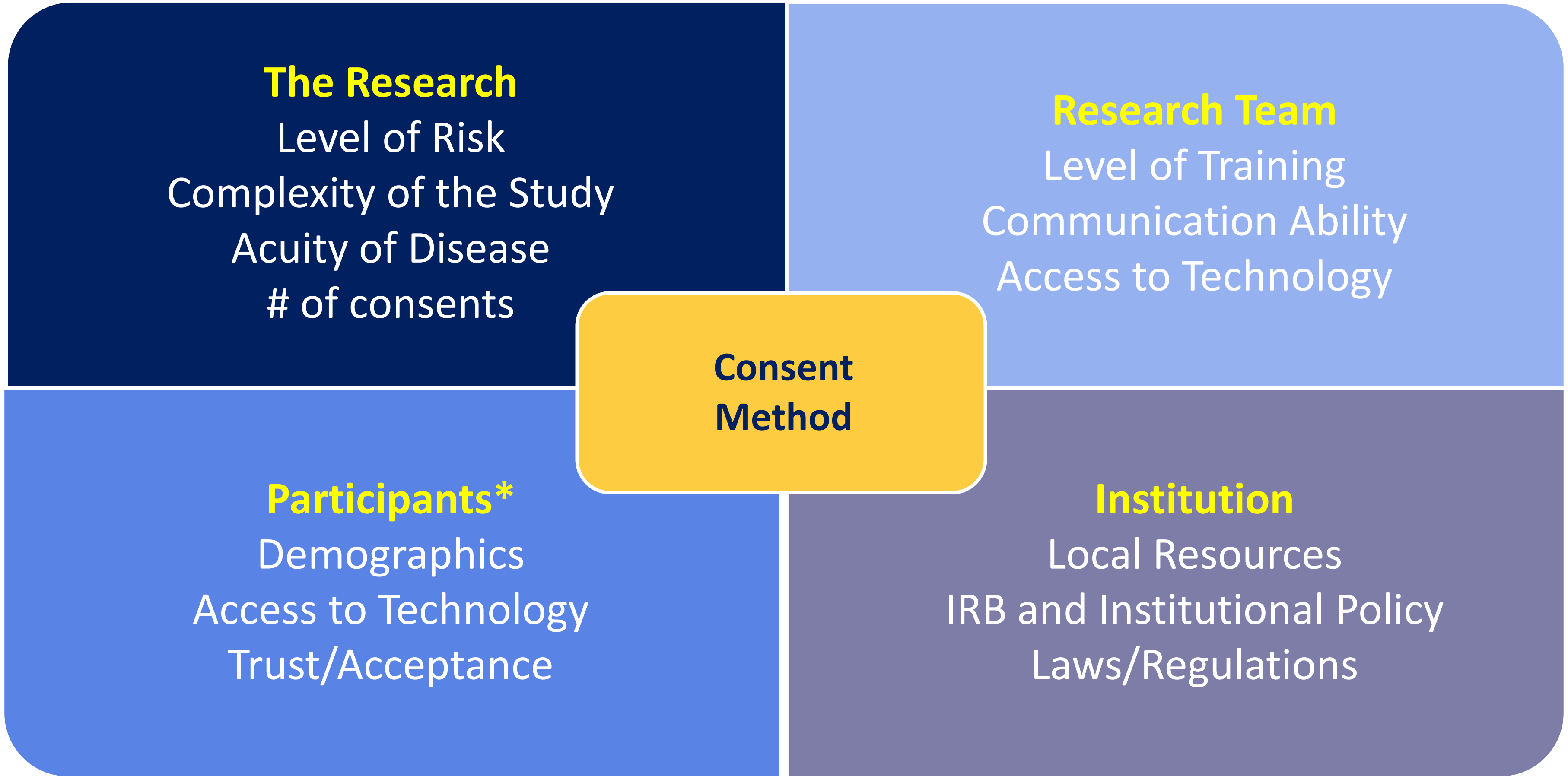
Remote and eConsent

- Combinations of remote, e-consent, and or e-signatures can be used together to accomplish the Informed Consent Process.

- How these modalities are implemented should take into consideration the qualities of the research study, the type of participants, regulations, institutional policies and the research team itself.

- There is no single “best way” to execute these consent modalities for all research.

Use of Remote and eConsent



The Research
Level of Risk
Complexity of the Study
Acuity of Disease
# of consents

Research Team
Level of Training
Communication Ability
Access to Technology

Consent Method

Participants*
Demographics
Access to Technology
Trust/Acceptance

Institution
Local Resources
IRB and Institutional Policy
Laws/Regulations

Chen C, Replacing Paper Informed Consent with Electronic Informed Consent for Research in Academic Medical Centers: A Scoping Review. (2020)

eConsent Best Practices

- “Whether part or all of the eIC process takes place on-site or remotely, the responsibility for obtaining informed consent remains with the investigator and the study personnel to which responsibility has been appropriately delegated. **The investigator cannot delegate the authority to obtain consent to the electronic system.**” FDA Guidance “Use of Electronic Consent” (2016)

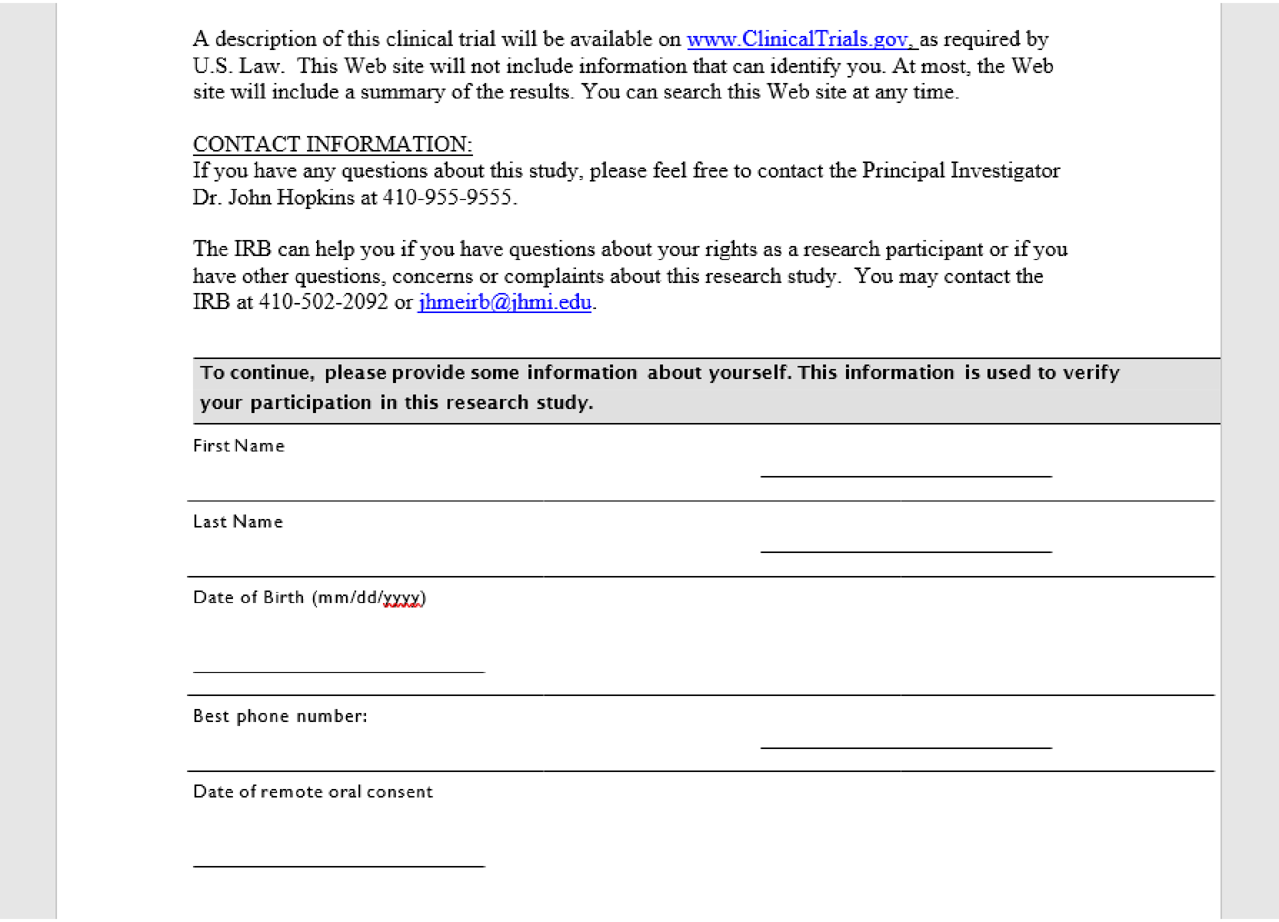
eSignatures are an important part of eConsent

- In-person eSignatures -personnel may verify the persons identity
- Remote eSignature -user name password combinations, computer readable ID cards, biometrics, or digital signatures.

REDCap versus DocuSign

REDCap Research Electronic Data Capture	DocuSign
Platform for presenting consent information electronically	Platform for obtaining electronic consent signatures
May included interactive elements: images, videos, narration or quizzes to aid and evaluate understanding	Uses standard pdf image of the consent document.
May not be used for FDA regulated research unless upgraded to Part 11 compliant	May be used for FDA regulated studies (and others.)
Signatures may not be legally valid (authentication and attestation needed)	Signatures legally valid

eConsent Script with REDCap ‘Agree to Participate’ Page



A description of this clinical trial will be available on www.ClinicalTrials.gov, as required by U.S. Law. This Web site will not include information that can identify you. At most, the Web site will include a summary of the results. You can search this Web site at any time.

CONTACT INFORMATION:
If you have any questions about this study, please feel free to contact the Principal Investigator Dr. John Hopkins at 410-955-9555.

The IRB can help you if you have questions about your rights as a research participant or if you have other questions, concerns or complaints about this research study. You may contact the IRB at 410-502-2092 or jhmeirb@jhmi.edu.

To continue, please provide some information about yourself. This information is used to verify your participation in this research study.

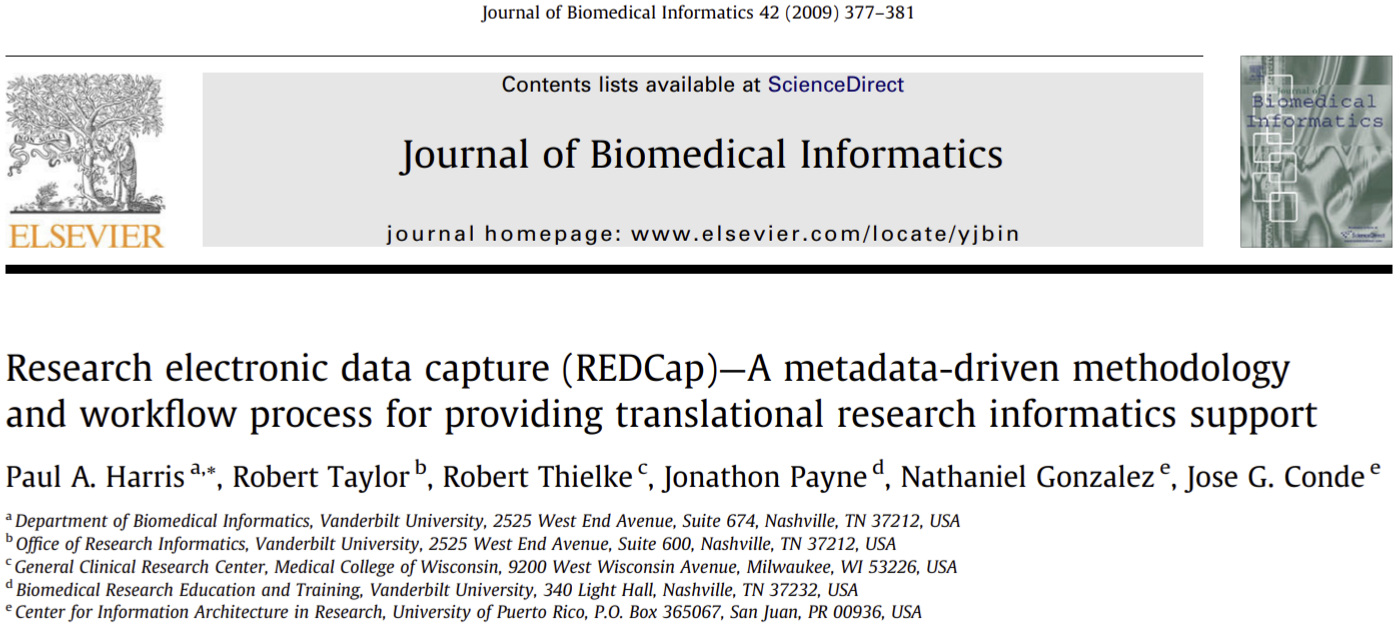
First Name ______________________

Last Name ______________________

Date of Birth (mm/dd/yyyy) ______________________

Best phone number: ______________________

Date of remote oral consent ______________________



Journal of Biomedical Informatics 42 (2009) 377–381

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support

Paul A. Harris [a,*], Robert Taylor [b], Robert Thielke [c], Jonathon Payne [d], Nathaniel Gonzalez [e], Jose G. Conde [e]

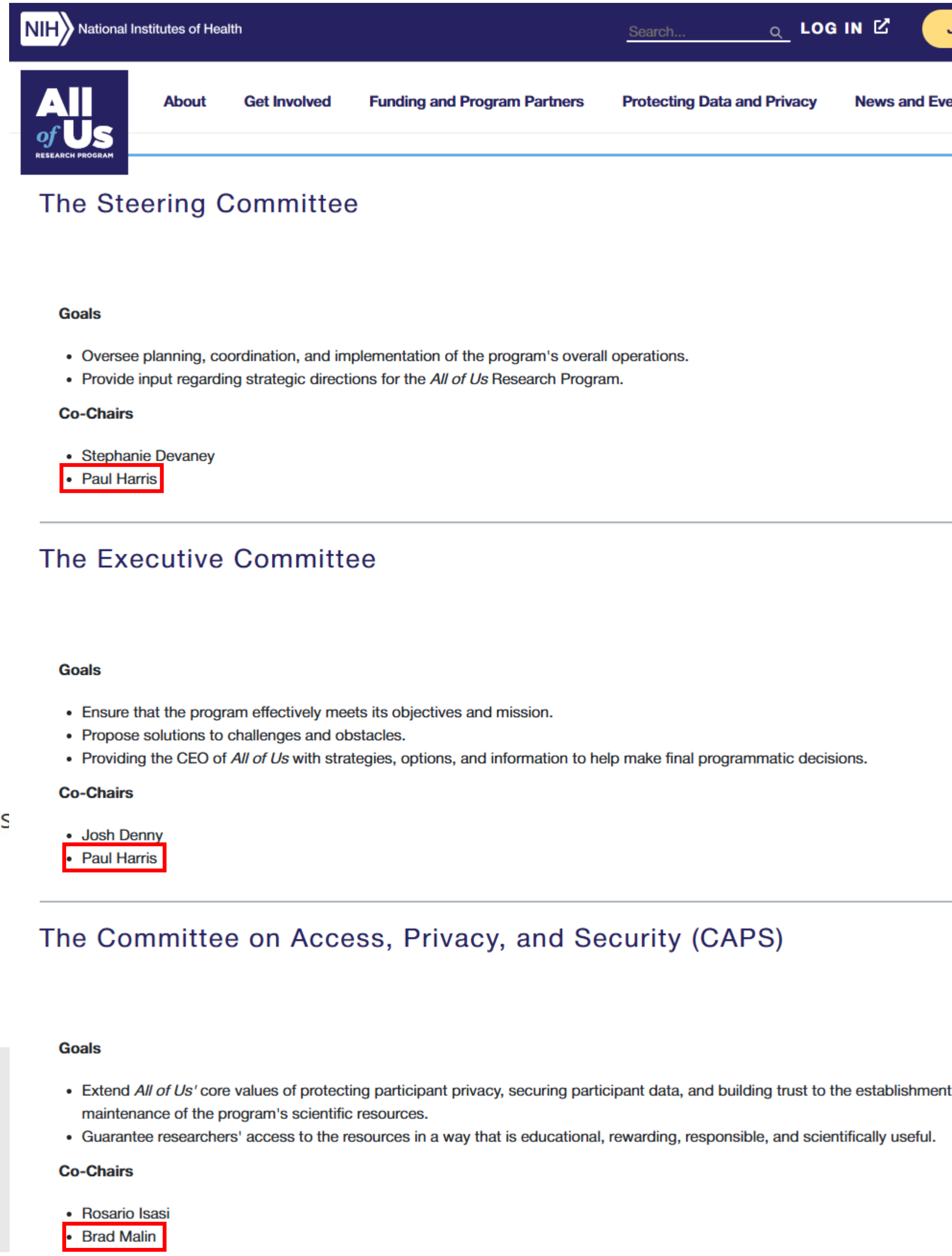
[a] Department of Biomedical Informatics, Vanderbilt University, 2525 West End Avenue, Suite 674, Nashville, TN 37212, USA
[b] Office of Research Informatics, Vanderbilt University, 2525 West End Avenue, Suite 600, Nashville, TN 37212, USA
[c] General Clinical Research Center, Medical College of Wisconsin, 9200 West Wisconsin Avenue, Milwaukee, WI 53226, USA
[d] Biomedical Research Education and Training, Vanderbilt University, 340 Light Hall, Nashville, TN 37232, USA
[e] Center for Information Architecture in Research, University of Puerto Rico, P.O. Box 365067, San Juan, PR 00936, USA

Paul Harris, PhD, FACMI, FIAHSI

Professor, Department of Biomedical Informatics
Professor, Department of Biomedical Engineering
Professor, Department of Biostatistics
Vice President for Research Informatics, VUMC Office of Research Informatics

2525 West End Avenue
Nashville, TN, Tennessee
37203

Dr. Harris devised and created **REDCap**, a data collection platform that has seen widespread adoption by more than 3600 institutional partners and over 1 million end-users across 131 countries.
He also created and runs **ResearchMatch**, a national program designed to match individuals wishing to volunteer for studies and researchers recruiting patients for studies and trials. ResearchMatch serves approximately 150,000 research volunteers and 165 research institutions.



NIH National Institutes of Health
Search...
LOG IN
All of Us RESEARCH PROGRAM
About
Get Involved
Funding and Program Partners
Protecting Data and Privacy
News and Eve

The Steering Committee

Goals

- Oversee planning, coordination, and implementation of the program's overall operations.
- Provide input regarding strategic directions for the *All of Us* Research Program.

Co-Chairs

- Stephanie Devaney
- Paul Harris

The Executive Committee

Goals

- Ensure that the program effectively meets its objectives and mission.
- Propose solutions to challenges and obstacles.
- Providing the CEO of *All of Us* with strategies, options, and information to help make final programmatic decisions.

Co-Chairs

- Josh Denny
- Paul Harris

The Committee on Access, Privacy, and Security (CAPS)

Goals

- Extend *All of Us'* core values of protecting participant privacy, securing participant data, and building trust to the establishment maintenance of the program's scientific resources.
- Guarantee researchers' access to the resources in a way that is educational, rewarding, responsible, and scientifically useful.

Co-Chairs

- Rosario Isasi
- Brad Malin

Future of Medical Data Privacy (Venues)

- Biomedical Informatics
  - AMIA (American Medical Informatics Association) Annual Symposium
  - IEEE International Conference on Healthcare Informatics (ICHI)
  - AIME (Artificial Intelligence in Medicine)
- Privacy & Security
  - USENIX Security Symposium
  - Privacy Enhancing Technologies Symposium (PETS)
  - iDASH Privacy & Security Workshop

Future of Medical Data Privacy (Venues)

- AI
  - AAAI/ACM Conference on AI, Ethics, and Society (AIES)
  - NeurIPS – Datasets and Benchmarks Track / Privacy in ML Workshop
- Biomedical Big Data
  - IEEE BIBM (Bioinformatics and Biomedicine)
  - Health Informatics Meets Digital Health (Hi-DH) Workshop @ ECML-PKDD

Future of Medical Data Privacy (Journals)

- Biomedical Informatics

  Health informatics, privacy, EHR security, data governance

  - Journal of the American Medical Informatics Association (JAMIA) (IF:7)
  - npj Digital Medicine (IF:15) Digital health, AI, clinical informatics, and data privacy
  - Journal of Biomedical Informatics (IF:6.5) Computational methods in biomedical data privacy and security
  - IEEE Journal of Biomedical and Health Informatics (JBHI) (IF:5)

    Privacy-preserving machine learning, wearable health data privacy
  - Computers in Biology and Medicine

    Applied computing for health data, including secure and ethical data handling
- Privacy & Security
  - ACM Transactions on Privacy and Security (TOPS)

    Formal security/privacy models, data access, anonymization in health data

Future of Medical Data Privacy (Journals)

- AI

  Ethical AI in medicine, fairness, explainability, privacy in LLMs
  - Nature Machine Intelligence (IF:25)
  - AI in Healthcare Responsible AI, secure AI implementation, ethical data-driven health systems
  - Patterns Data science for societal good, including healthcare privacy & transparency
- Multidisciplinary
  - Science Advances (IF:13)

  Interdisciplinary science, including societal impact of health data sharing and privacy

Readings for the Next Week

- 1. Wiest IC, Leßmann ME, Wolf F, Ferber D, Treeck MV, Zhu J, Ebert MP, Westphalen CB, Wermke M, Kather JN. **Deidentifying Medical Documents with Local, Privacy-Preserving Large Language Models: The LLM-Anonymizer**. *NEJM AI*. 2025 Mar 27;2(4):AIdbp2400537.

- Optional
  - ❑ Li H, Chen Y, Luo J, Wang J, Peng H, Kang Y, Zhang X, Hu Q, Chan C, Xu Z, Hooi B. **Privacy in large language models: Attacks, defenses and future directions**. *arXiv preprint* arXiv:2310.10383. 2023 Oct 16.
  - ❑ Du H, Liu S, Zheng L, Cao Y, Nakamura A, Chen L. **Privacy in Fine-tuning Large Language Models: Attacks, Defenses, and Future Directions**. *arXiv preprint* arXiv:2412.16504. 2024 Dec 21.
  - ❑ Das BC, Amini MH, Wu Y. **Security and privacy challenges of large language models: A survey**. *ACM Computing Surveys*. 2025 Feb 10;57(6):1-39.

Feedback Survey

- One thing you learned or felt was valuable from today's class & reading
- Muddiest point: what, if anything, feels unclear, confusing or "muddy"
- https://www.wjx.cn/vm/hX0mIro.aspx