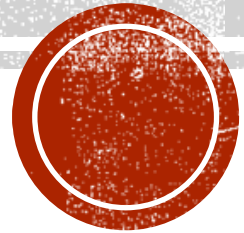
DATA PRIVACY



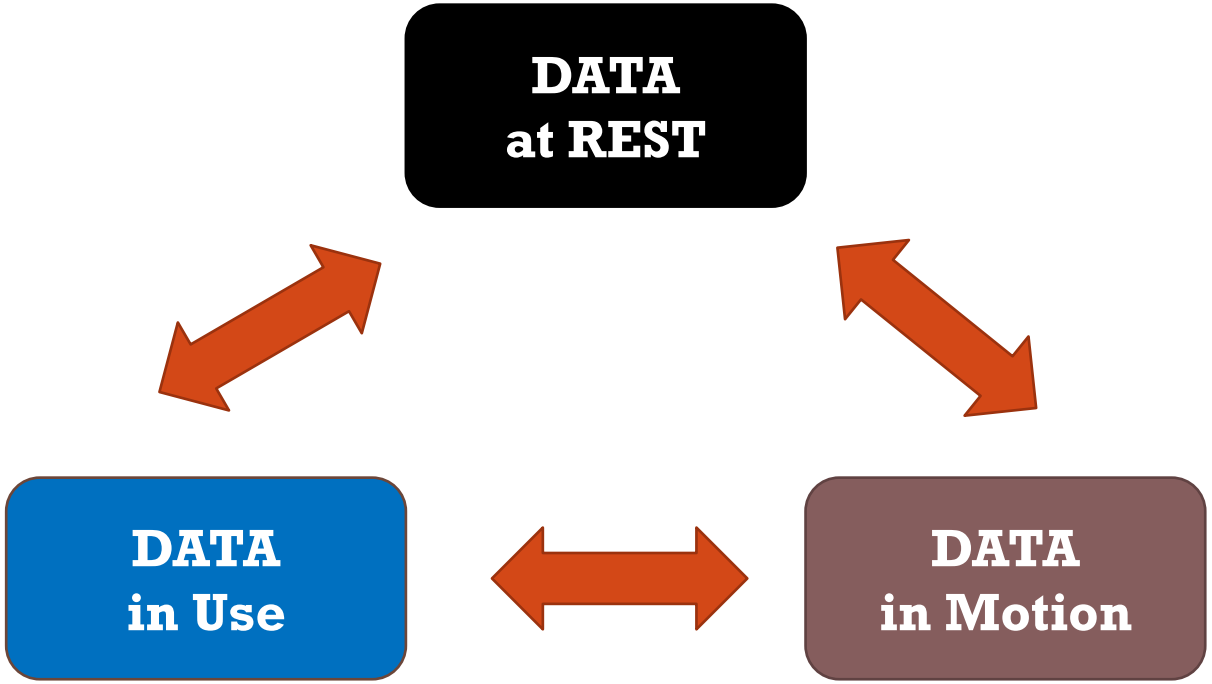
CS 361S

Fall 2021

REFERENCES

- ***HAC*** – *Handbook of Applied Cryptography* by Alfred J. Menezes, Paul C. van Oorschotm and Scott A. Vanstone (CRC Press, 2001)
- ***O'Keefe*** – *Ethical Data and Information Management* by Katherine O'Keefe and Daragh O Brien (Kogan Page, 2018).
- ***McGilvray*** - *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information* ™ by Danette McGilvray (Morgan Kaufmann, 2008)
- ***English*** – Larry English, Improving Data Warehouse and Business Information Quality (John Wiley & Sons, 1999).

THREE STATES OF DIGITAL DATA



SECURING DATA IN USE

Data-In-Use:
information in CPU, RAM, registers, etc. for current processing and applications

Security approaches: full memory encryption, secure enclaves, isolated systems, homomorphic encryption

This sounds interesting…

SECURING DATA IN MOTION

Data-At-Motion:
*information moving across communications channels including **within a computer.***

Security approaches: encryption, entity authentication, key management and ephemeral keys, and conscientious governance

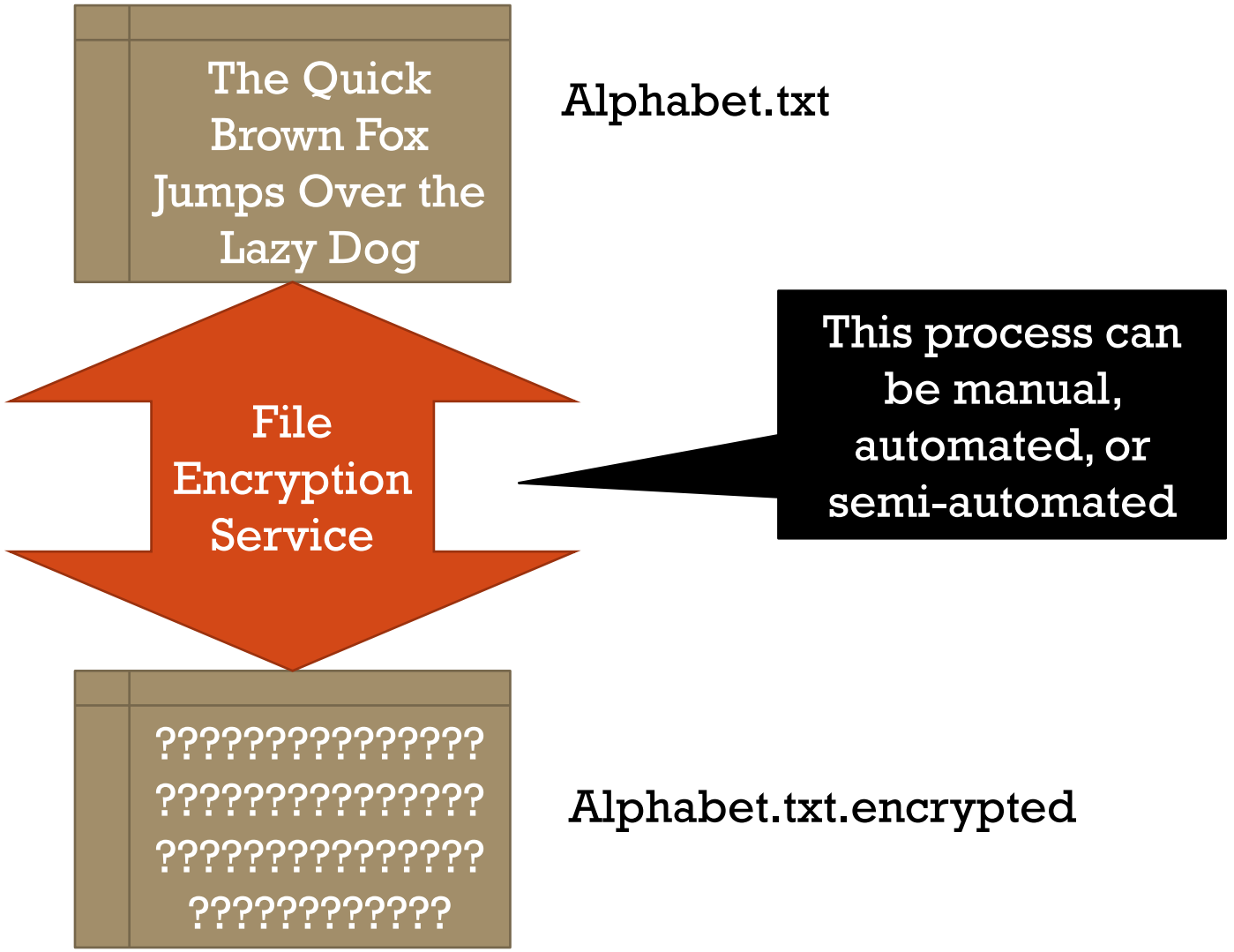
SECURING DATA AT REST

Data-At-Rest:
inert information stored on physical media such as disks, tapes, databases, etc.

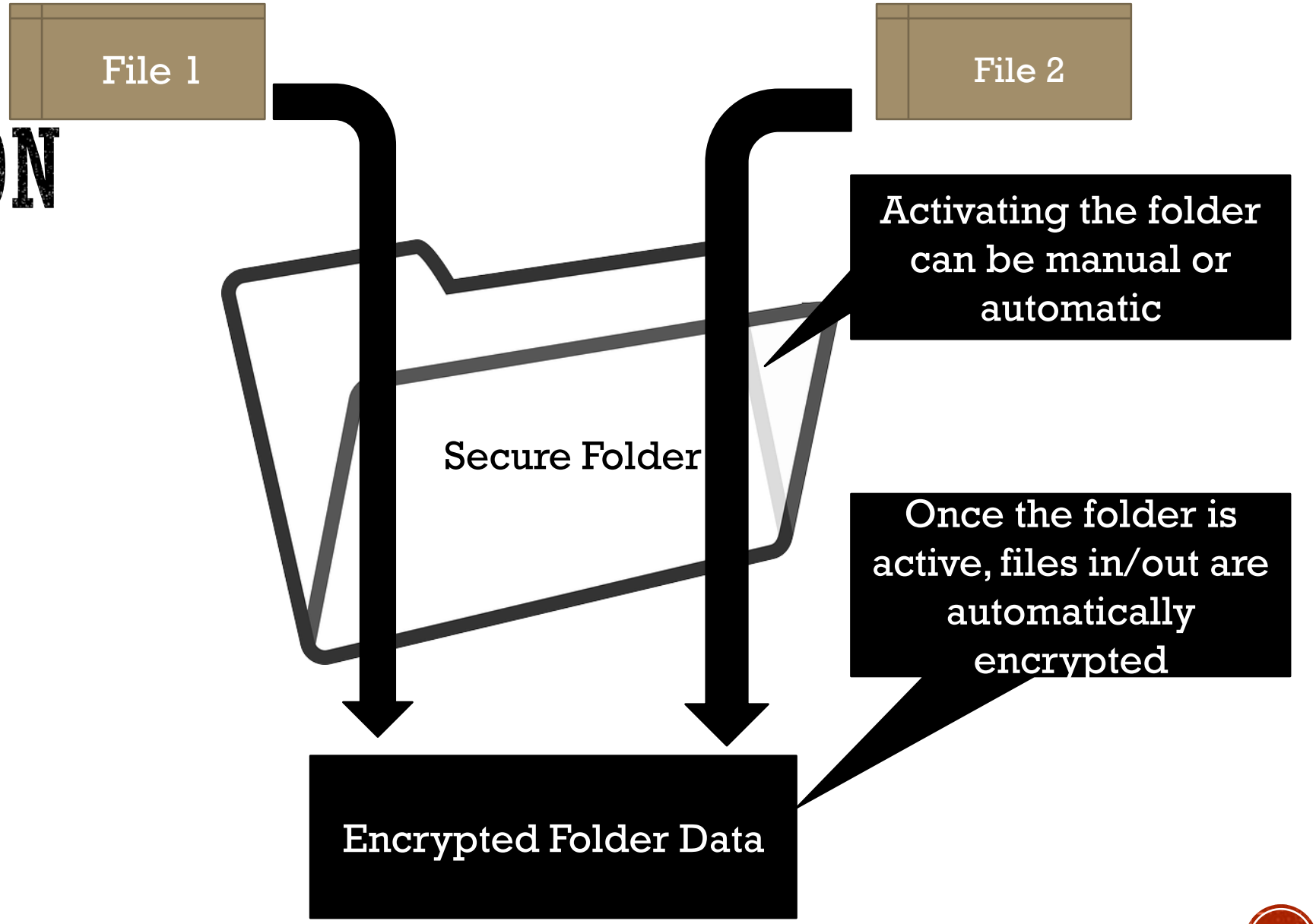
Security approaches: encryption, access controls, key management, audits, tokenization, and conscientious governance

Now this is more like it!

FILE ENCRYPTION

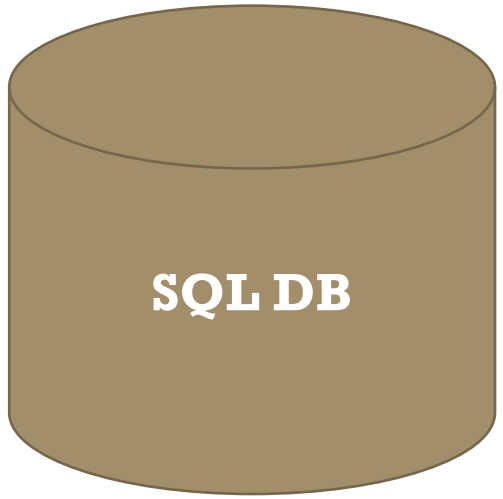


FOLDER ENCRYPTION



APPLICATION ENCRYPTION (E.G. DATABASE)

An application, like a MySQL DB, can be manually configured to store data encrypted*



INSERT INTO table… VALUES
(Alice, 1/1/1971, 1 Encryption Rd,
AES_ENCRYPT(555-55-5555, key))

SELECT AES_DECRYPT(SSN, key) as SSN
FROM table WHERE Name=Bob

Name	DOB	Address	SSN
Alice	1/1/1971	1 Encryption Rd.	????????????????
Bob	2/2/1972	2 Security Way	????????????????
…			

**Encrypted Data cannot be index, searched, etc.*

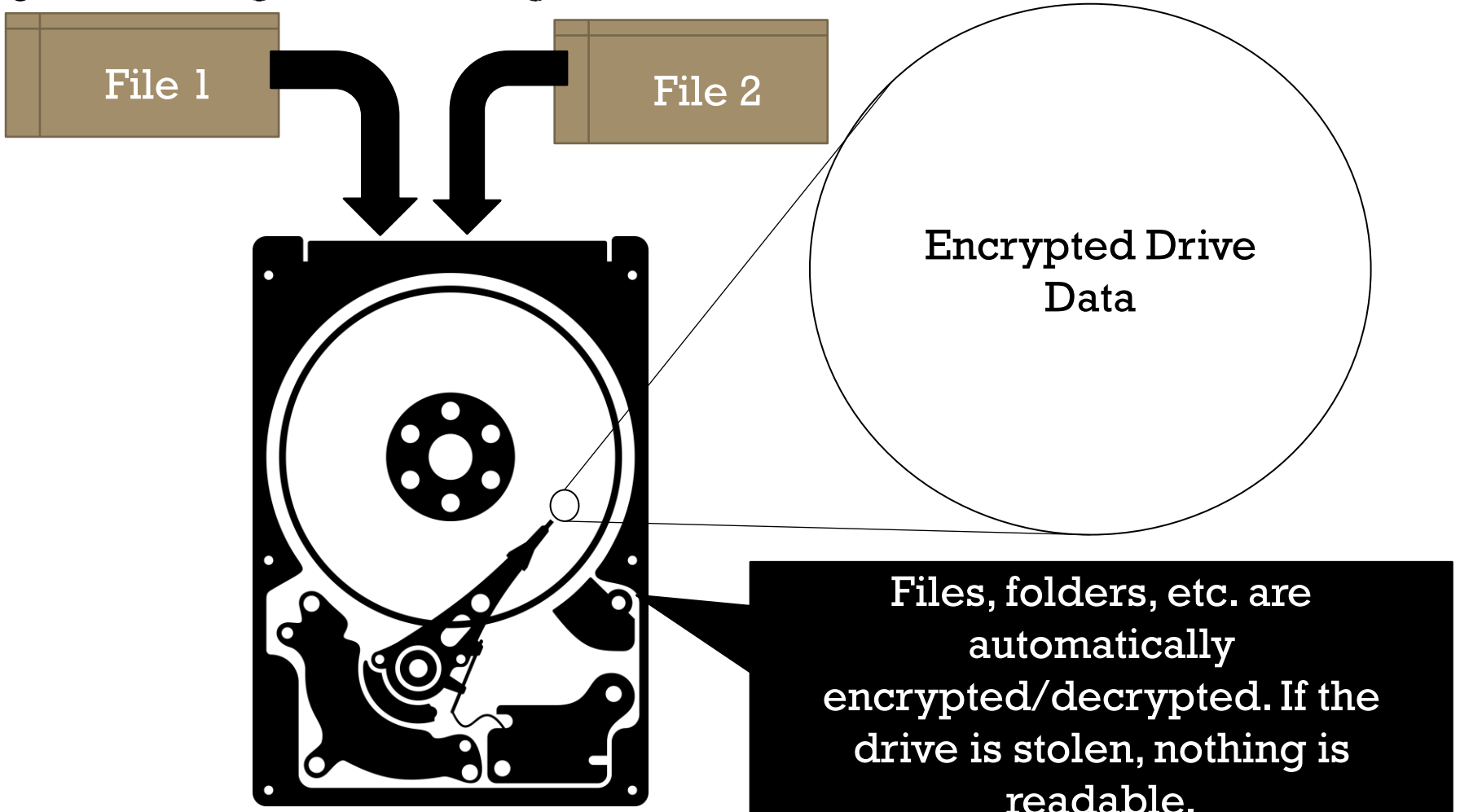
WHOLE DATABASE ENCRYPTION

- MySQL can also be configured with whole-db encryption
- Example: Enterprise Transparent Data Encryption (TDE)
- When enabled, data is automatically encrypted/decrypted
- This data can be searched, indexed, etc.
- The point is that application encryption varies widely

SIDE NOTE!!

- MySQL provides a number of options for AES_ENCRYPT
- ***MANY OF THEM ARE UNSAFE!***
- You should consult a cryptography expert before using!!!!

WHOLE DISK ENCRYPTION



OTHER DATA-AT-REST ENCRYPTION ISSUES

- Strong cipher (e.g., AES) with a large key size (256 bit)
- Should *Fail-Secure:* on failure, data remains encrypted

THE OTHER SECURITY CONCERNS

- Recall that encryption does not "create" security
- Other security components required to enforce policy
  - **access controls** – limiting who has access to data
  - **key management** – managing a key's lifecycle
  - **audits** – tracking crypto, access controls, keys over time

THE NEW WORLD OF BIG DATA, CLOUD STORAGE, ETC.

- The tech world has changed drastically within the last decade
- Companies are accelerating moving data resources to the cloud
- Big data is… well, ***big***. And technologies are changing to match
- New technologies are introducing new security challenges
- We'll talk about just two:
  - Data Lakes
  - Cloud Storage in General

DATA LAKE OVERVIEW

A *data lake* stores raw data, from wide input sources, into a single logical store. Using search and "big data" engines, it provides discovery, analytics, reporting, and so forth.

Alice's engineering team

Bob's Sales Office

Interconnected Storage Devices (raw, unformatted data)

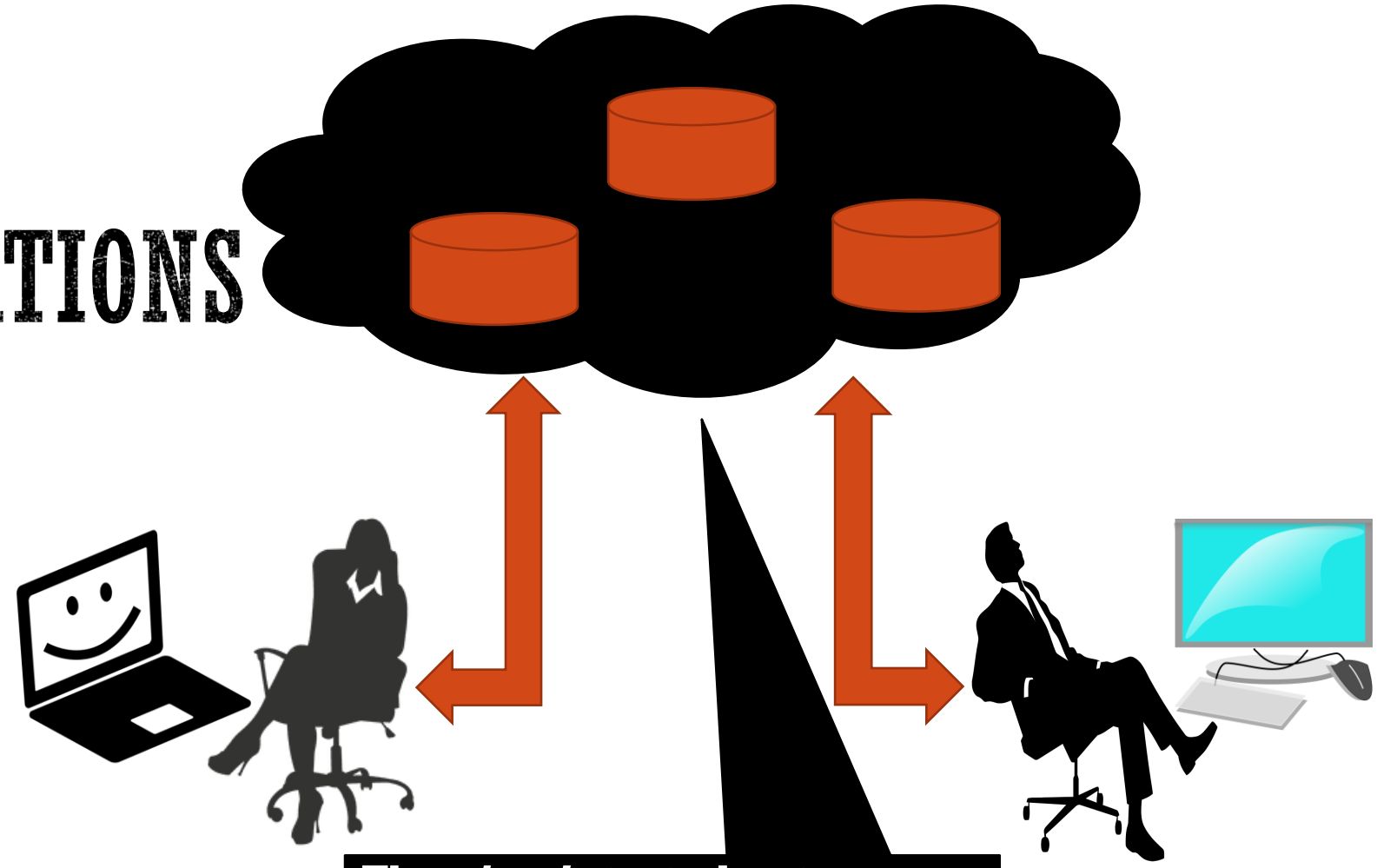
DATA LAKE SECURITY CHALLENGES

- Data stored together ***IN ALL THREE STATES!***
  - Where did data come from?
  - Who touched it?
  - Who is authorized to access it?
- Encryption questions abound, especially for processing
- Access control questions outside, *and inside*, the lake
  - Most of the advice I find is about outside access
  - But a "Data Lake" is a concept on top of hardware. Who has access?
- ***Some experts recommend not storing PII in Data Lakes!***

THE POINT

- I'm not criticizing Data Lakes
- But when tech changes, security implications change too
- Note: poorly used Data Lakes are called Data Swamps
- Can data in a Data Swamp be properly secured?

CLOUD CONSIDERATIONS

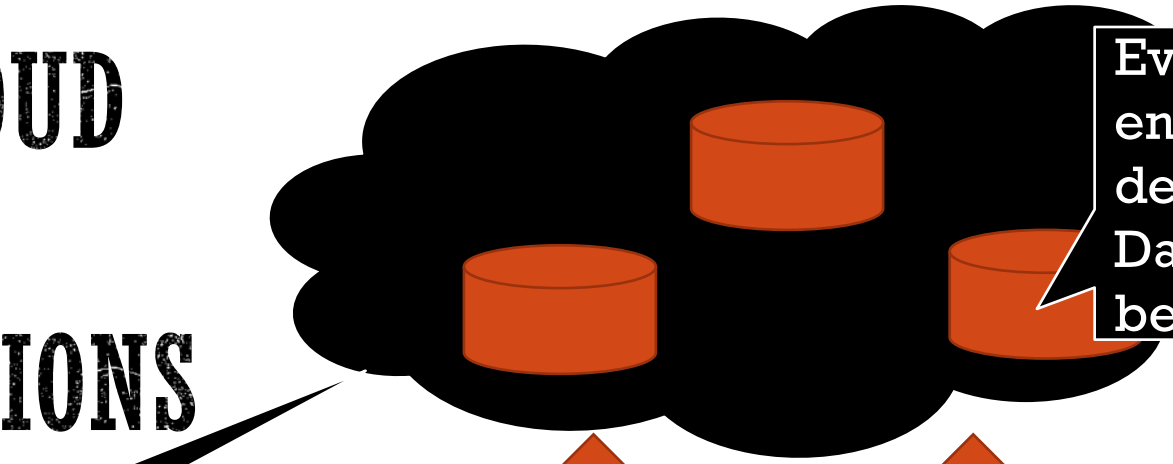


The *cloud* started out as purely storage, but now is used for processing and entire enterprise infrastructures.

CLOUD SECURITY MINDSET CONCERNS

- Clouds ***can*** be more secure than many enterprises
- However incentivize users to ***stop thinking about security***
- A CEO told me: "We're secure. ***We use the cloud***."
- There is no security free lunch

SAMPLE CLOUD SECURITY CONSIDERATIONS



Even when data is encrypted, it is often decrypted for processing. Data-in-use security can be a concern.

Some cloud vendors don't encrypt the data-at-rest. But for those who do, the enterprise still needs a key management architecture.

User Access Control Policies are still determined by the enterprise, not the cloud provider.

Privacy requirements, regulatory burdens, and accountability remain with the enterprise.

CLOUD USABILITY/ SECURITY TRADE-OFF

- There is fundamental tension between usability & security
- For example, take big data.
  - Bigger data sets
  - Search and analysis engines reveal valuable insights
  - Hence, why data is the "oil of the 21st century"
- Problem: data must (usually) be decrypted for analysis
- This is a potential security, privacy, regulatory hazard!

THE GMAIL EXAMPLE

- I still use Google mail for personal and business
- I do not end-to-end encrypt my mail
  - It is encrypted "at rest" on Google servers
  - But it is un-encrypted and analyzed by Gmail search servers
- I could use ***proton mail*** for end-to-end security but I don't.
- Why? Because I've come to rely on Gmail search.
  - I'm not sure I could function without this search capability
  - Unfortunately, I have to trust Google with my data for this

DATA PRIVACY

Data Privacy:
is the relationship between dissemination of data and the gathering/use/management thereof. It includes legal, policy, and technical issues.

For our class, we will only talk about the technology issues.

Who cares?

WHY IT MATTERS

Data Privacy:
is the relationship between dissemination of data and the gathering/use/management thereof. It includes legal, policy, and technical issues.

For our class, we will only talk about the technology issues.

People care because data collected about them could be used to manipulate, rob, embarrass, blackmail, or even control them.

DATA AS A MEANS OF CONTROL

Yes, control. Some experts are concerned that genetics might be used to control where you live or go to school.

Where you live? Go to school? How?

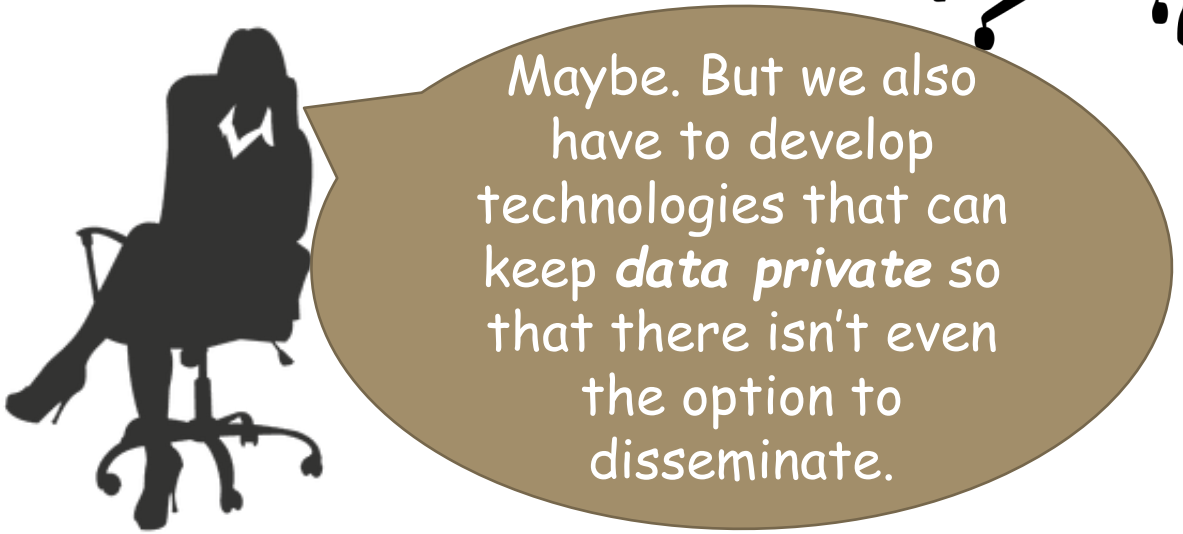
A condo association forces you to submit to a DNA test. If you have a predisposition to Alzheimer's disease, you can't live there.

That's terrible!

And right now, some feel Big Data is being used to exploit individuals with addiction issues. Is that "control" or just "manipulation"?

Ugh! It doesn't matter!

TECHNOLOGY VS LAW VS POLICY



DATA OWNERSHIP VS STEWARDSHIP

- Who "owns" data about you? (legal/political question)
- In Europe, laws require that you own the data about you
- In the United States, laws are generally moving towards this
- One who handles data for another: a ***data steward****

* ***McGilvray,*** pp. 53-54
O'Keefe, pp. 102-105, 236-244

DATA PRIVACY TECHNOLOGY GOALS*

- Enable identification of ownership and stewardship of data
- Enable owners to maintain policy for their own data
- Enable stewards to communicate data handling to owners
- Enable data handling by a steward to adhere to owner policy
- Enable permitted data handling to expose minimal privacy risk
- Enable accountability of data stewards to data owners
- Enable transparency of data, handling, stewardship to owners

*Others have expressed similar goals differently.
These are Dr. Nielson's formulations.

PERSONAL INFORMATION/PII

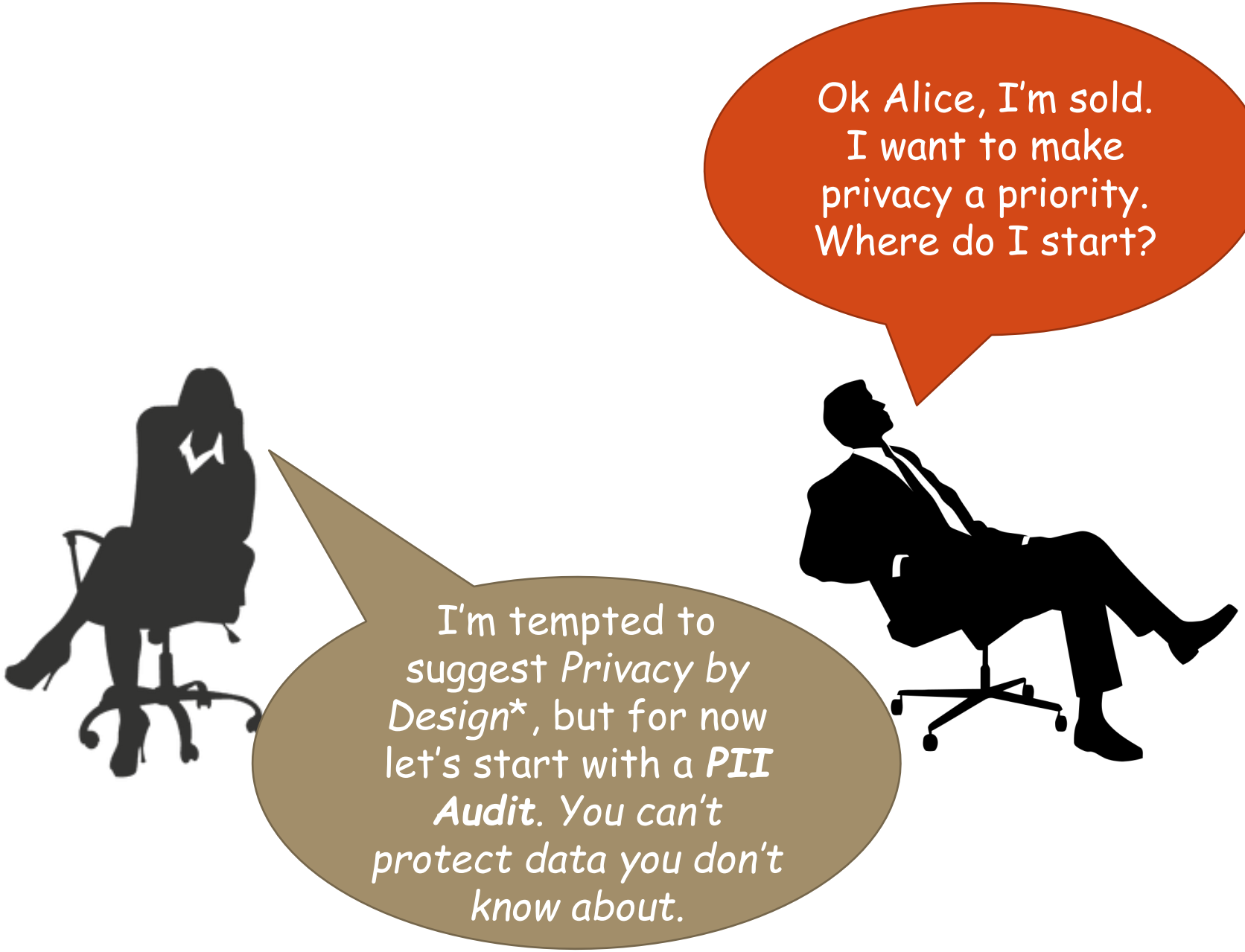
- Data privacy begins with identifying "personal" data
- The defined set of personal data varies by legal jurisdiction
- Example: in Europe an IP addr is personal but not in the US

PII IN THE UNITED STATES

"*any information about an individual maintained by an agency, including (1) any information that can be used to* ***distinguish or trace an individual's identity****, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2)* ***any other information that is linked or linkable to an individual****, such as medical, educational, financial, and employment information.*"

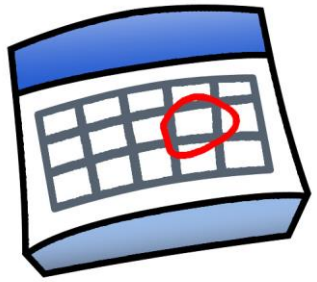
(NIST Special Publication 800-122, emphasis added)

PII AUDIT



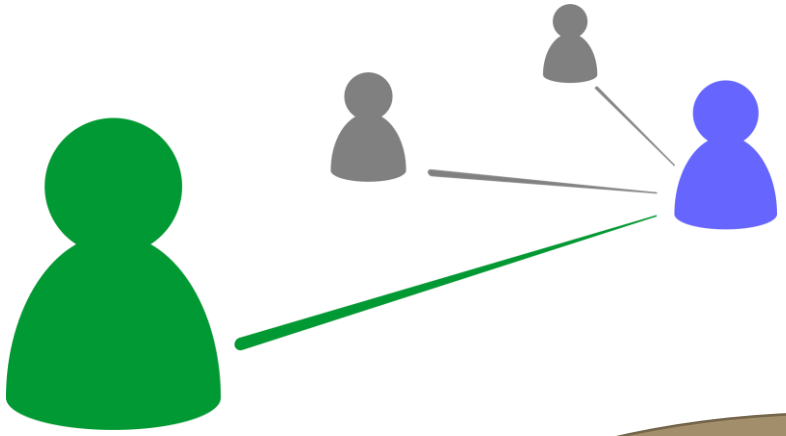
* See ***O'Keefe***, pp. 259-260, 265

DISTINGUISHING DATA



Obviously, any data that directly identifies someone is PII. NIST calls this "distinguishing" data. It includes name, DOB and even biometrics.

TRACING DATA



PII also includes data that could be used to determine an individual's activities or status. This includes ***log files*** or camera recordings.

LINKING DATA (LINKED OR LINKABLE)

This data is already linked to the student

Student	Grade	Height	Weight
Bob Jr.	5	4’5”	100lbs

This data could be linked to the student

School Uniforms Online Store

Grade	Height	Weight	Purchases
5	4’5”	100lbs	$100.00

“Linked” data is data already linked to the person. “Linkable data” is data that *could* be linked to the person.

PII AUDIT SOLUTIONS

That's a lot of PII! How can I find all of it?

Start by reading NIST SP 800-122. It has a number of good starting points. If you're doing business in Europe, you may need a GDPR specialist.

PII SAFEGUARDS*

- Privacy-Specific Safeguards
  - Minimizing the Use, Collection, and Retention of PII
  - De-Identifying Information
  - Anonymizing Information
- Security Controls
  - Access Enforcement
  - Auditable Events
  - Information System Monitoring
  - Media Sanitization

* This is a subset of safeguards described in NIST SP 800-122

MINIMIZING PII USE, COLLECTION, AND RETENTION

The first, ***and perhaps most important step***, in protecting PII is ***NOT TO COLLECT, USE, and/or RETAIN*** it.

That's impossible!

It very well might be! If you're a medical records company, you have to. But start with this mindset.

We're a data company. Processing data is what we do.

Start by remembering it isn't ***your*** data. You should only have data you absolutely need.

MINIMIZING PII USE, COLLECTION, AND RETENTION (2)

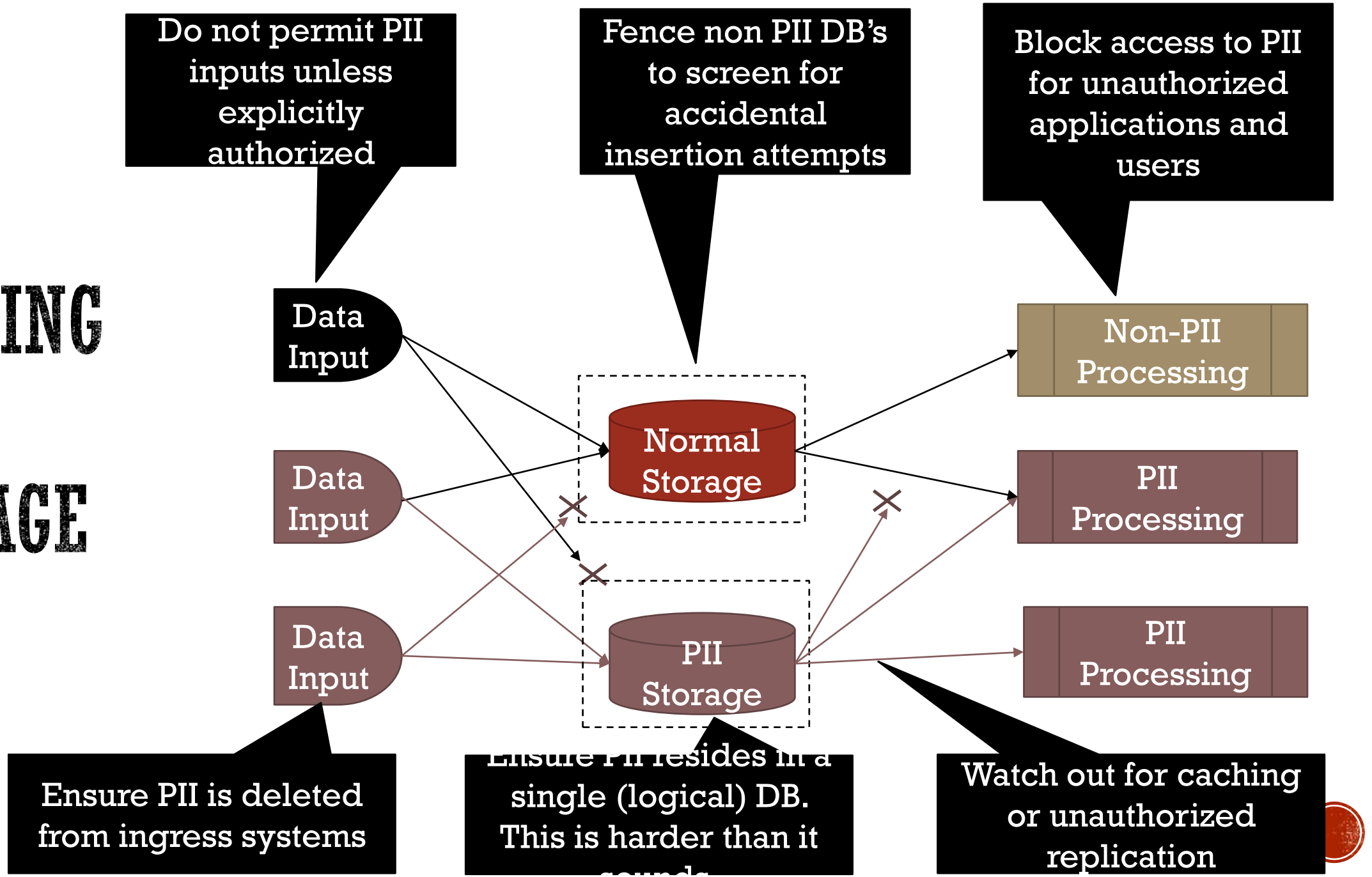
If you absolutely must have PII data, minimize all engagement with the data.

Limit the roles that can access the data, limit the users assigned to those roles. Limit PII processing and limit where it is stored.

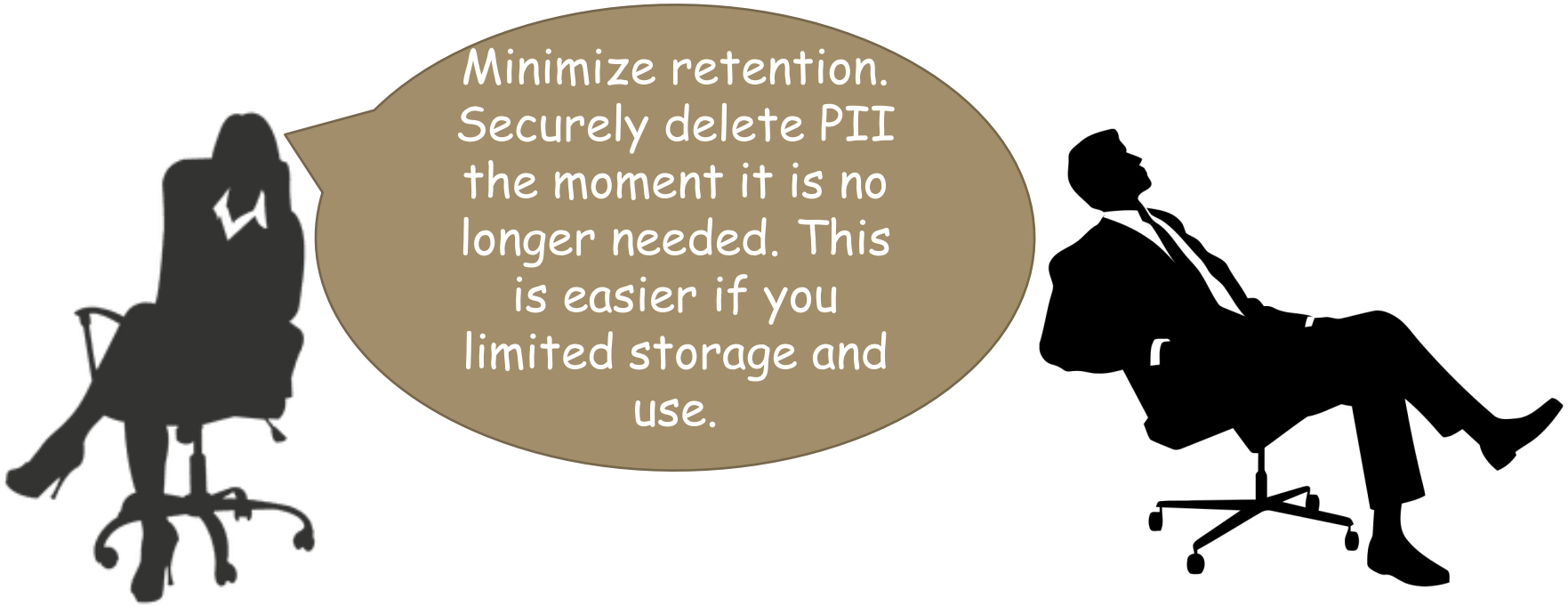
How do you limit where it is stored?

Meaning?

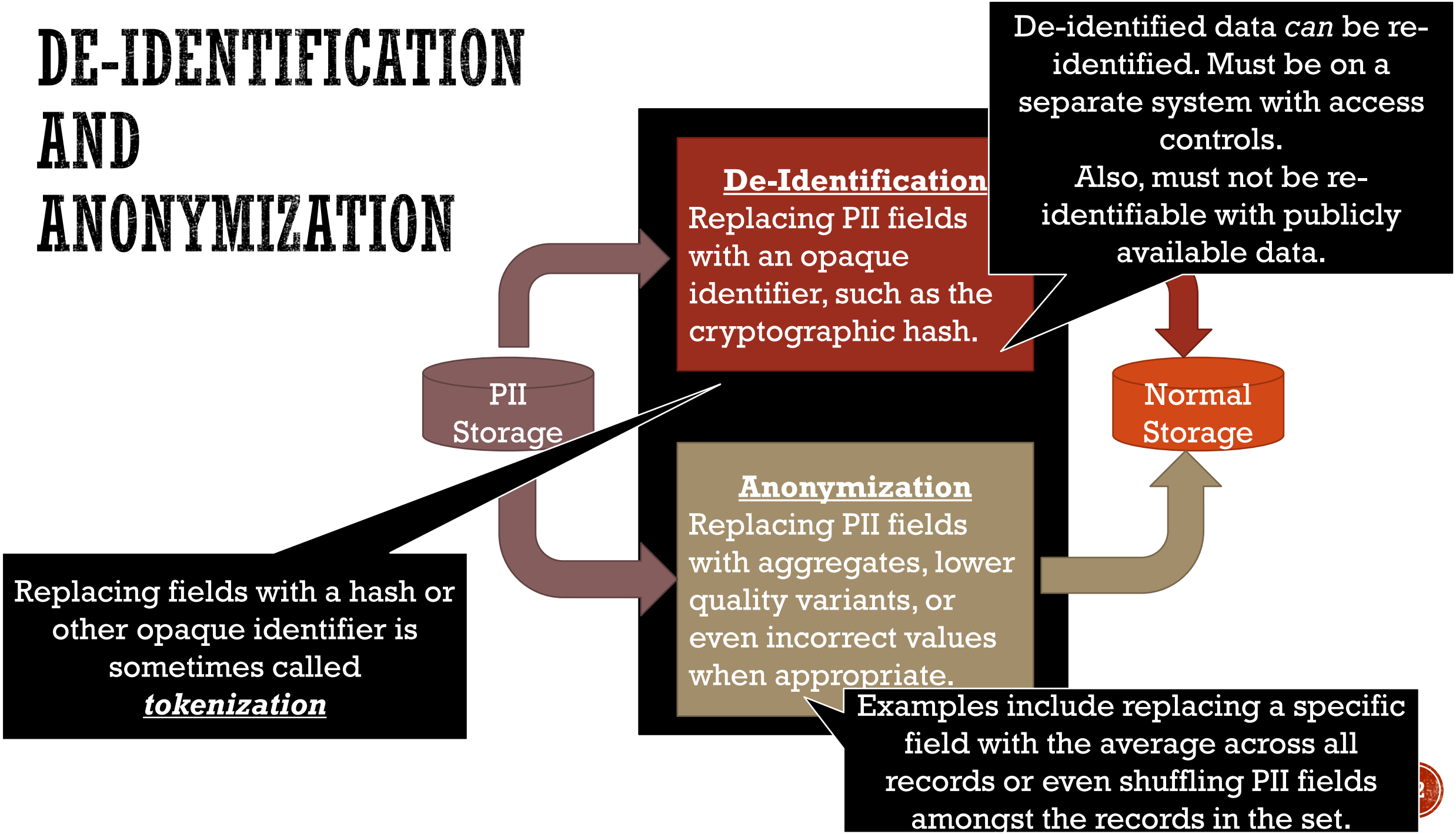
LIMITING PII STORAGE



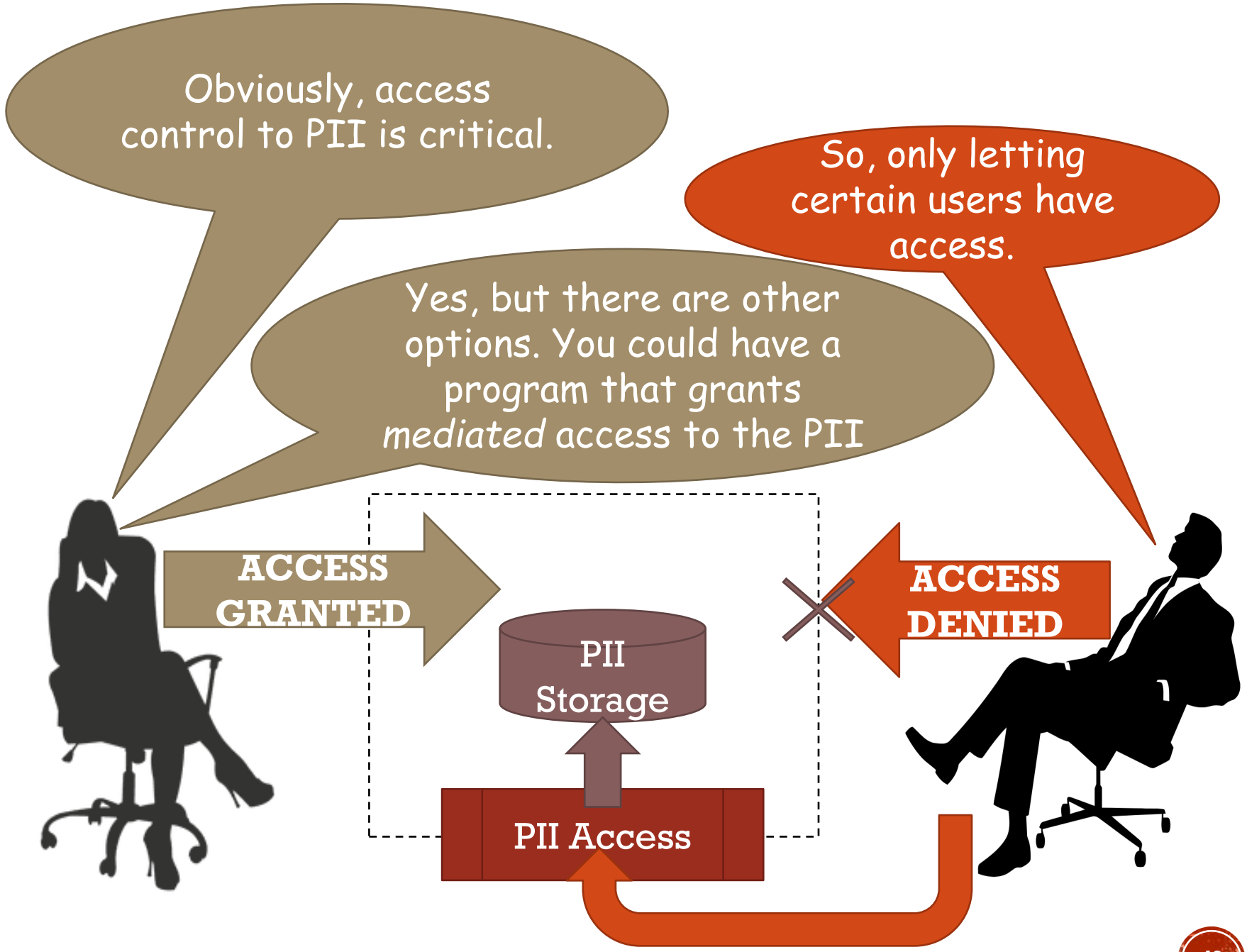
MINIMIZING PII USE, COLLECTION, AND RETENTION (3)



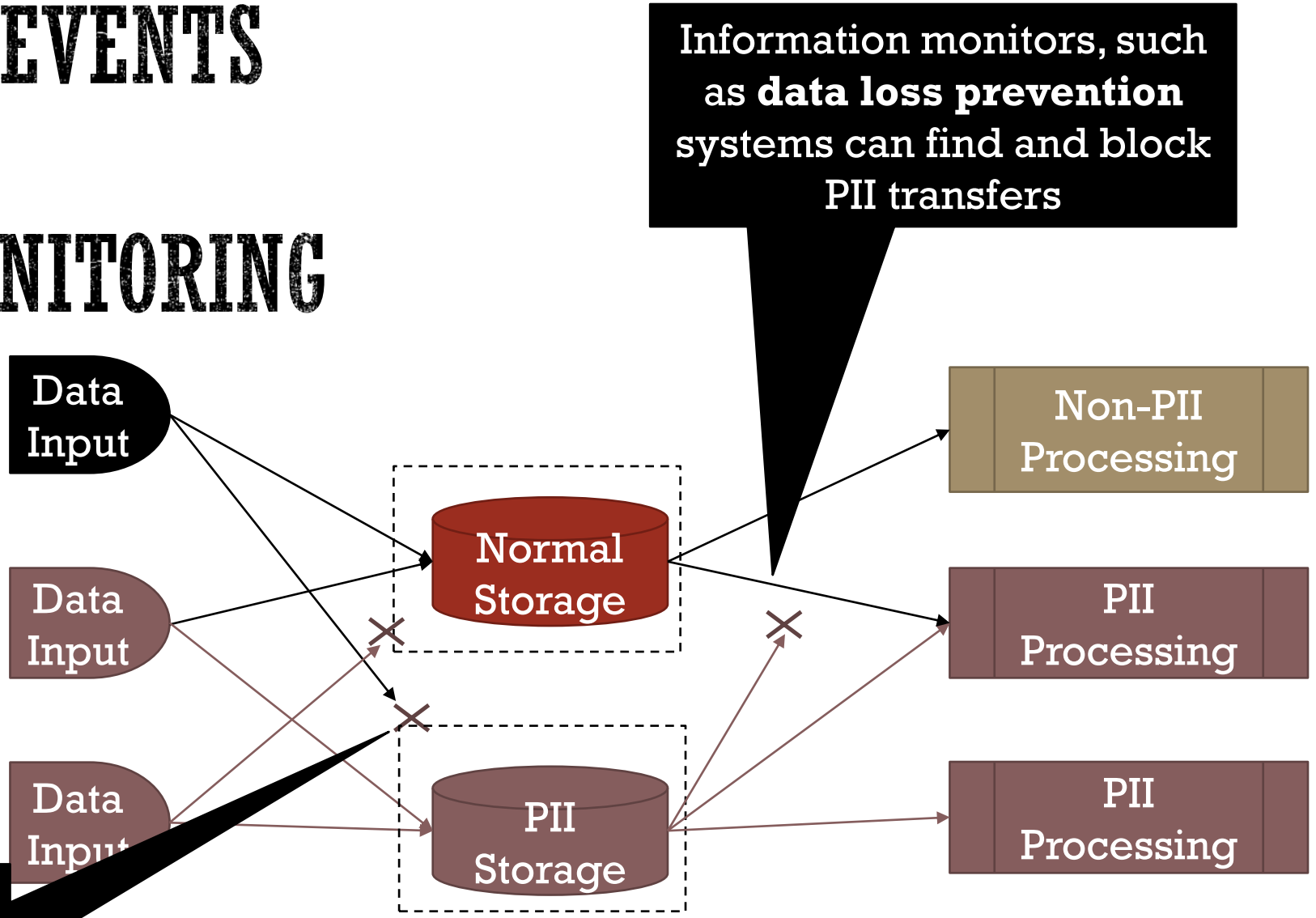
DE-IDENTIFICATION AND ANONYMIZATION



ACCESS CONTROLS



AUDITABLE EVENTS AND SYSTEM MONITORING



Any PII activity that violates policy, ***even if unsuccessful***, should be audited.

SECURITY AND PRIVACY SUMMARY

- We've covered a lot of ground for both security and privacy.
- One point that should be clear: both are complex subjects
- Your organization may need an SME to help you navigate

- But, as the data person, ***you*** hold the keys to the most critical part!