Vendor Management in the Era of Big Data and Machine Learning

Speaker

Dr. Daniel Fabbri, Founder & CEO, Maize Analytics, Inc.

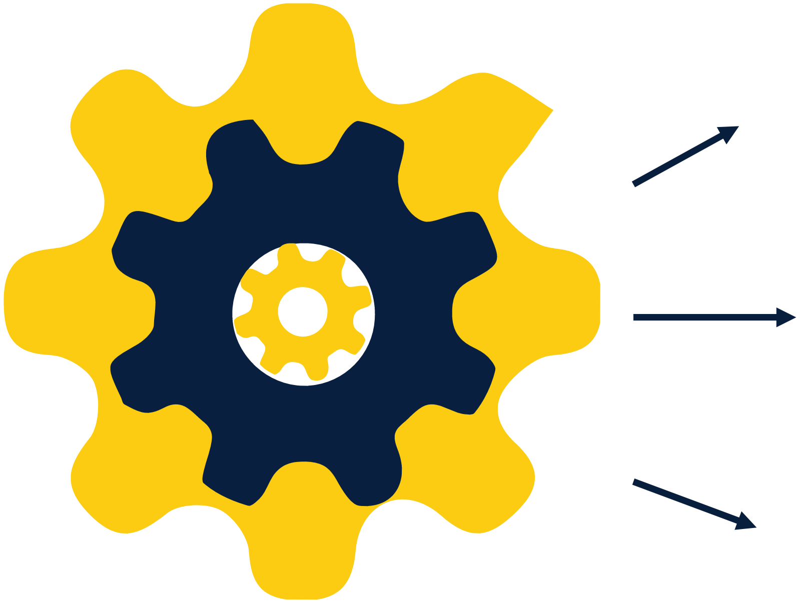
- PhD in Computer Science from University of Michigan, Ann Arbor
- Assistant Professor of Biomedical Informatics and Computer Sciences, Vanderbilt University

Big Data and Machine Learning

Predictive Analytics

Improve Care

Optimize Treatment



Machine Learning

Billing

Improve Treatment

Population Health

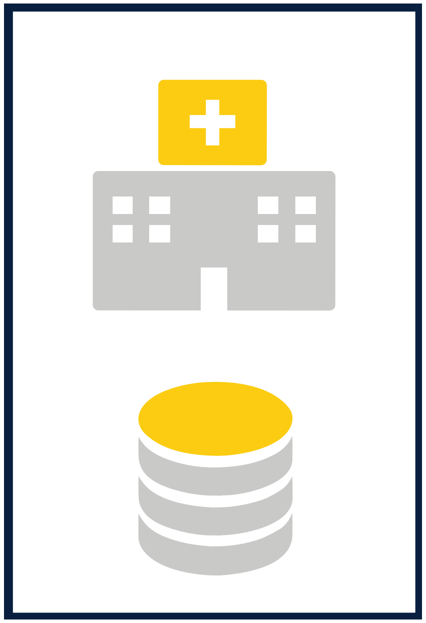
Moving to the Cloud

- Health information stored in the cloud doubled, 2014-2016
- Vendor's have data expertise: billing, population health, diagnostics
- Covered entities are still responsible for vendor's usage of their data

Source: *HIMSS Media: The Cloud Evolution in Healthcare*

Background: Onsite Deployment

- Application deployed within provider’s firewall
- Strict data controls and high data visibility
- Lacks the cloud’s scalability



Background: Cloud Deployments

Covered Entity-Managed Cloud:

- CE controls environment
- CE have better visibility into data movement and controls
- Vendors must manage application in multiple locations

Vendor-Managed Cloud:

- Vendor manages all resources
- CE loses visibility into their data once sent
- Economies of scale benefit the Vendor

Background: NIST Supply Chain Risk

Supply chain risks are often associated with an "organization's decreased visibility into, and understanding of, how the technology that they acquire is developed, integrated, and deployed."

Does the lack of visibility put your organization at risk?

NIST revision 2 of NIST SP 800-37, Risk Management Framework, December 2018

Risks of Vendors in the Cloud

Data Mixing

Machine Learning Model Mixing

Data Repurposing

Example In the News

- Social media firm released user data to researchers
- Once user data left their servers, control was lost
- Data were used for unintended purposes

Risks of Vendors in the Cloud

Data Mixing

Machine Learning Model Mixing

Data Repurposing

Vendor-Hosted Solution

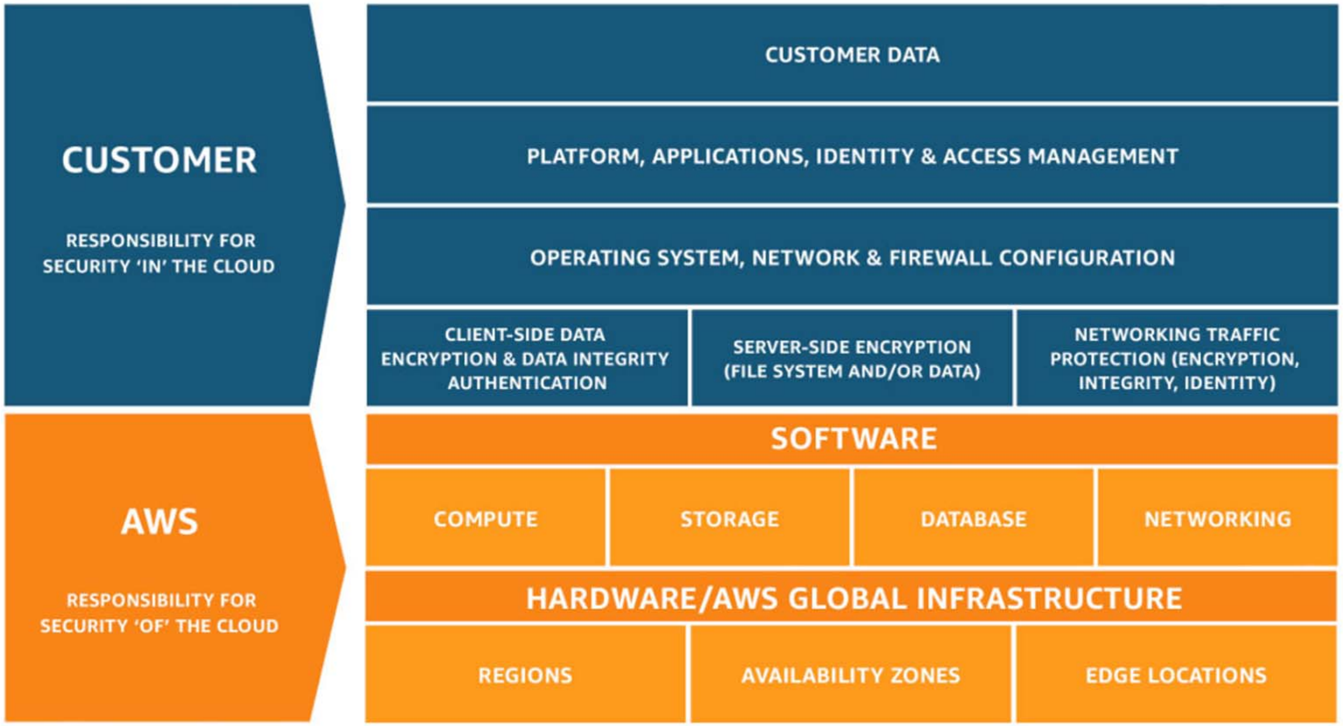
Shared Responsibility Model

- Security ***of*** the cloud is managed by the cloud provider (e.g. AWS).
- Security ***in*** the cloud is the vendor's responsibility.
- Third party vendor controls data placement.

Data Organizations **in** the Cloud:

- Single-tenant data storage
- Multi-tenant data storage

Shared Responsibility Model



CUSTOMER

RESPONSIBILITY FOR SECURITY 'IN' THE CLOUD

CUSTOMER DATA

PLATFORM, APPLICATIONS, IDENTITY & ACCESS MANAGEMENT

OPERATING SYSTEM, NETWORK & FIREWALL CONFIGURATION

CLIENT-SIDE DATA ENCRYPTION & DATA INTEGRITY AUTHENTICATION

SERVER-SIDE ENCRYPTION (FILE SYSTEM AND/OR DATA)

NETWORKING TRAFFIC PROTECTION (ENCRYPTION, INTEGRITY, IDENTITY)

AWS

RESPONSIBILITY FOR SECURITY 'OF' THE CLOUD

SOFTWARE

COMPUTE

STORAGE

DATABASE

NETWORKING

HARDWARE/AWS GLOBAL INFRASTRUCTURE

REGIONS

AVAILABILITY ZONES

EDGE LOCATIONS

Source: Amazon AWS

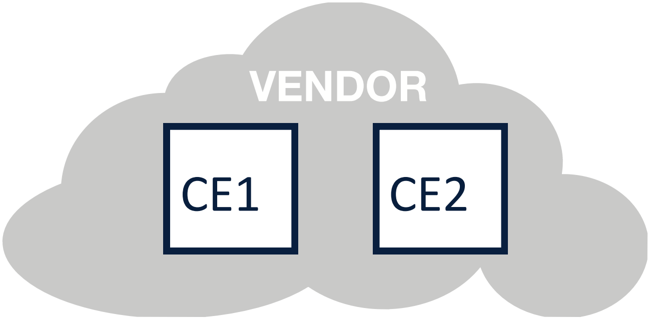
Single-Tenant

- **Logical separation** between CE's data
- Data are never mixed in a database or other storage system
- **Dedicated Hardware:** Logical and physical separation

VENDOR
CE1
VENDOR
CE2

Multi-Tenant

- Two or more CE's data in a single data storage system
- Software controls limit who can see what
- Bugs risks inadvertent exposure
- Vendors can prefer multi-tenant -- fewer systems to manage



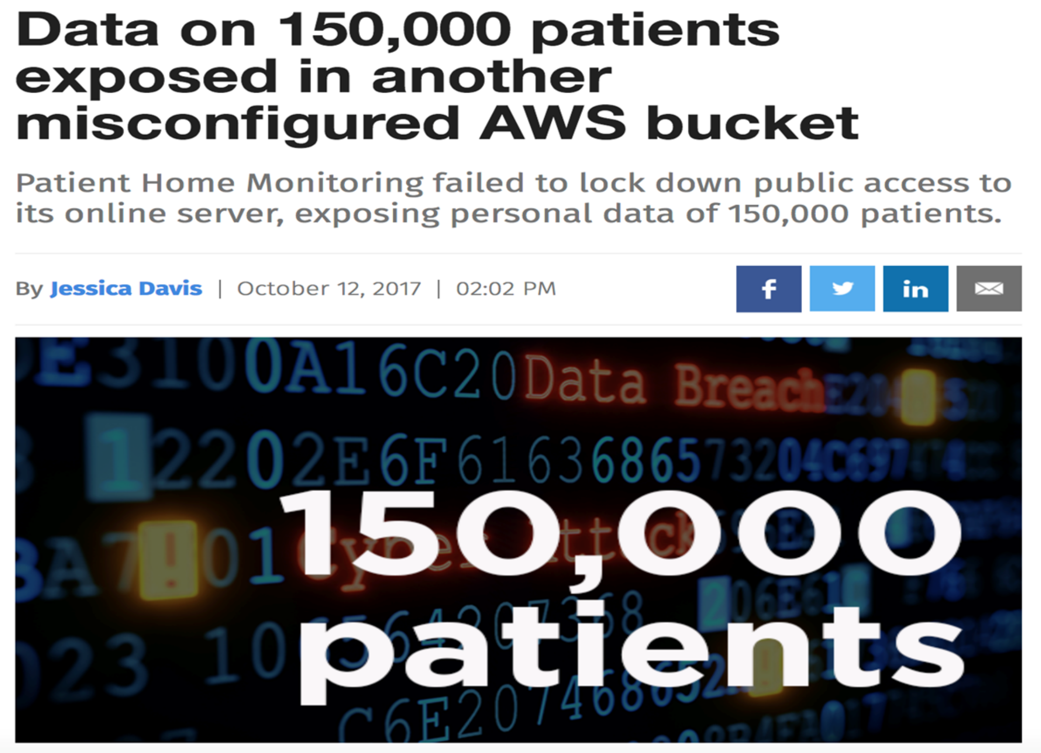
Reminders of Shared Hardware Risks

Applications run on shared hardware

Compromised hardware / applications can leak sensitive data

Source: https://meltdownattack.com/

Cloud Misconfiguration - Example



Data on 150,000 patients exposed in another misconfigured AWS bucket

Patient Home Monitoring failed to lock down public access to its online server, exposing personal data of 150,000 patients.

By **Jessica Davis** | October 12, 2017 | 02:02 PM

Source: Healthcare IT News

Steps To Take

- Ask if vendor deploys in single or multi-tenant environment
- Add contractual language to require specific tenancy

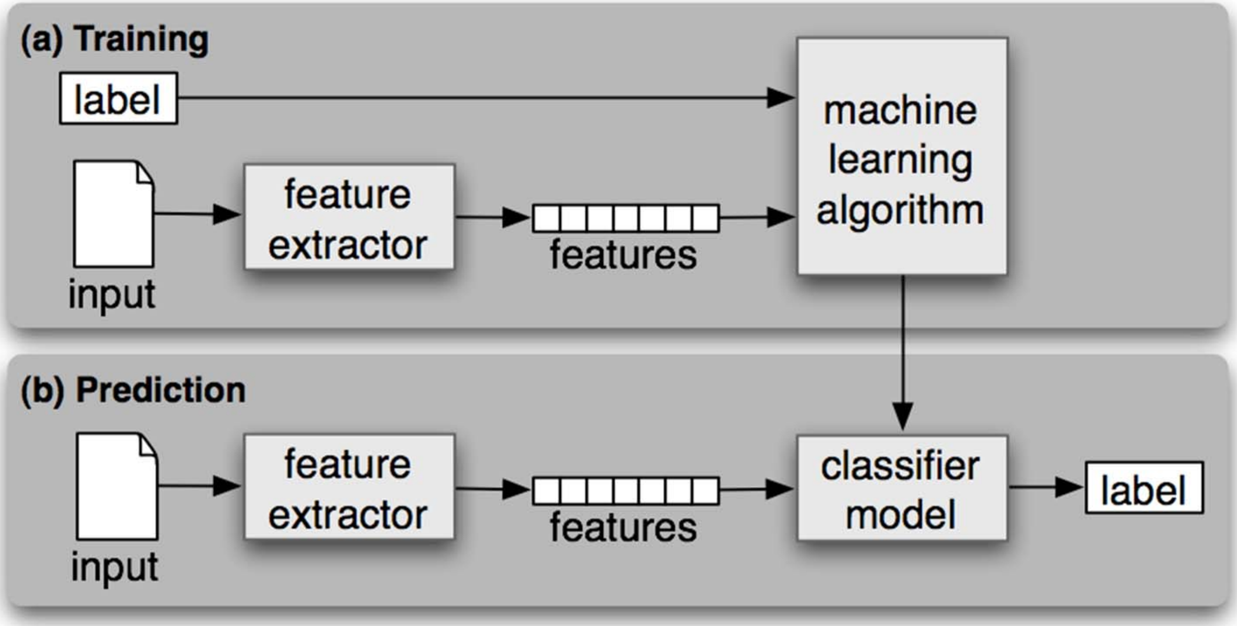
Risks of Vendors in the Cloud

Data Mixing

Machine Learning Model Mixing

Data Repurposing

Machine Learning Pipeline



Models benefit from more training data

Source: http://www.nltk.org/book/ch06.html

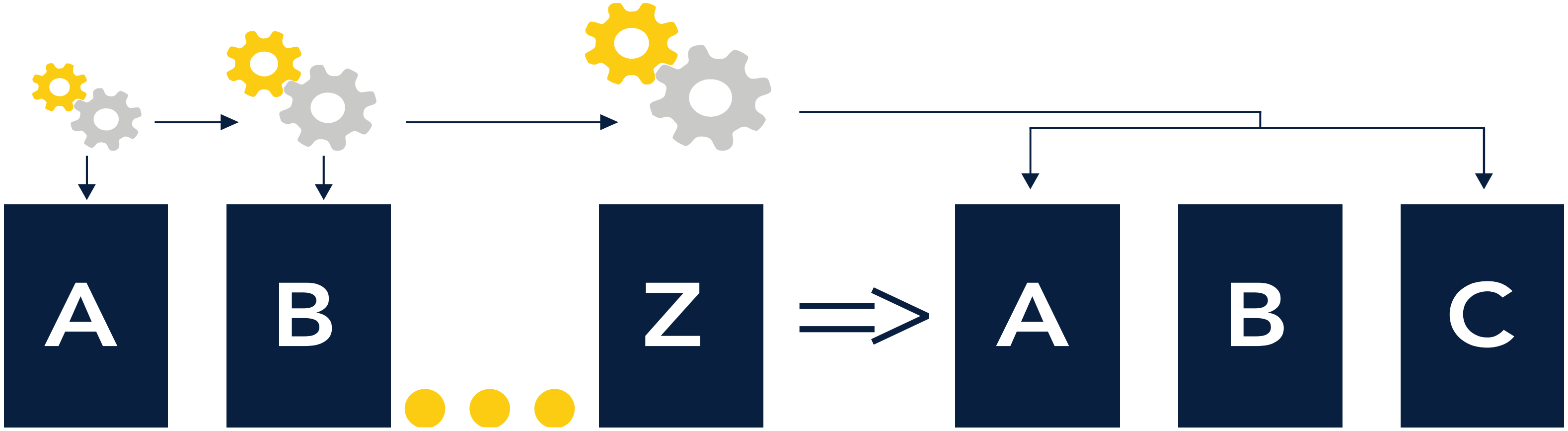
Training Data Set Construction

- **Multi-Tenant:** Data are already aggregated
- **Single-Tenant with Shared DB:** Copy to data warehouse and train
- **Single-Tenant:** Iterative training

Iterative Machine Learning Training

Train model at Covered Entity A’s data, Transfer model to B

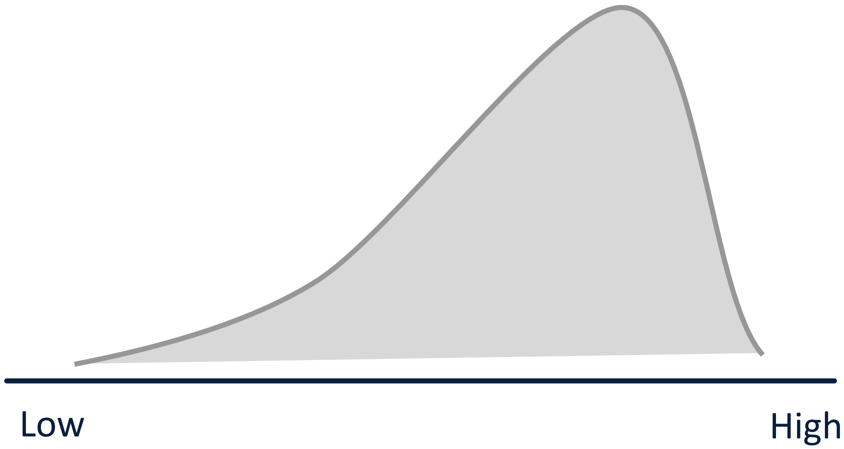
Train model at Covered Entity B’s data, Transfer model to C



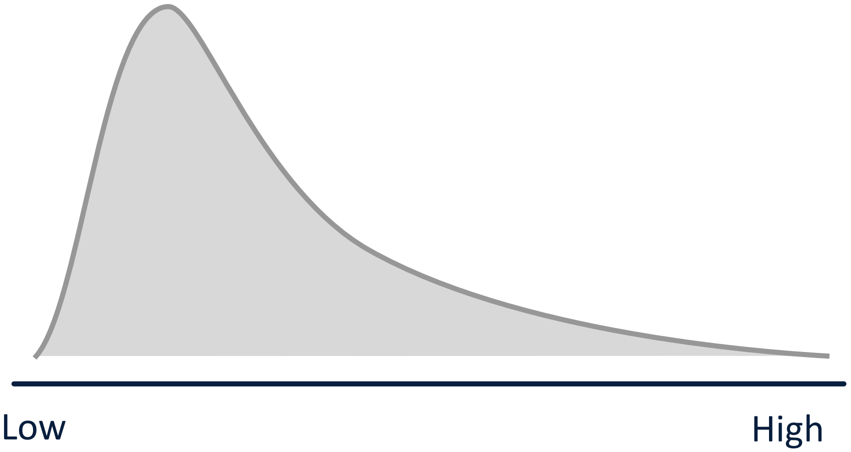
Apply the full model to Covered Entities A, B, C…

Risks of Iterative Machine Learning Training

Incorrect prediction due to
different data distributions between covered entities



CE1 – Diabetes Risk

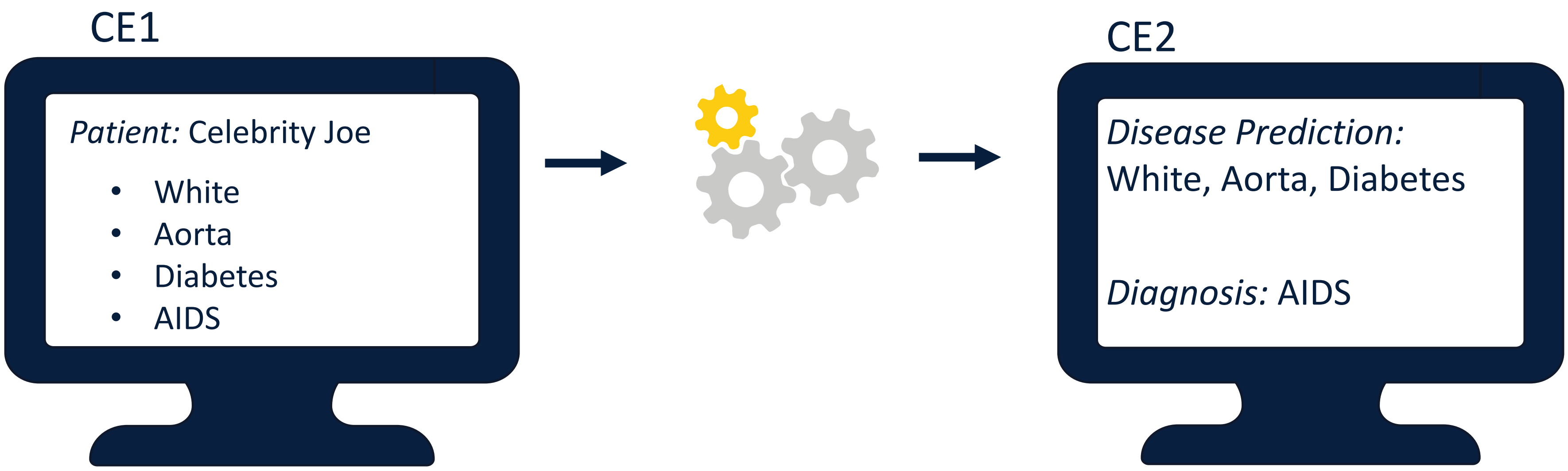


CE2 – Diabetes Risk

Prediction of Diabetes

Risks of Iterative Machine Learning Training

Inadvertent exposure of patient information



CE1

Patient: Celebrity Joe

- White
- Aorta
- Diabetes
- AIDS

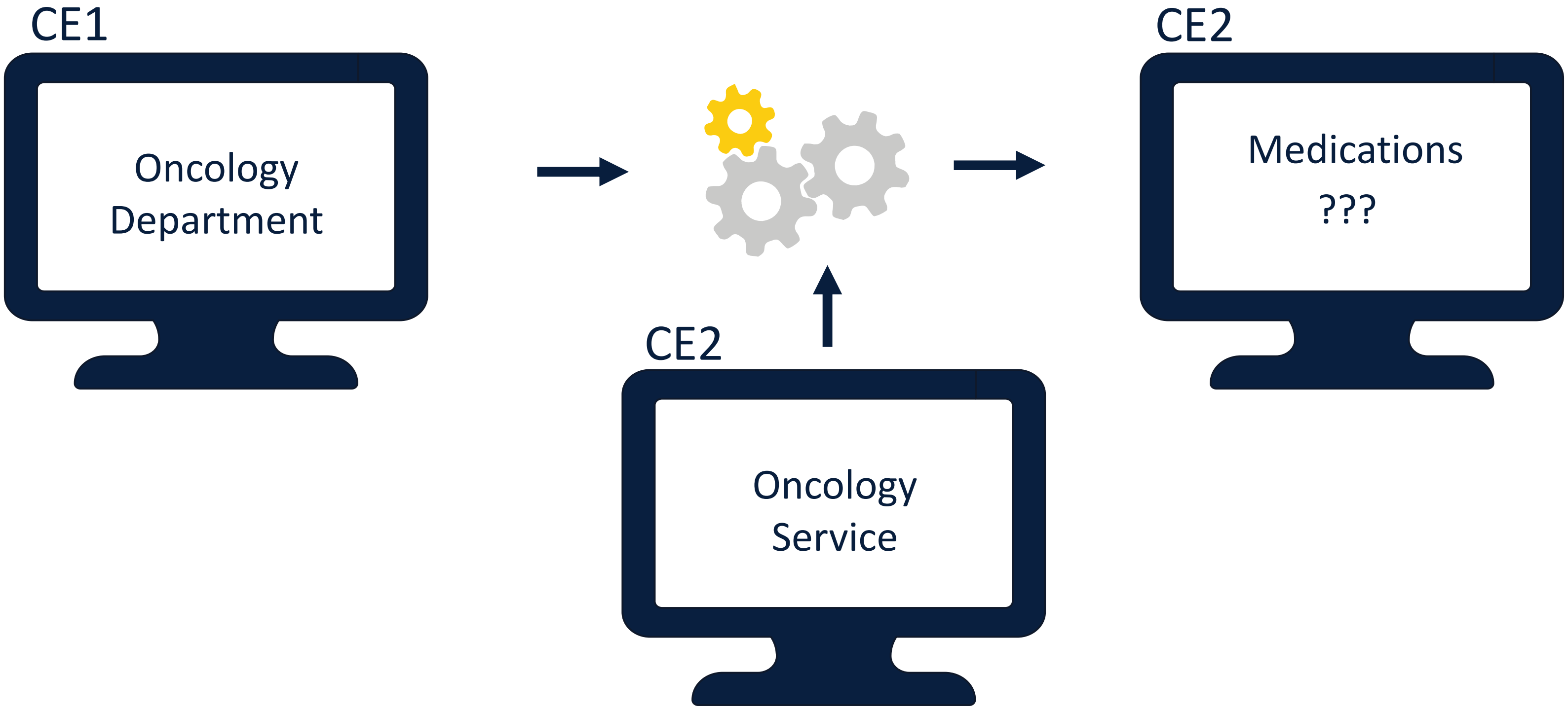
CE2

Disease Prediction:
White, Aorta, Diabetes

Diagnosis: AIDS

Risks of Iterative Machine Learning Training

Incorrect conclusions due to differing semantics



CE1

Oncology Department

CE2

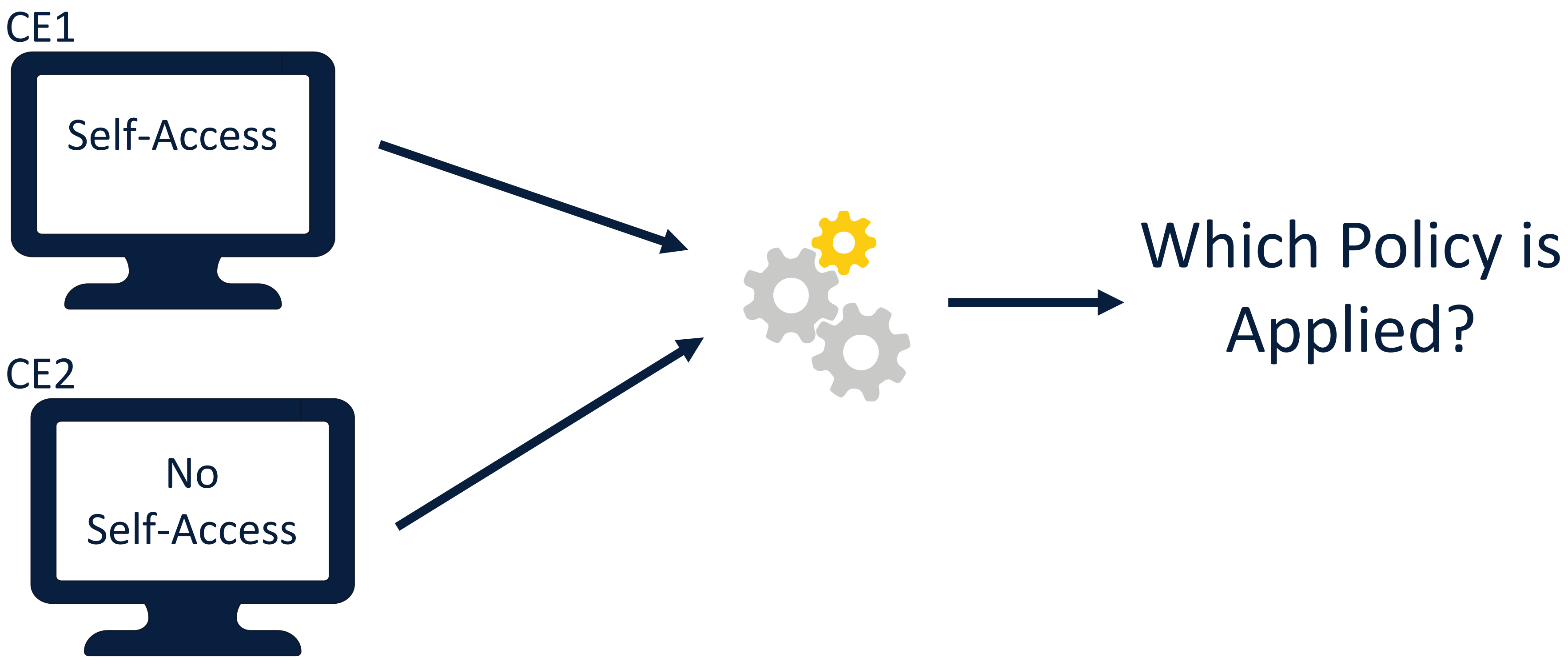
Oncology Service

CE2

Medications ???

Risks of Iterative Machine Learning Training

Incorrect application of Covered Entities' policies



Steps To Take

- Ask if only your data or other CE data will be used to train
- Add contractual language to restrict model training / sharing

Risks of Vendors in the Cloud

Data Mixing

Machine Learning Model Mixing

Data Repurposing

Data Repurposing

- Vendors using data for non-contracted purposes
- Difficult to detect as CEs lack visibility into data usage

Vendor Data Monitoring

Covered entities need better visibility into vendors’ data management

Visibility includes:

- What data are sent?
- Where are data stored?
- What operations are performed on the data?

Types of Monitoring

Application Monitoring

- Accesses to applications by vendor employees

Backend Monitoring

- Queries to backend data management system

Data Governance Monitoring

- Data sent to each vendor

Require access to these logs as part of your contracting process

Example: Vendor Breach Remediation

- Covered entities have an obligation to notify each patient
- Need to identify what PHI data the Vendor held
- Manual retrospective reviews of feeds are slow and often inaccurate

Would you benefit from knowing which MRNs Vendors receive?

Vendor Data Usage Monitoring

- Legal requirements are not enough – **Trust, Monitor, Verify.**
- Need visibility into how vendors use data

Steps To Take

- Contractually require that data are not repurposed
- Contractually require access to vendor logs and monitor

Vendor Management in Era of Big Data

Manage Vendors through contracting:

- Data Mixing
- Machine Learning Model Mixing
- Data Repurposing

Trust but Monitor:

- Application Monitoring
- Back-end Query Monitoring
- Data Governance Monitoring

Questions?