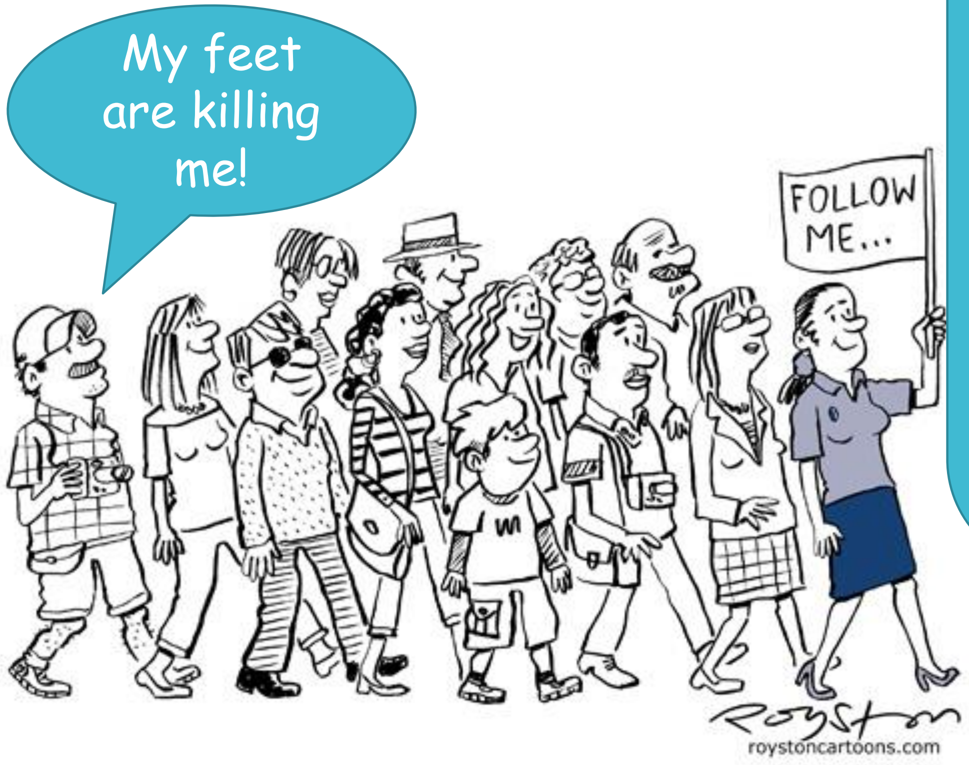
Data Security and Privacy Basics

Network Security Fall 2019

Seth James Nielson

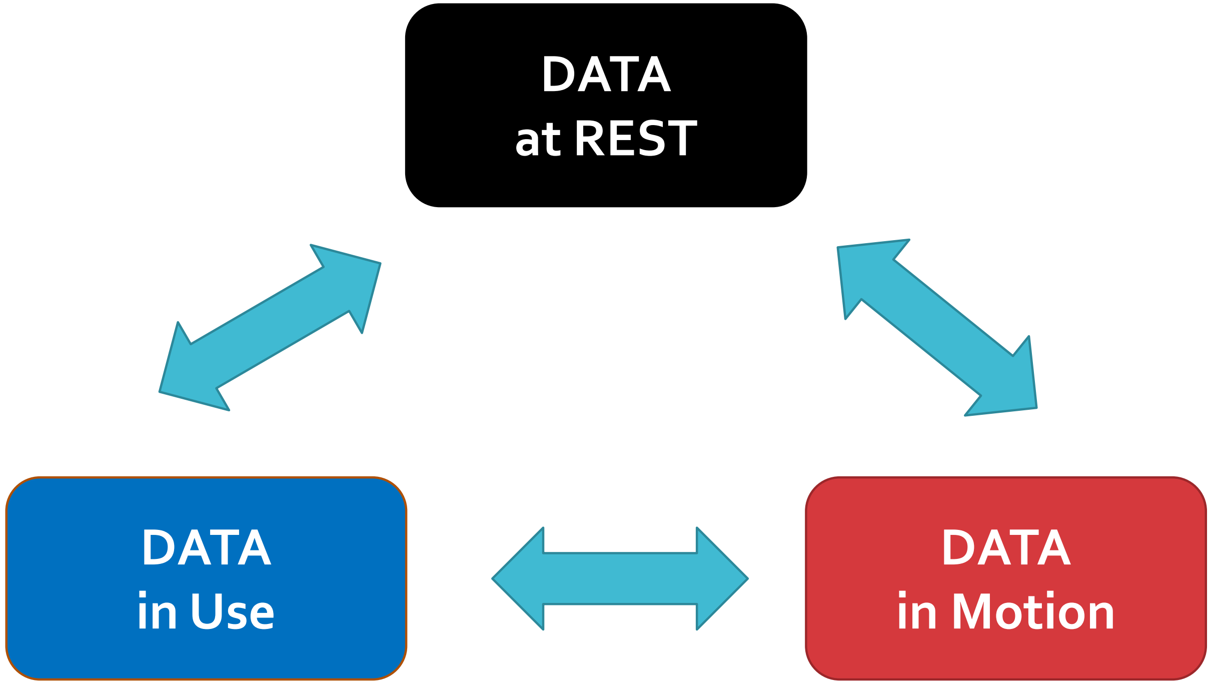
Part I
Data Security Basics



This Photo by Unknown Author is licensed under CC BY-SA-NC

- Data in Use (brief)
- Data in Motion
- Data at Rest

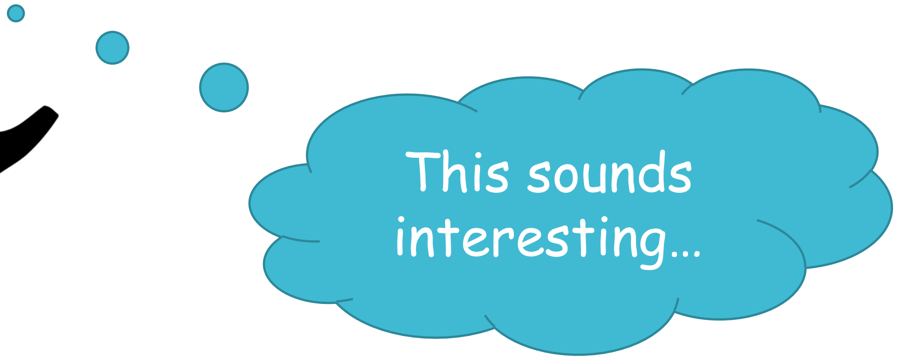
Three States of Digital Data



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



Skipping this idea…

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

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

We'll come back to this later.

Today's a tight lecture anyway. Data-in-Use is easiest to explain with examples that will come up later.

Why?

Securing Data in Motion

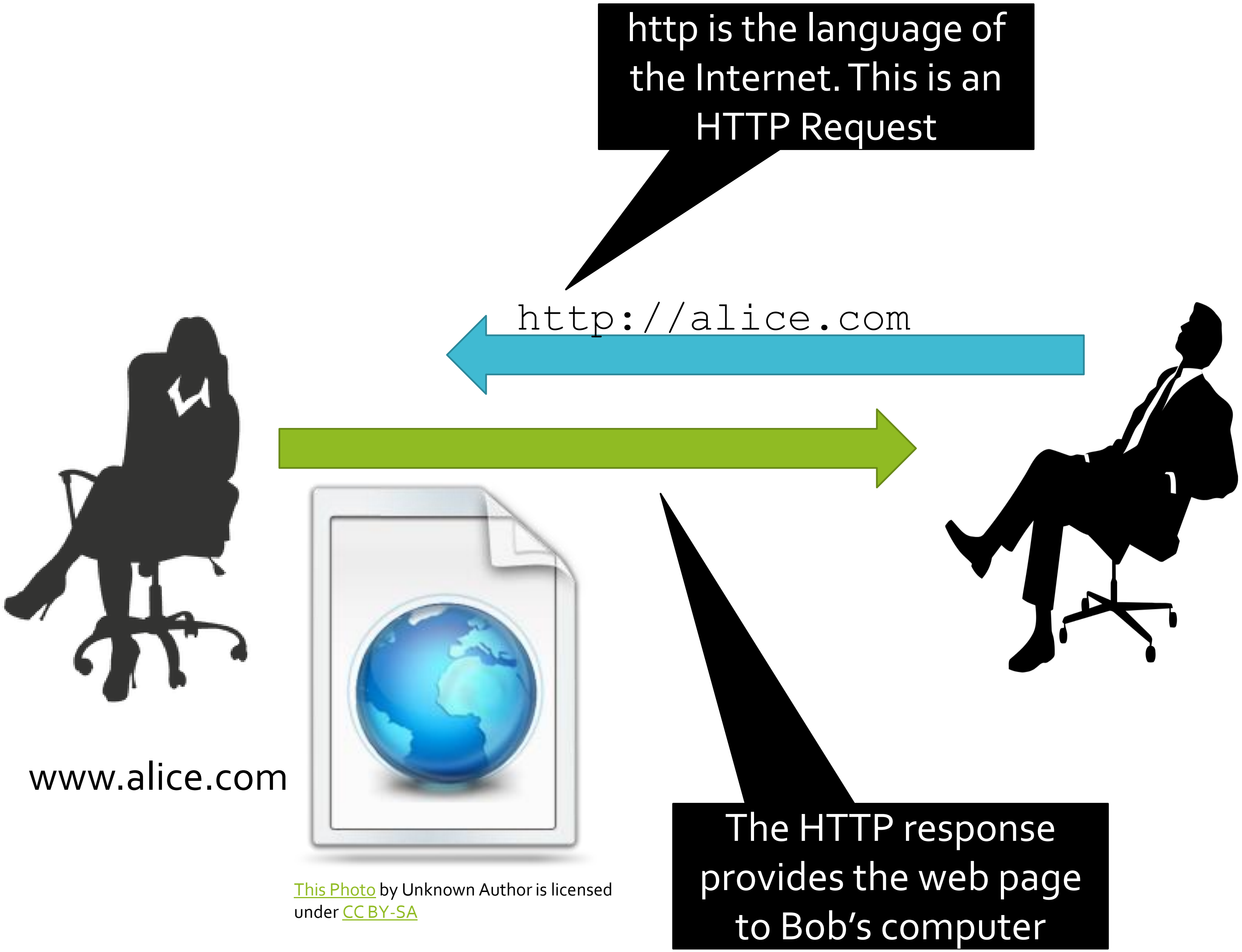
Data-At-Motion:
information moving across communications channels including ***within a computer.***

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

The Origin of Data Security

- Almost all of data security was historically for data in motion
- Nowadays, the “bad guys” would prefer to steal bulk data in bulk
- For this audience Data-at-Rest is probably more interesting too
- But securing data-in-motion is still important; let’s discuss TLS
- This will illustrate a lot of data-in-motion issues

Standard HTTP Messages

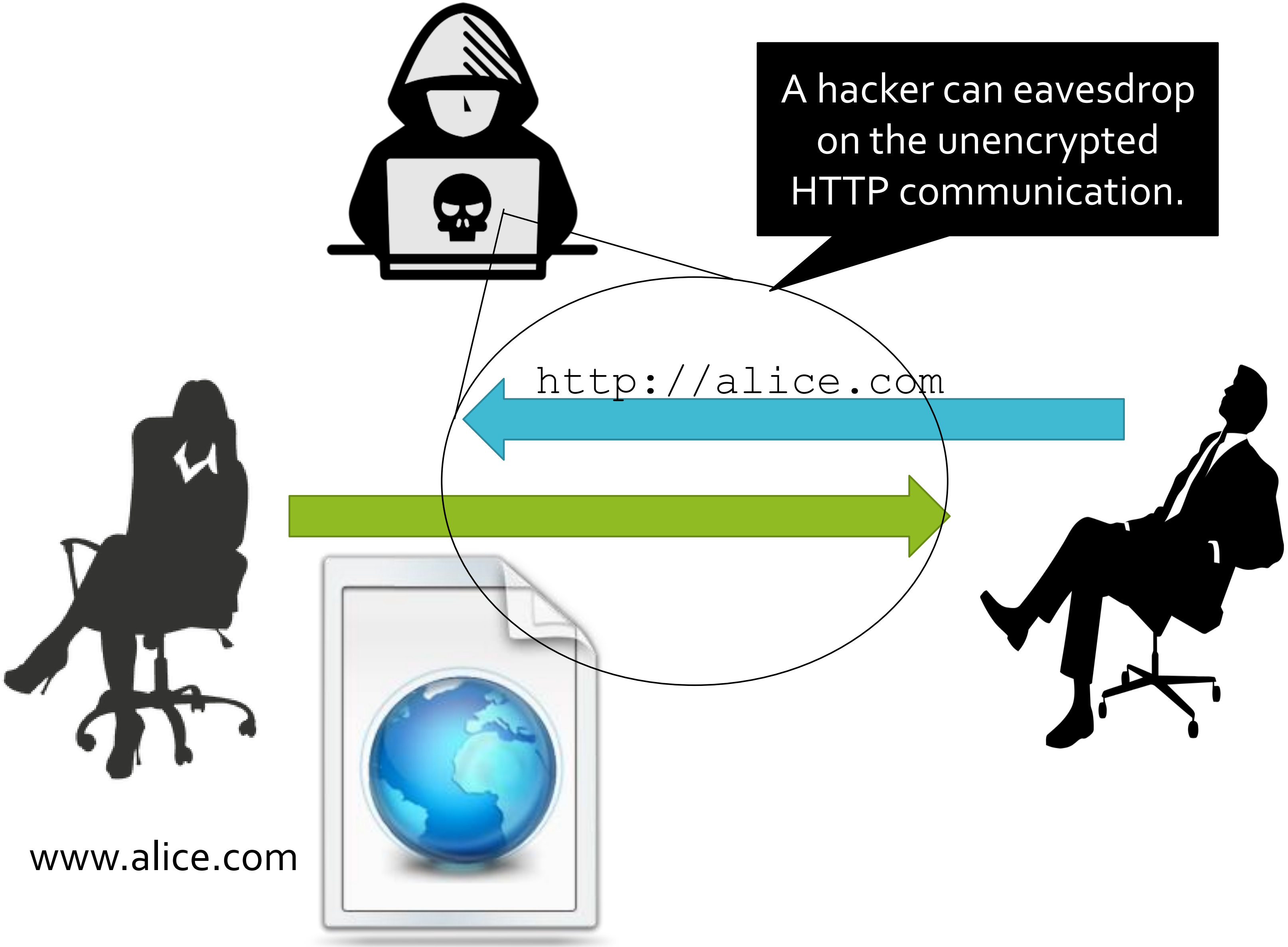


http is the language of the Internet. This is an HTTP Request

The HTTP response provides the web page to Bob's computer

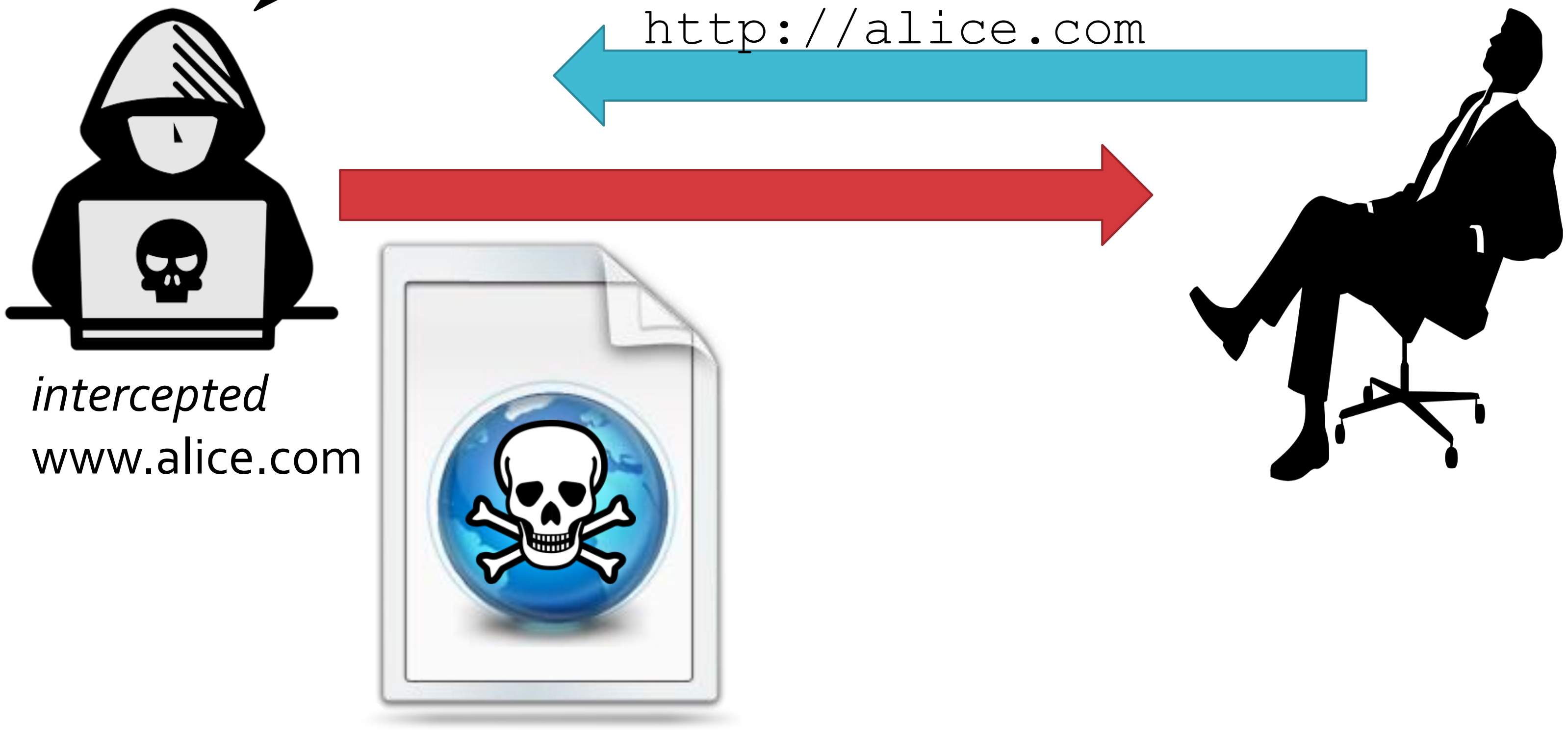
This Photo by Unknown Author is licensed under CC BY-SA

Unencrypted Channels


A hacker can eavesdrop on the unencrypted HTTP communication.
http://alice.com
www.alice.com

Un-authenticated Channels

Worse, a hacker could intercept the request and then impersonate Alice.



http://alice.com

intercepted
www.alice.com

Minimum Data-in-Motion Security

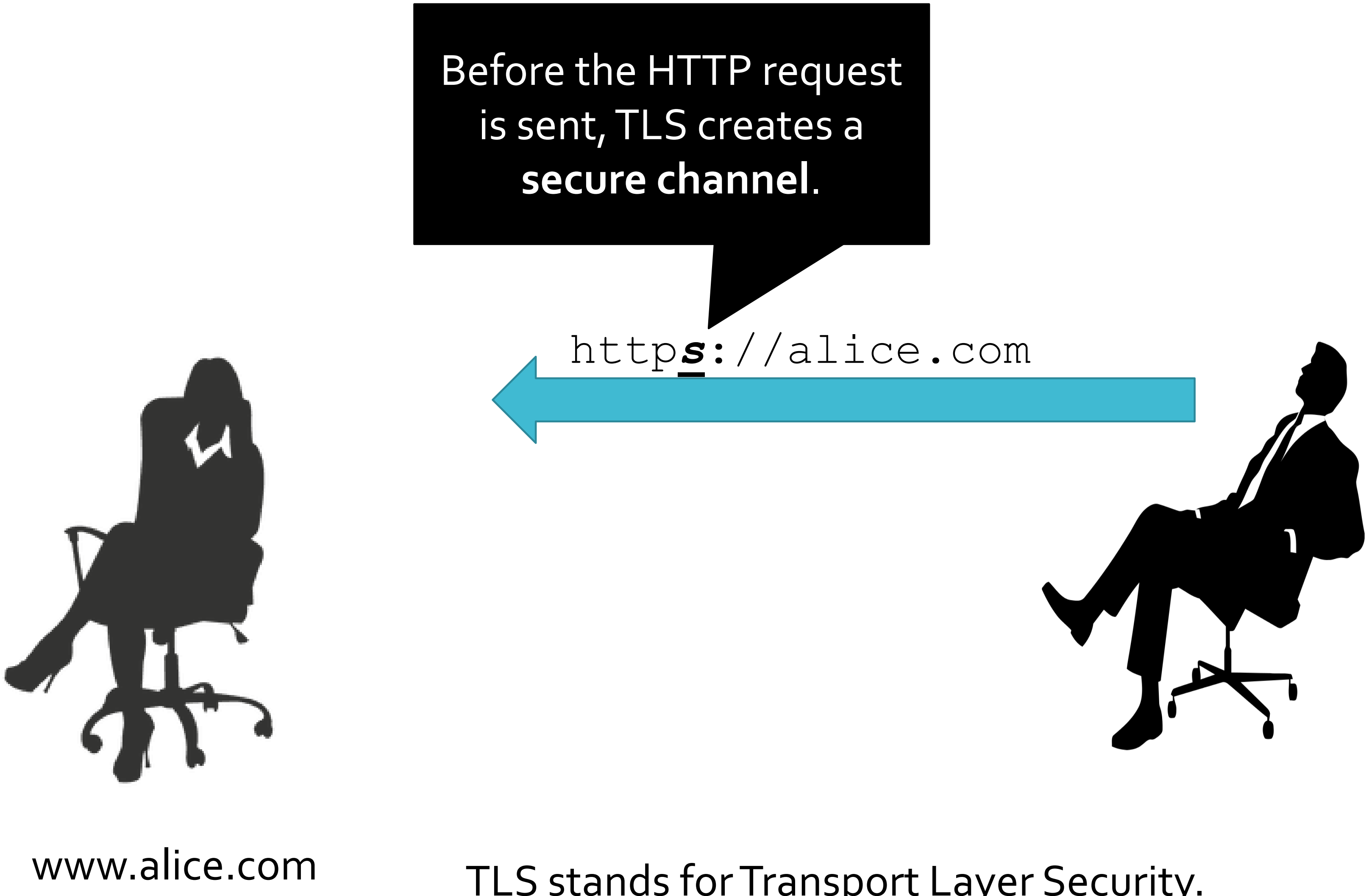
Entity Authentication

Confidentiality

- Bob needs to know he's talking to the real Alice,
- Bob needs to know his communications with Alice are secret
- Bob needs to know that communications can't be altered

Data-origin Authentication

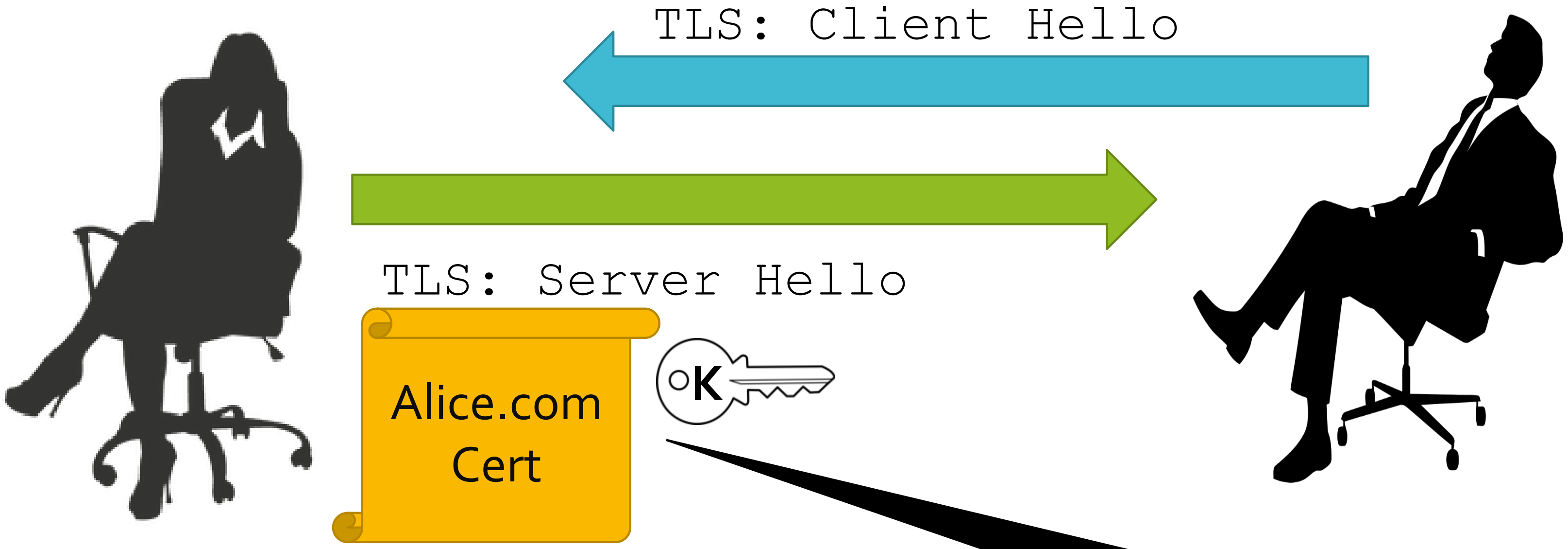
HTTPS:
HTTP over TLS



Before the HTTP request is sent, TLS creates a **secure channel**.

TLS stands for Transport Layer Security.
It replaced SSL, Secure Socket Layer,
although that name is still used.

TLS*: Start with a Firm Handshake



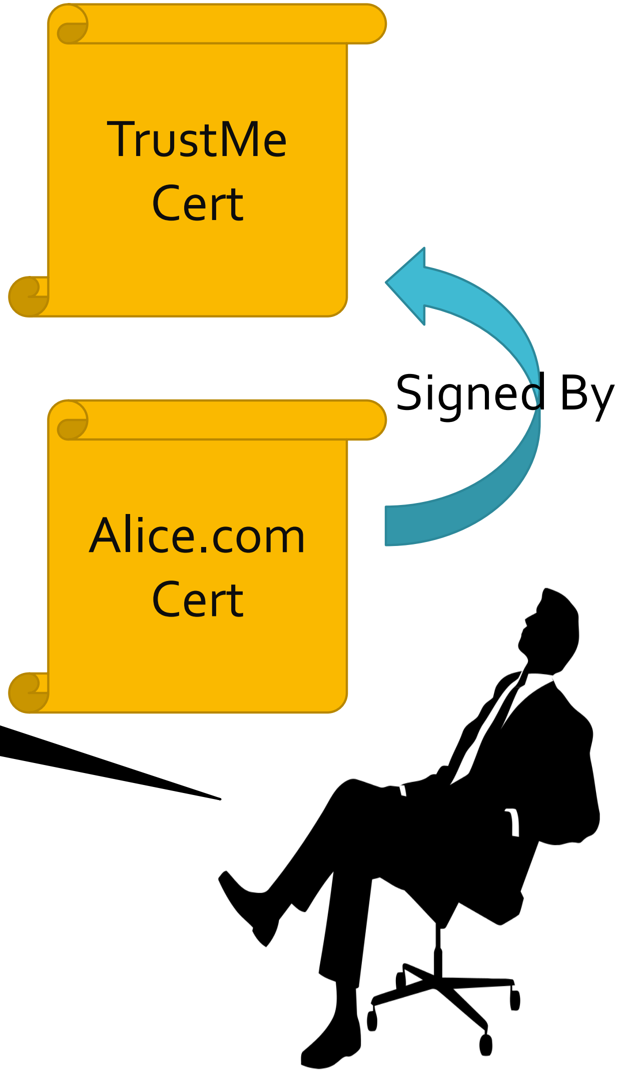
www.alice.com

Alice sends her ***certificate*** and an ***asymmetric key*** that will be used to *create* a temporary symmetric key.

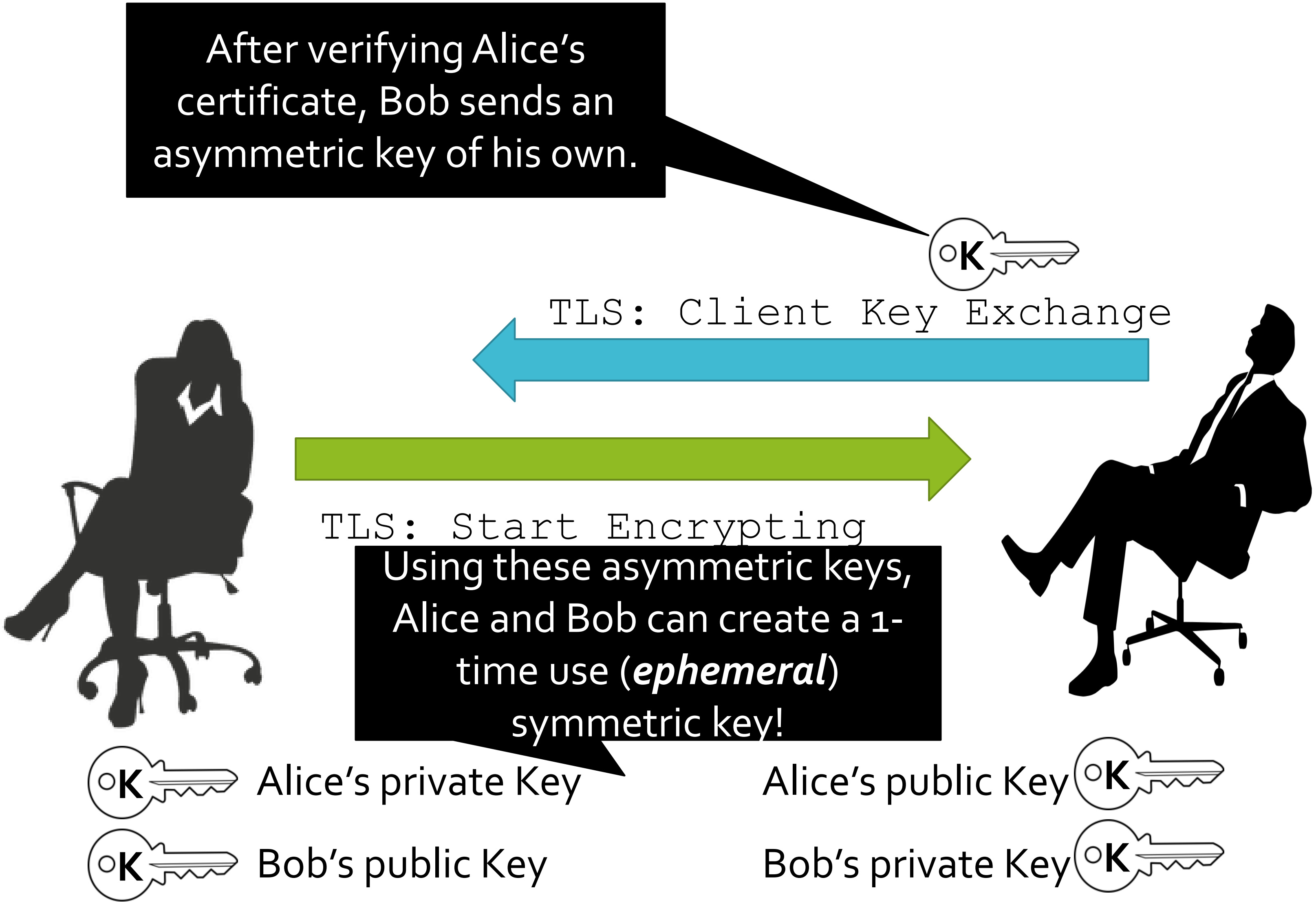
*For simplicity, these TLS examples relate to version 1.2

TLS: Stranger Danger

Bob ***verifies*** *the certificate by checking if it is signed by someone he trusts and has the right data (i.e., for Alice)*



TLS: Get a Temporary Key



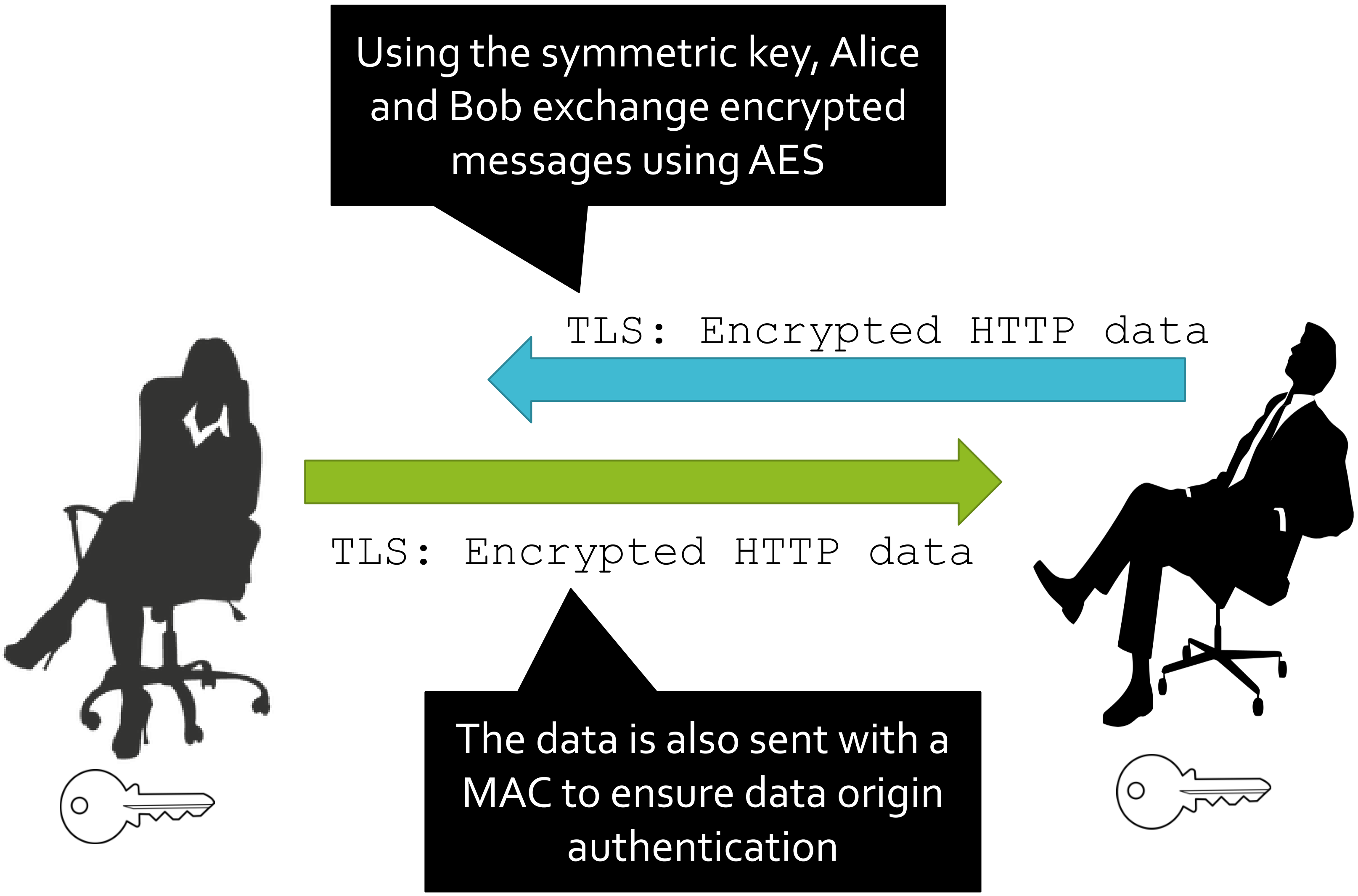
After verifying Alice's certificate, Bob sends an asymmetric key of his own.

`TLS: Client Key Exchange`

`TLS: Start Encrypting`

Using these asymmetric keys, Alice and Bob can create a 1-time use (***ephemeral***) symmetric key!

TLS:
Bulk Data



Using the symmetric key, Alice and Bob exchange encrypted messages using AES

TLS: Encrypted HTTP data

TLS: Encrypted HTTP data

The data is also sent with a MAC to ensure data origin authentication

Other Data-in-Motion Issues

- Data-in-Motion shows up in:
    - A single system when data moves from the hard drive to RAM
    - An enterprise system as data flows between systems
    - A system made up of multiple computers (e.g., **data lake**)
    - One enterprise system to another
- TLS provides an overview and an intro to data-in-motion issues
- There are other security concerns, of course
- Many of these concerns show up in data-at-rest
- So for now, let's move on
- (We'll see some data-in-motion examples near the end of class!)

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

Skipping this Idea too…

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

We aren't going to talk much about this today.

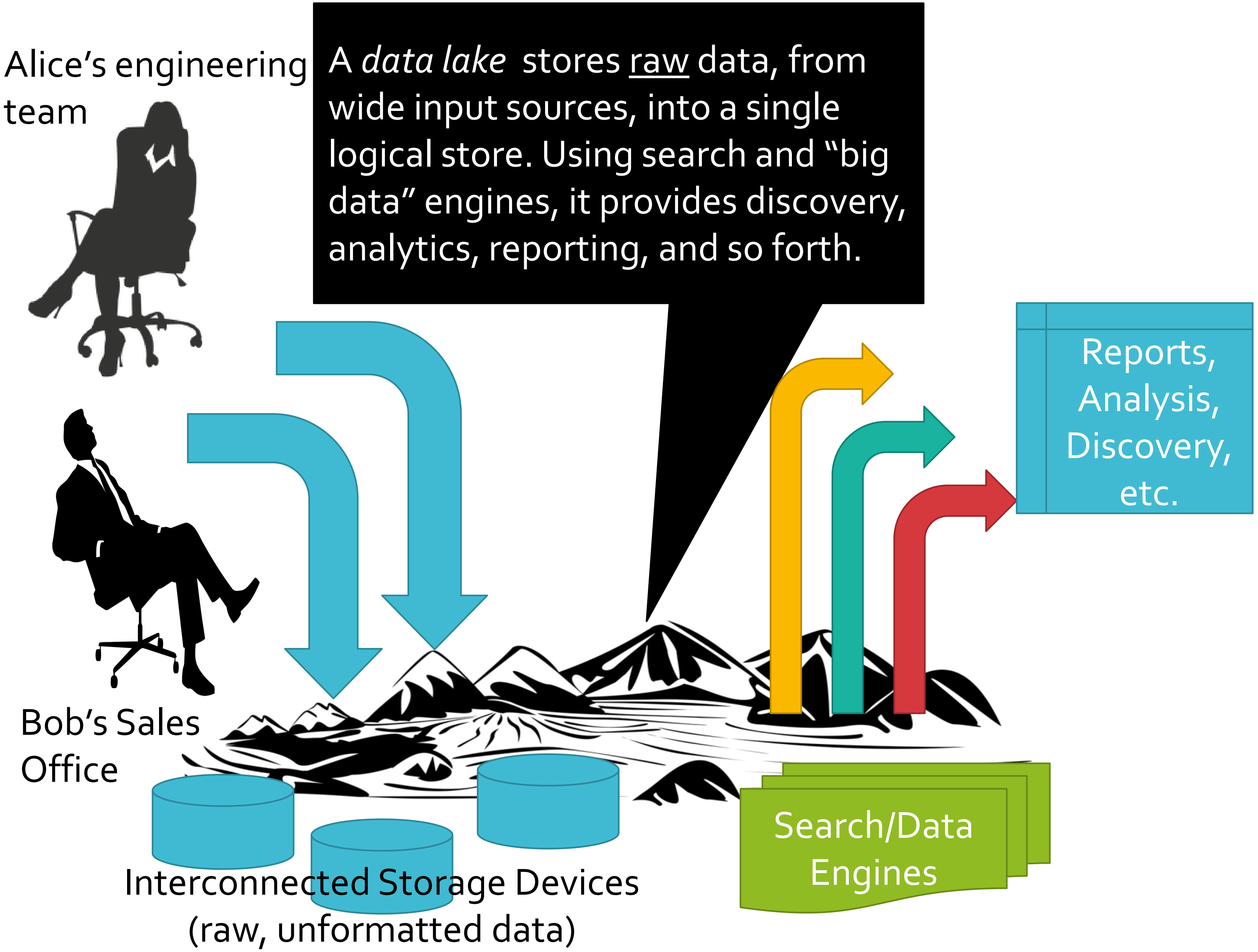
Data in Motion isn't as applicable to "Network" Security

Why not?!

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
- For example, “Data Lakes” have to protect data in all 3 states!

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)
Search/Data Engines
Reports, Analysis, Discovery, etc.

Data Lake Security Challenges

- Wide variety of data stored together.
  - Where did data come from?
  - Who touched it?
  - Who is authorized to access it?
- ***All three states of data!*** (rest, motion, use)
- 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 data experts recommend not storing PII in the Data Lake!***

The Gmail Example

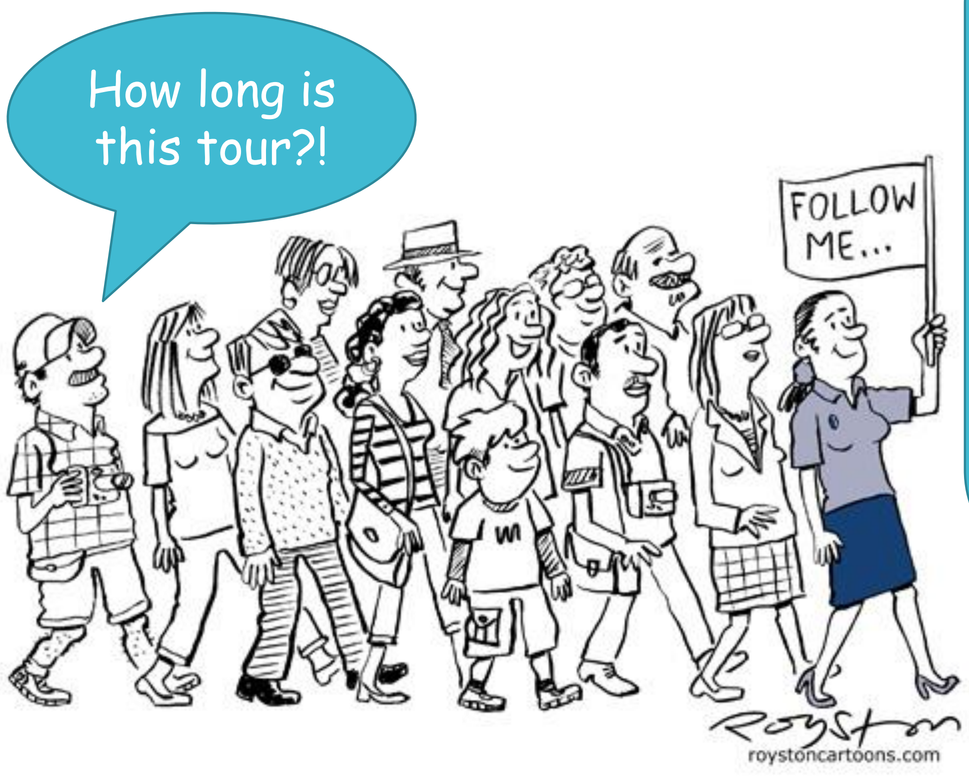
- I still use Gmail for personal email, and Google for my business
- I do not end-to-end encrypt my mail
    - It is encrypted “at rest” on Gmail servers
    - But it is un-encrypted and analyzed by Gmail search servers
- I could use ***proton mail*** for completely secure email, but I don’t.
- Why? 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

Part II Summary

- We’ve talked primarily about securing Data-in-Motion
- But all three states matter for network security at least indirectly
- Data Lakes deal with security in all states

- The focus has been security; now we need to talk about ***Privacy***.

Part II
Data Privacy Basics



This Photo by Unknown Author is licensed under CC BY-SA-NC

- What is Privacy?
- Data Sensitivities
- Regulatory Issues
- Query Controls

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.

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.

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

But shouldn’t we just pass laws banning these kinds of activities?

Maybe. But we also have to develop technologies that can keep ***data private*** so that there isn’t even the option to disseminate.

Data Ownership vs Stewardship

- Who “owns” data about you? This is a legal/political question
- In Europe, laws generally support that you own the data about you
- In the United States, laws are generally moving towards this
- For purposes of this class, we assume a user owns their own data
- We will call 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

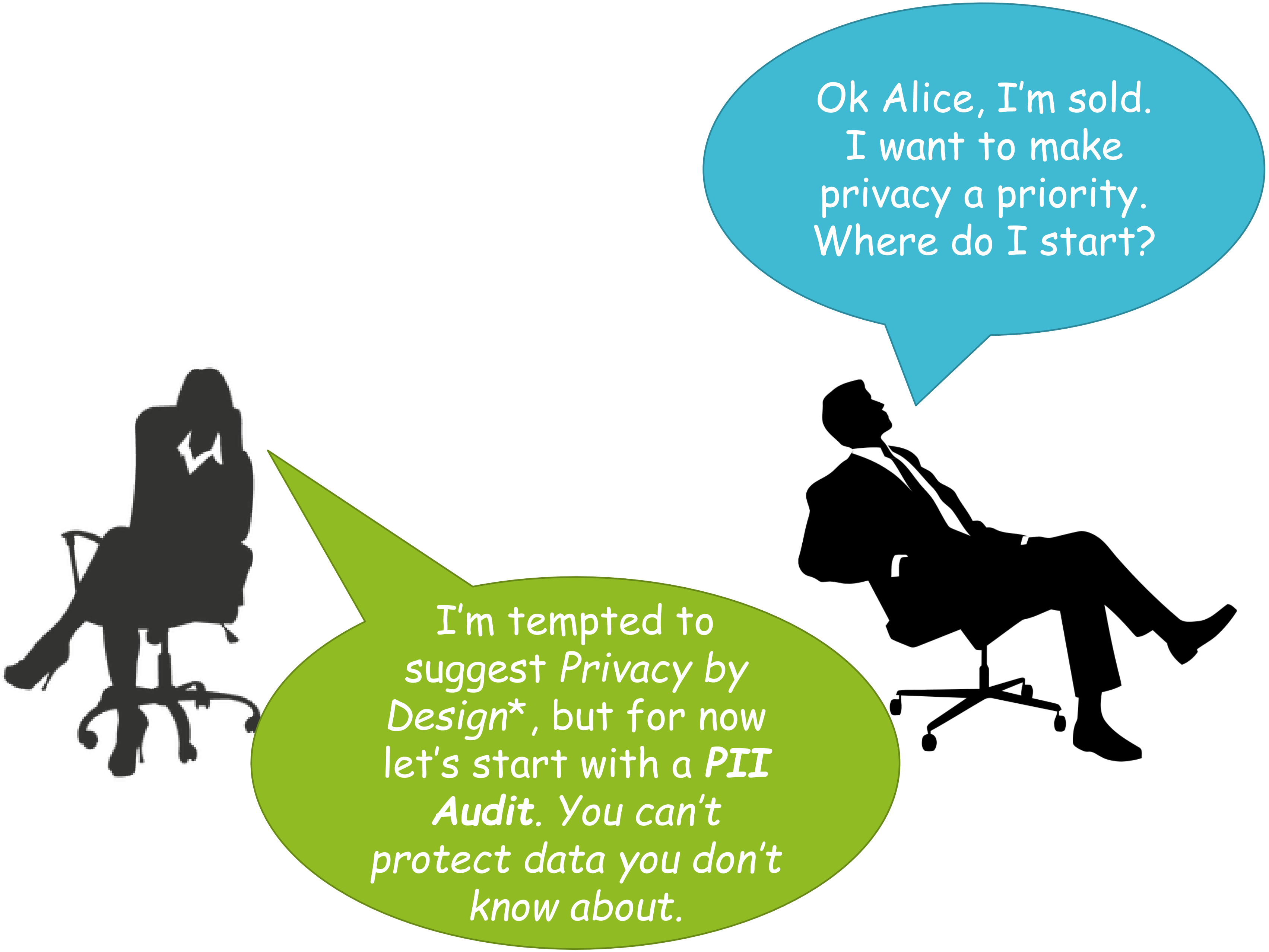
- In practice, data privacy begins with identifying “personal” data
- The defined set of personal data varies by legal jurisdiction
- For example, in Europe an IP address is personal, but not in the US

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."*

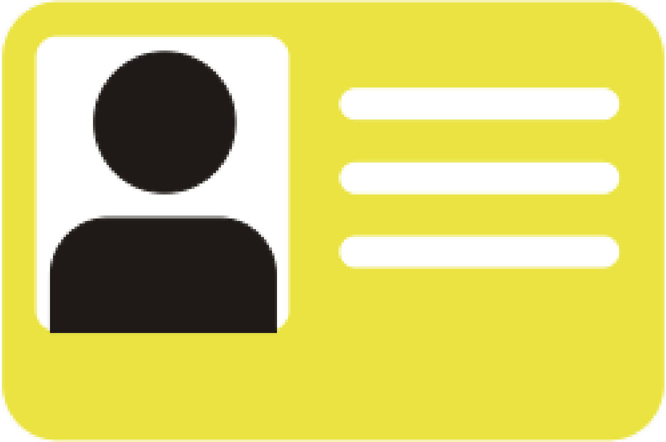
(NIST Special Publication 800-122, emphasis added)

PII Audit

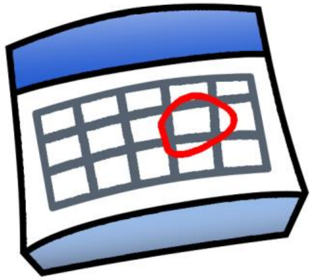

Ok Alice, I'm sold. I want to make privacy a priority. Where do I start?
I'm tempted to suggest *Privacy by Design**, but for now let's start with a ***PII Audit***. *You can't protect data you don't know about.*

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

Distinguishing Data



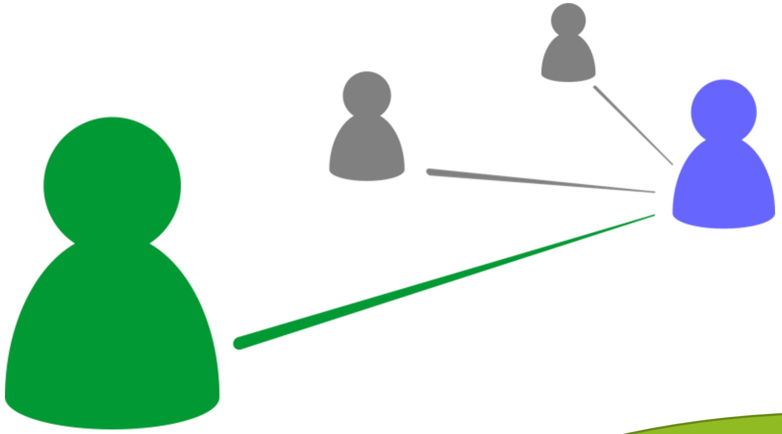
This Photo by Unknown Author is licensed under CC BY-NC

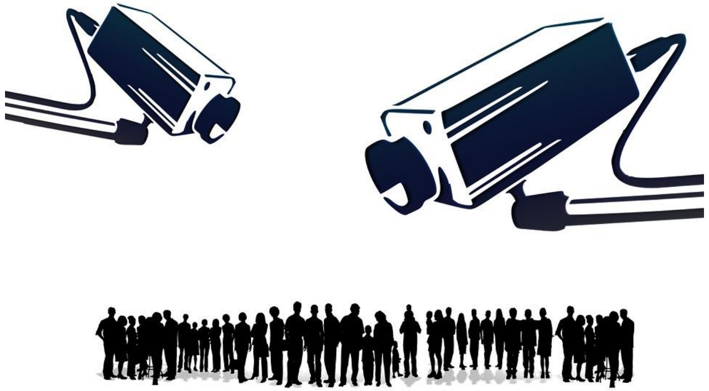


This Photo by Unknown Author is licensed under CC BY

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

Tracing Data

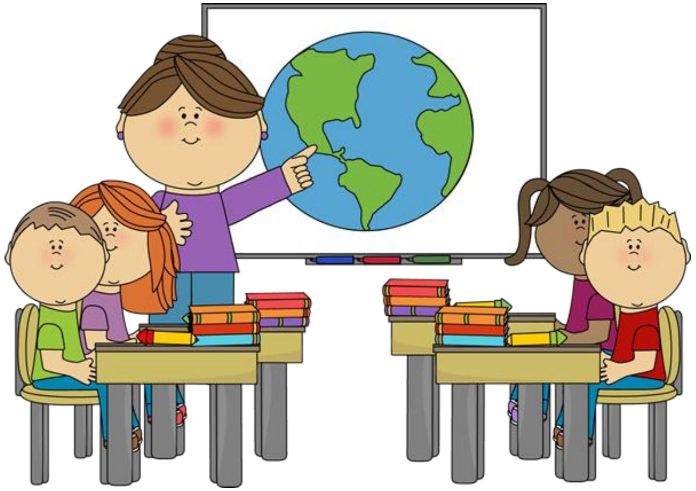




PII also includes data that could be used to determine an individual's activities or status. This includes ***log files*** or camera recordings.

Linking Data (linked or linkable)

This data is already linked to the student



This Photo by Unknown Author is licensed under CC BY-SA-NC

Student	Grade	Height	Weight
Bob Jr.	5	4'5"	100lbs

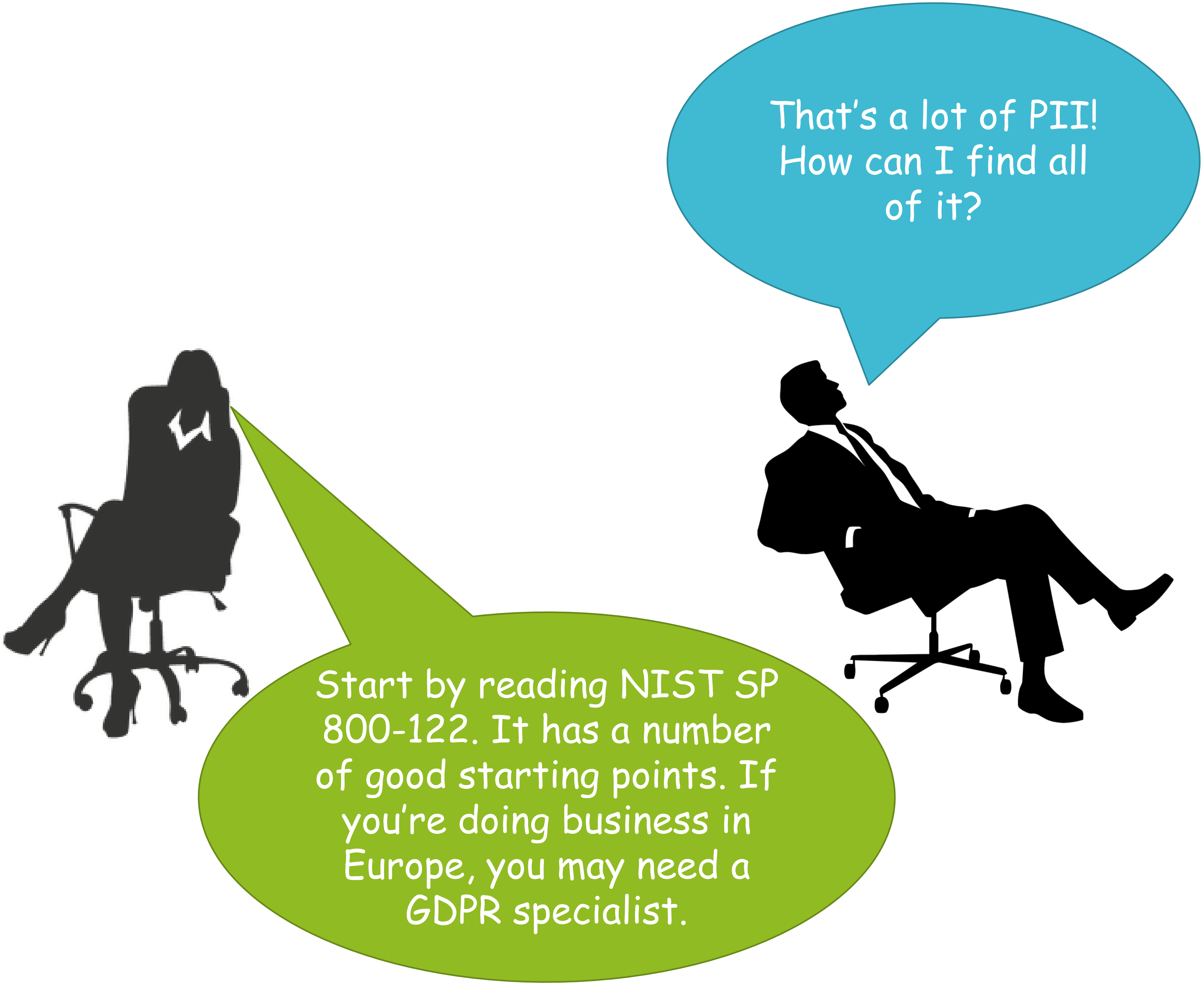
This data could be linked to the student

School Uniforms Online Store

Grade	Height	Weight	Purchases
5	4'5"	100lbs	$100.00

Finally, "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?

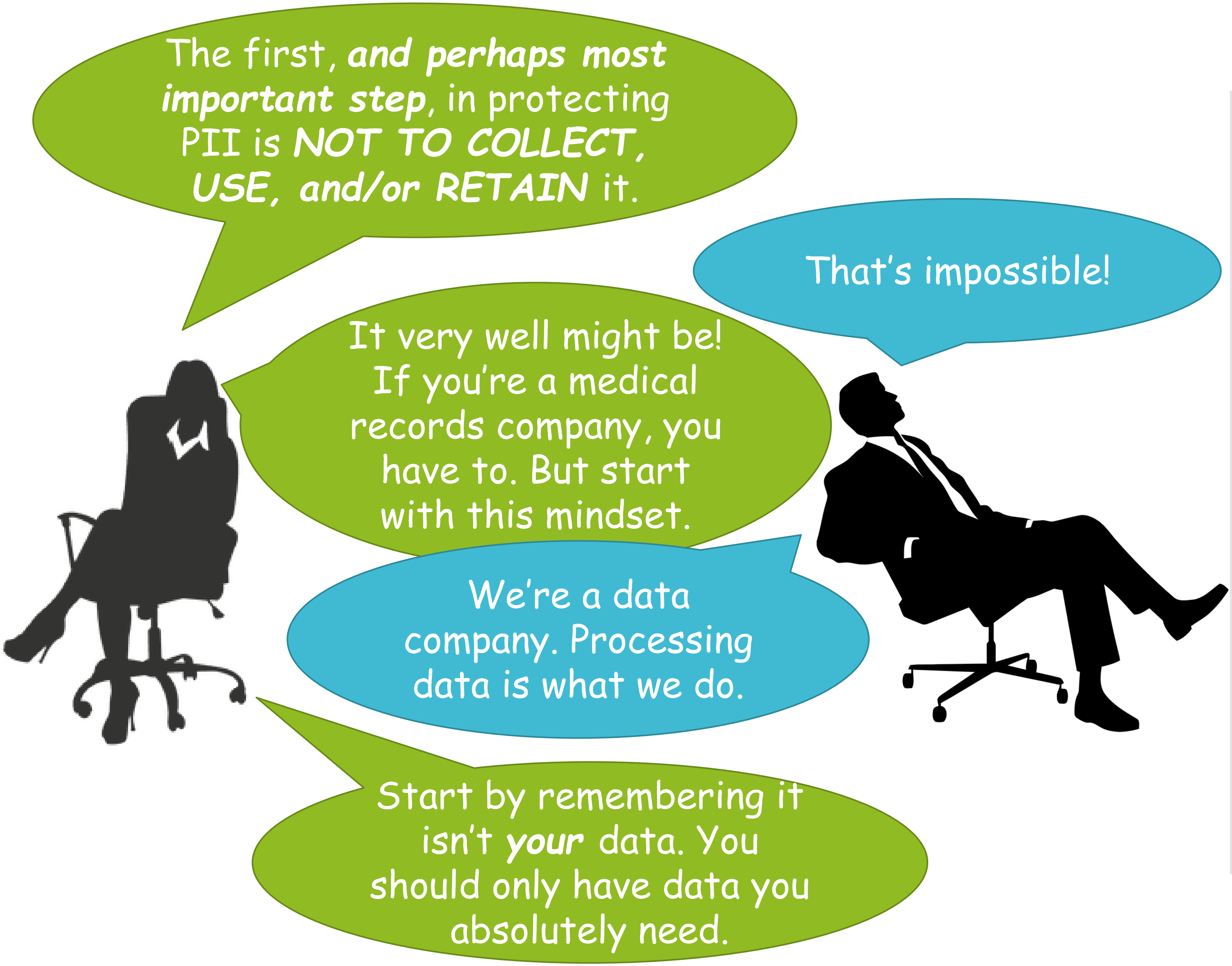
Start by reading NIST SP 800-122. It has a number of good starting points. If you're doing business in Europe, you may need a GDPR specialist.

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.

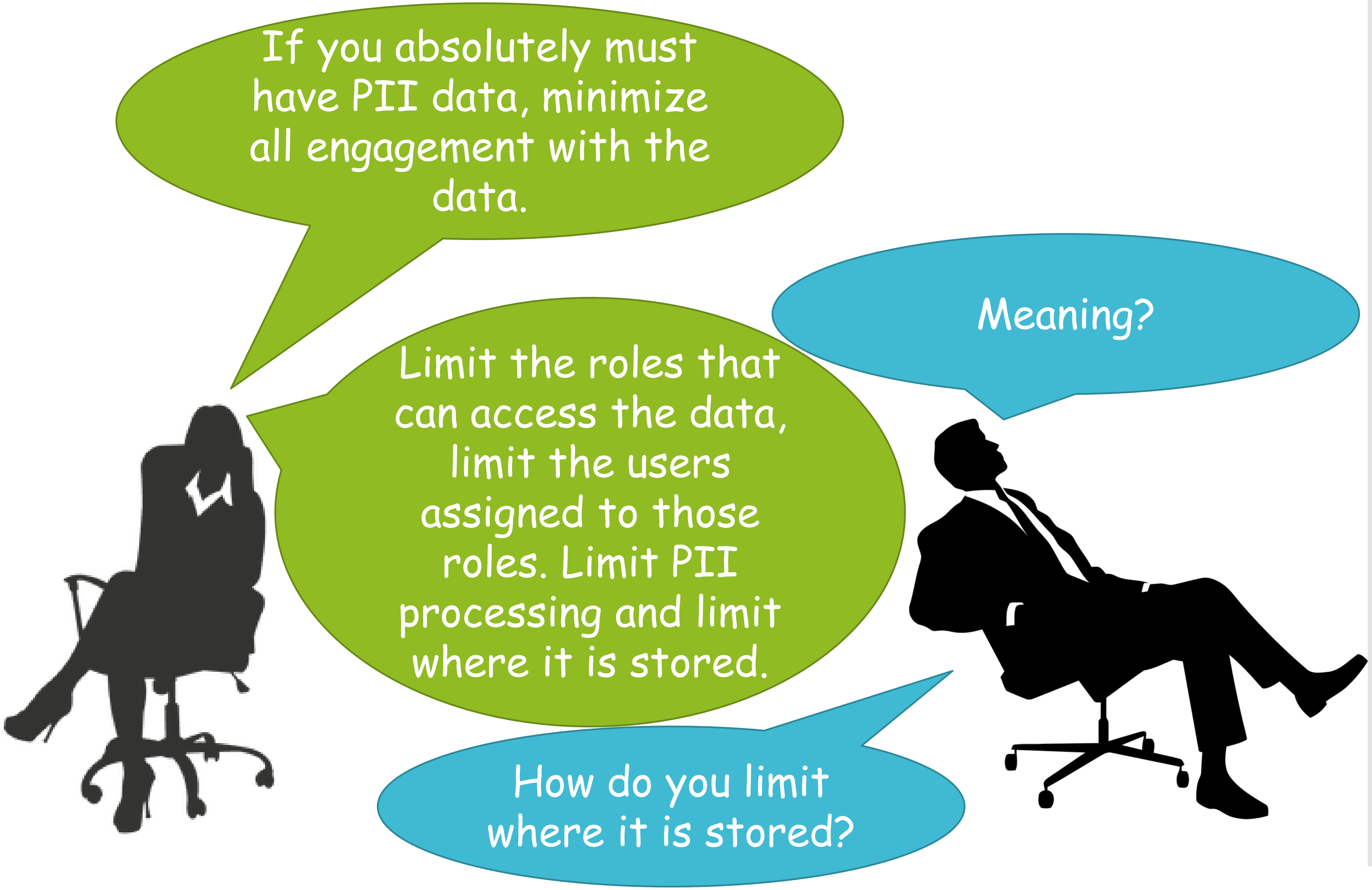
That's impossible!

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)



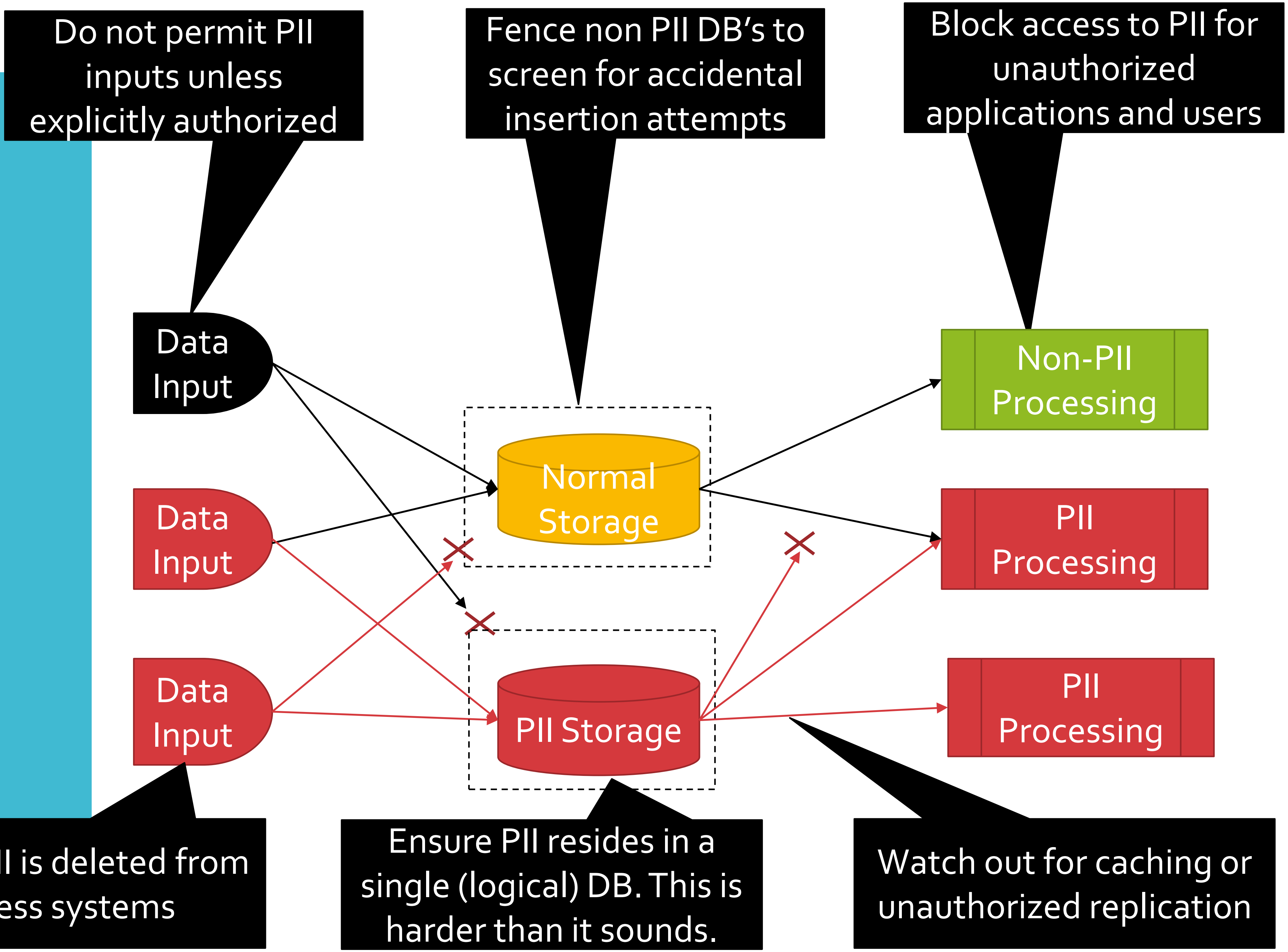
If you absolutely must have PII data, minimize all engagement with the data.

Meaning?

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?

Limiting PII Storage



Do not permit PII inputs unless explicitly authorized

Fence non PII DB's to screen for accidental insertion attempts

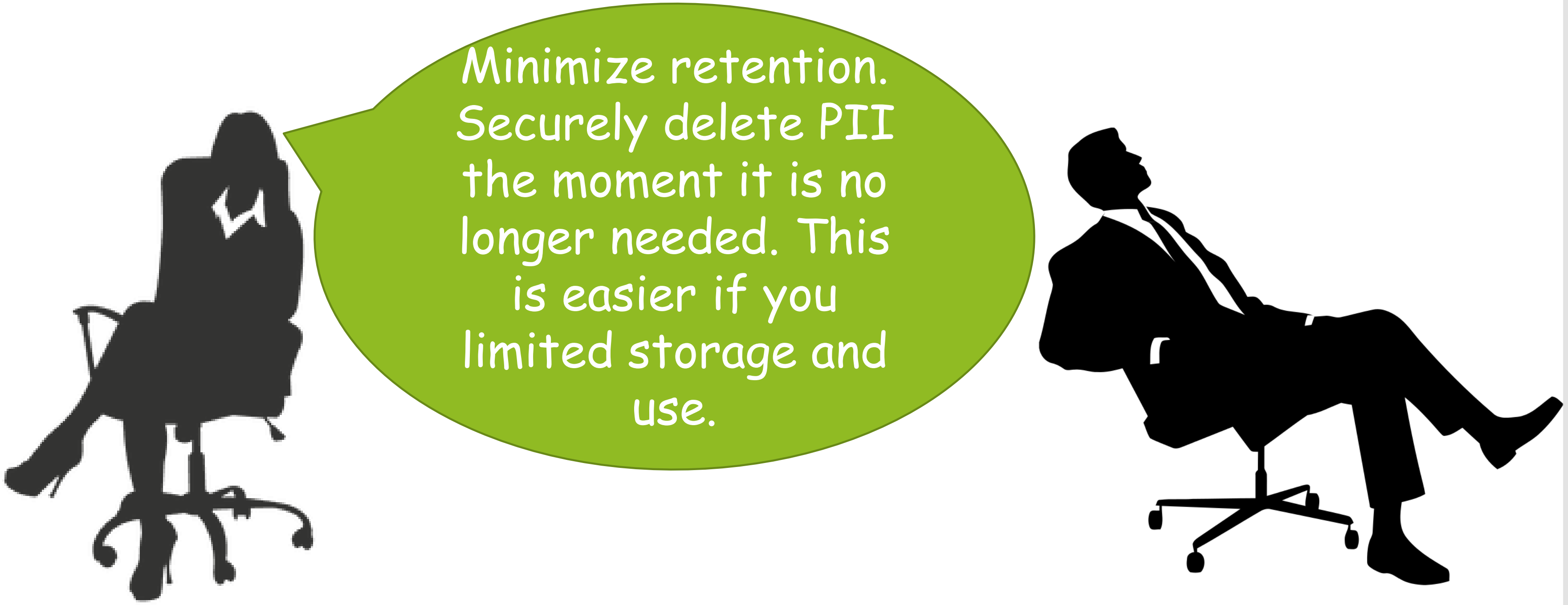
Block access to PII for unauthorized applications and users

Ensure PII is deleted from ingress systems

Ensure PII resides in a single (logical) DB. This is harder than it sounds.

Watch out for caching or unauthorized replication

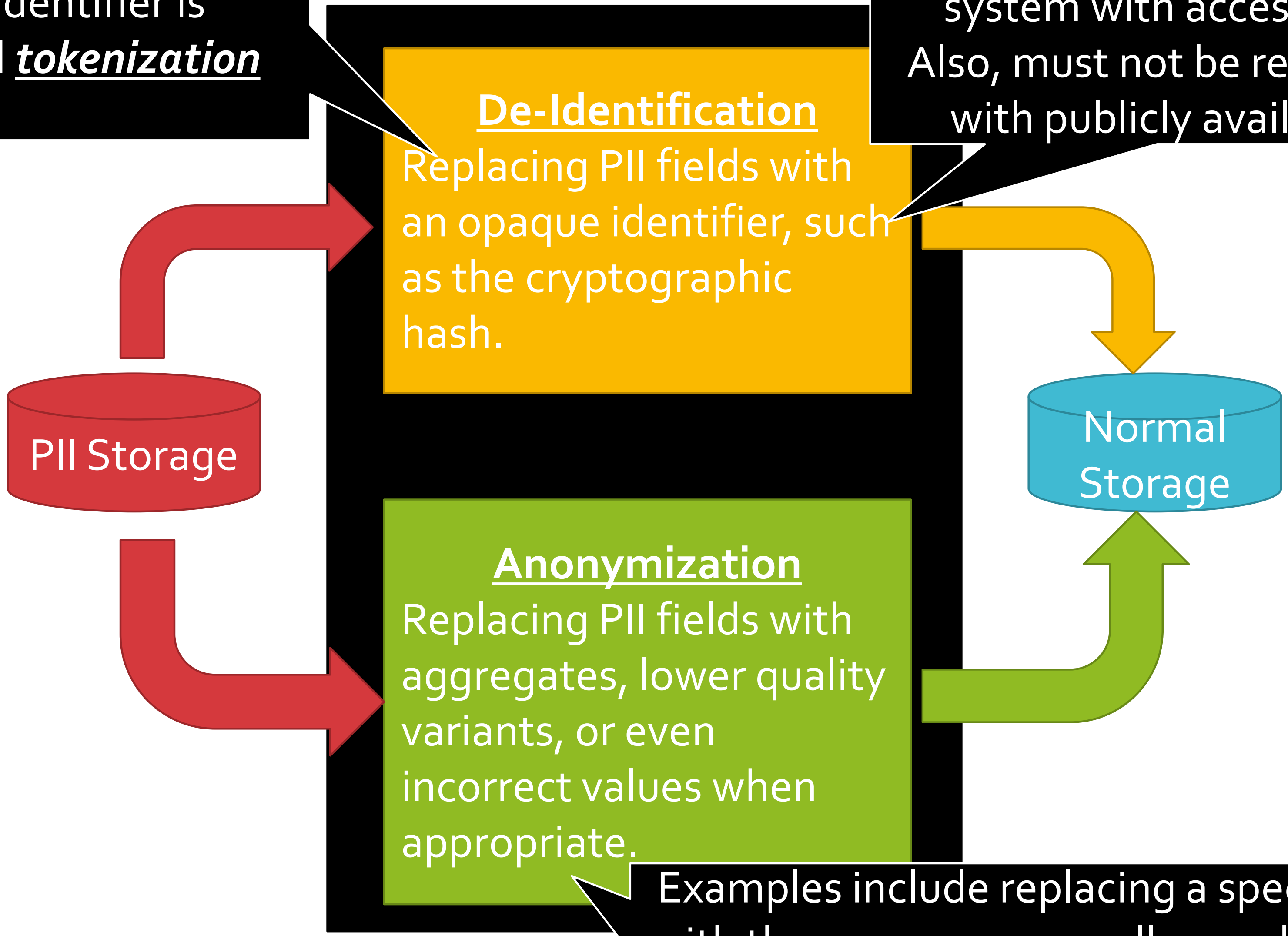
Minimizing PII Use, Collection, and Retention (3)



Minimize retention. Securely delete PII the moment it is no longer needed. This is easier if you limited storage and use.

De-Identification and Anonymization

Replacing fields with a hash or other opaque identifier is sometimes called ***tokenization***



PII Storage

De-Identification
Replacing PII fields with an opaque identifier, such as the cryptographic hash.

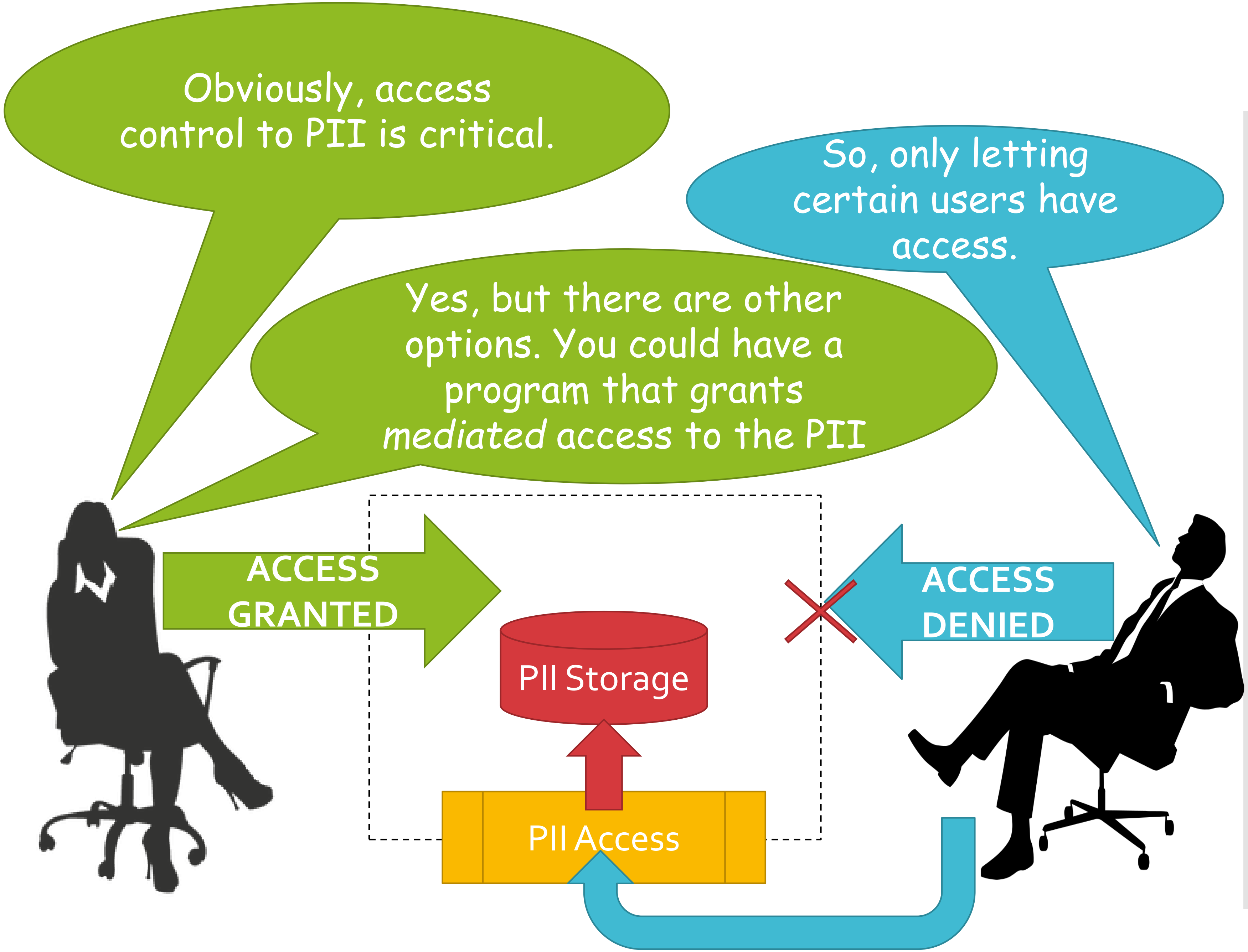
De-identified data *can* be re-identified. Must be on a separate system with access controls. Also, must not be re-identifiable with publicly available data.

Normal Storage

Anonymization
Replacing PII fields with aggregates, lower quality variants, or even incorrect values when appropriate.

Examples include replacing a specific field with the average across all records or even shuffling PII fields amongst the records in the set.

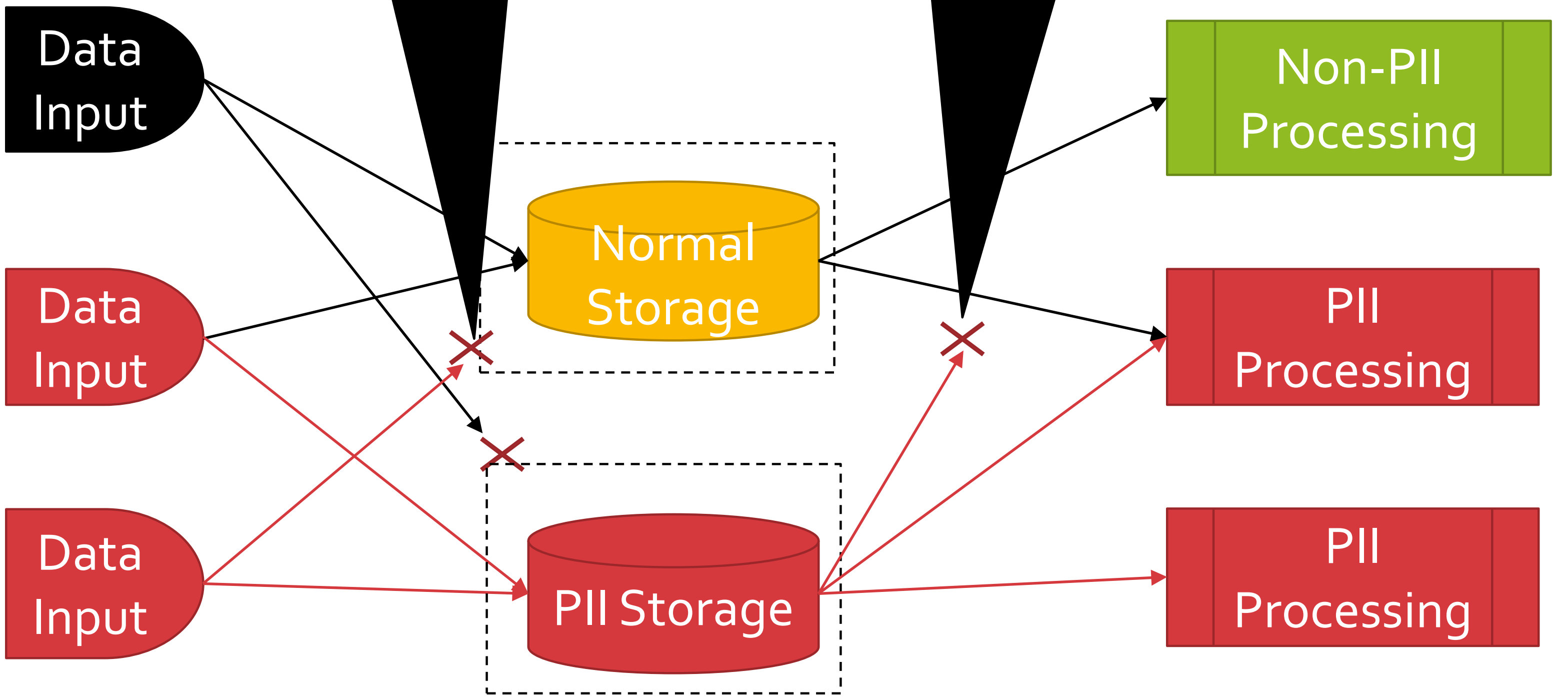
Access Controls


Obviously, access control to PII is critical.
Yes, but there are other options. You could have a program that grants *mediated* access to the PII
So, only letting certain users have access.
ACCESS GRANTED
ACCESS DENIED
PII Storage
PII Access

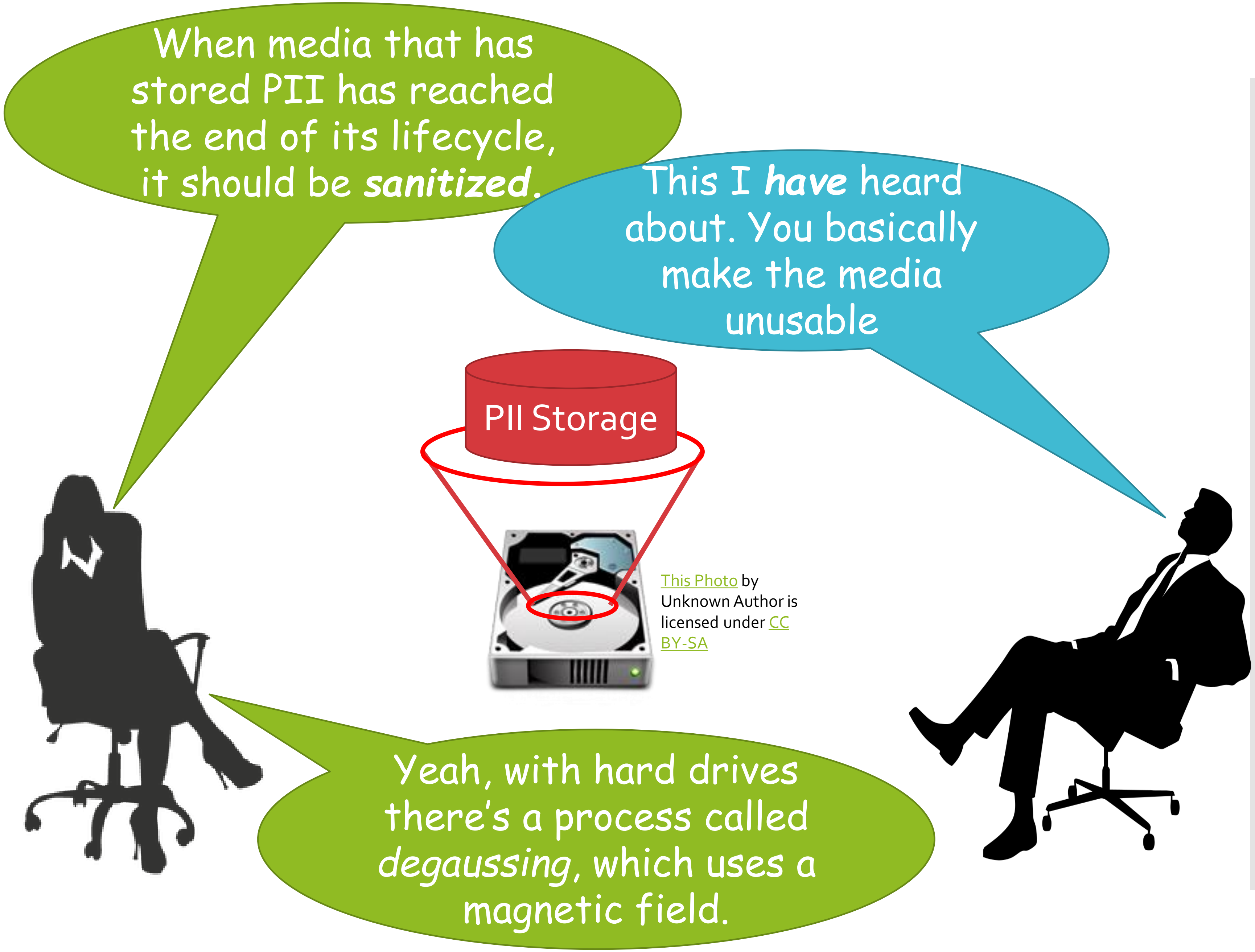
Auditable Events and System Monitoring

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

Information monitors, such as **data loss prevention** systems can find and block PII transfers



Media Sanitization



When media that has stored PII has reached the end of its lifecycle, it should be ***sanitized***.

This I ***have*** heard about. You basically make the media unusable

This Photo by Unknown Author is licensed under CC BY-SA

Yeah, with hard drives there's a process called *degaussing*, which uses a magnetic field.

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!