Privacy-Preserving Analytics and Secure Multiparty Computation

Ulf Mattsson

Chief Security Strategist

www.Protegrity.com

PROTEGRITY

(ISC)2 East Bay Chapter

Ulf Mattsson

- **Chief Security Strategist,** Protegrity
- **Chief Technology Officer,** Protegrity, Atlantic BT, and Compliance Engineering
- **Head of Innovation,** TokenEx
- **IT Architect,** IBM

- Develops **Industry Standards**
- **Inventor** of more than 70 issued US Patents
- **Products and Services**:
  - Data Encryption, Tokenization, and Data Discovery
  - Cloud Application Security Brokers (**CASB**) and Web Application Firewalls (**WAF**)
  - Security Operation Center (**SOC**) and Managed Security Services (**MSSP**)
  - Robotics and Applications

ISACA JOURNAL Nov 2021

How Innovative Businesses Win with Secure Machine Learning

ISACA JOURNAL May 2021

Privacy-Preserving Analytics and Secure Multi-Party Computation

ISACA JOURNAL May 2020

Practical Data Security and Privacy for GDPR and CCPA

Cloud Security Alliance

Quantum Computing

cloud security alliance

Cybersecurity and Cryptographic Solutions

Tokenization Management and Security

Cloud Management and Security

Quantum Computing Risk Study

Approved American National Standard

ANSI

American National Standard for Financial Services

ANSI X9.141–2020

Financial Services

Data Security Breach

Part 1: Data Protection

Part 1: Using Encryption Methods

Format-Preserving Encryption – Part 4

Payment Card Industry (PCI) Security Standards Council (SSC):

PCI Security Standards Council

1. **Tokenization Task Force**
2. **Encryption Task Force, Point to Point Encryption Task Force**
3. **Risk Assessment**
4. **eCommerce SIG**
5. **Cloud SIG, Virtualization SIG**
6. **Pre-Authorization SIG, Scoping SIG Working Group**

ISSA DEVELOPING AND CONNECTING CYBERSECURITY LEADERS GLOBALLY

Dec 2019

Data Security: On Premise or in the Cloud

By Ulf Mattsson – ISSA member, New York Chapter

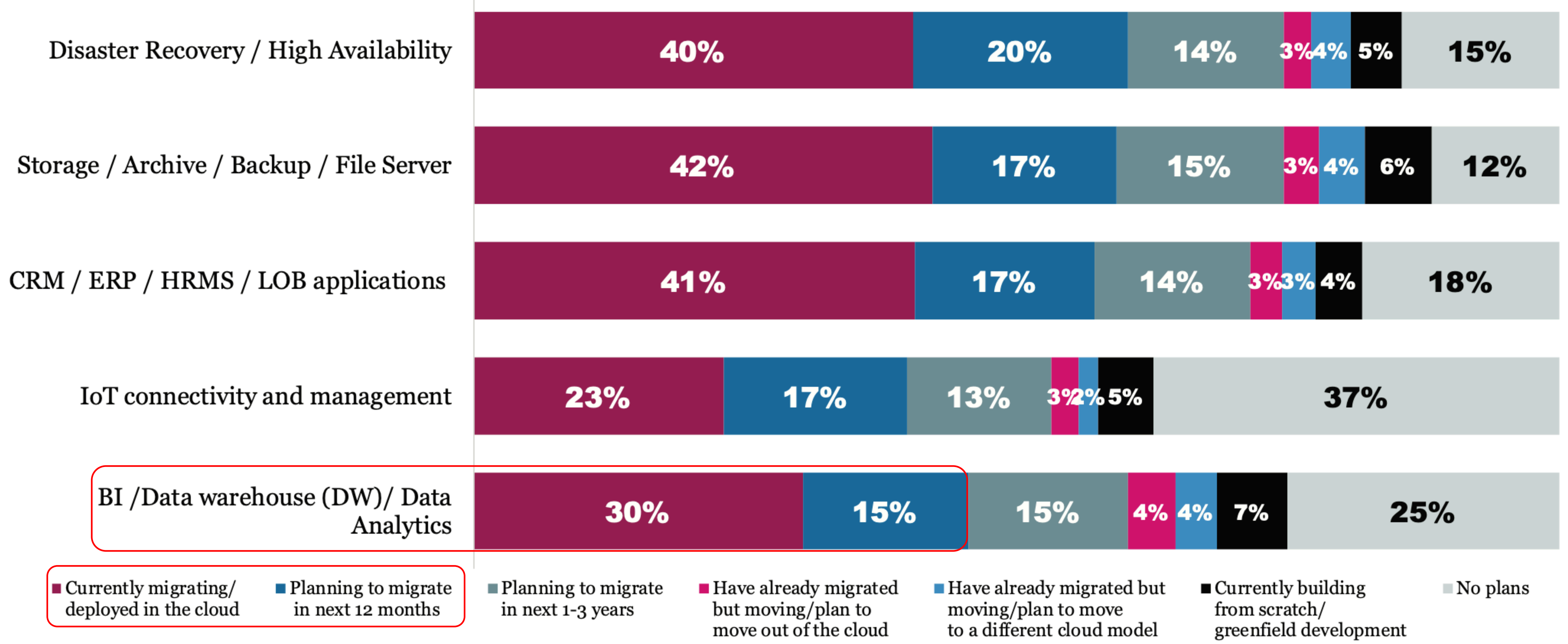
ISSA DEVELOPING AND CONNECTING CYBERSECURITY LEADERS GLOBALLY

May 2020

Data Privacy: De-Identification Techniques

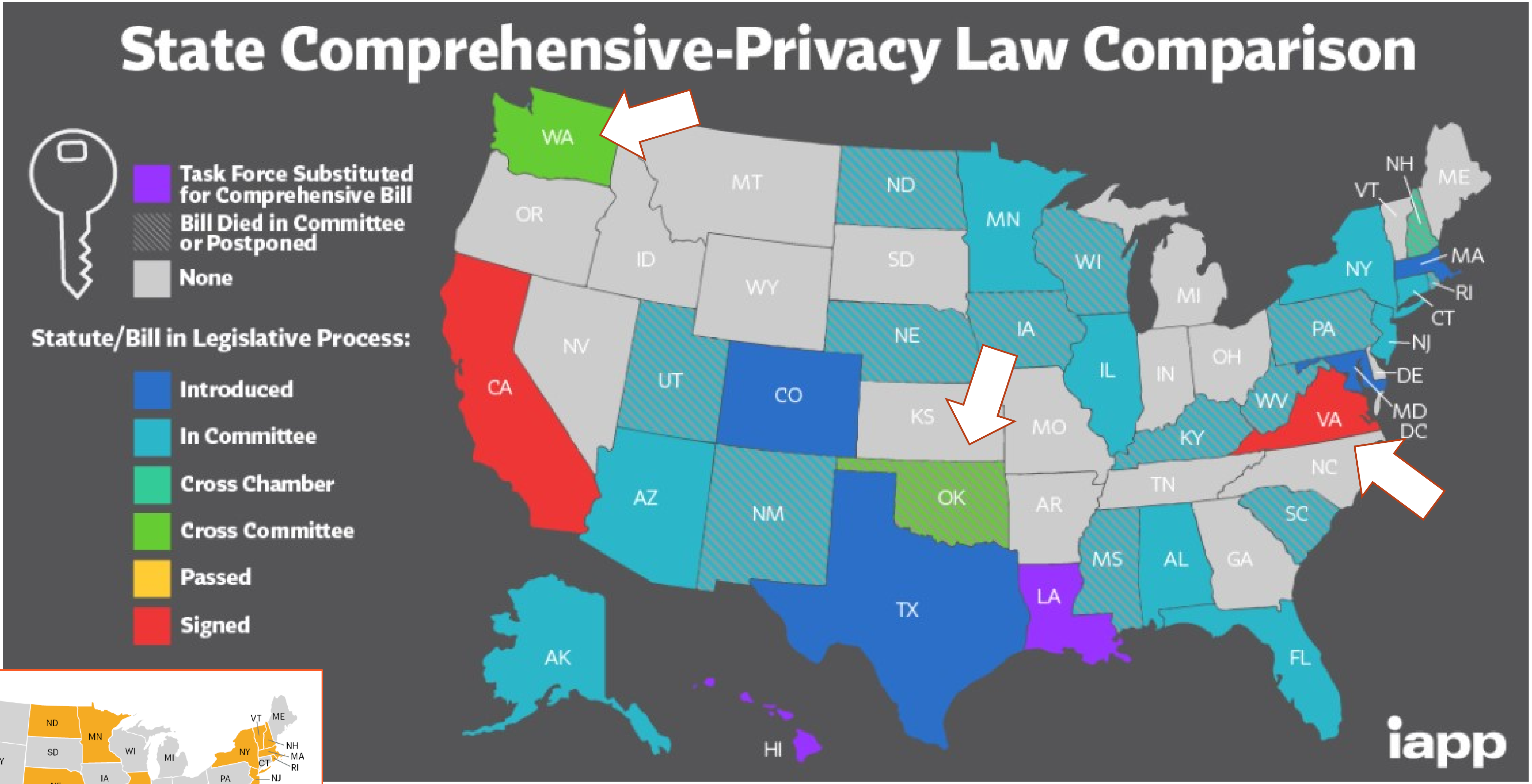
By Ulf Mattsson – ISSA member, New York Chapter

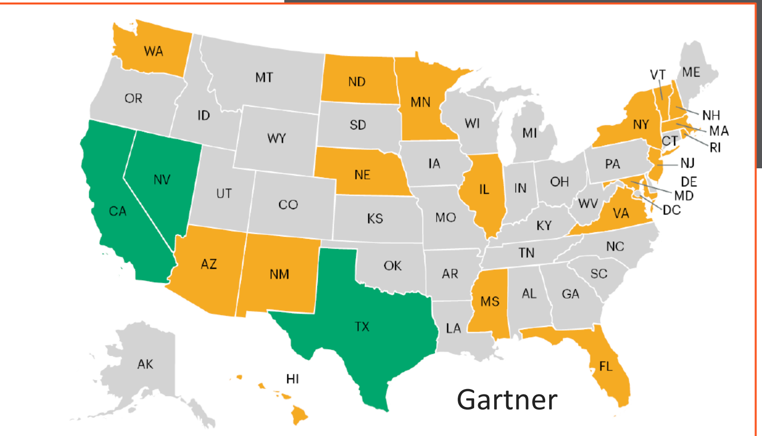
Applications & Services being migrated to the Cloud



The 2020 IDG Cloud Computing Survey

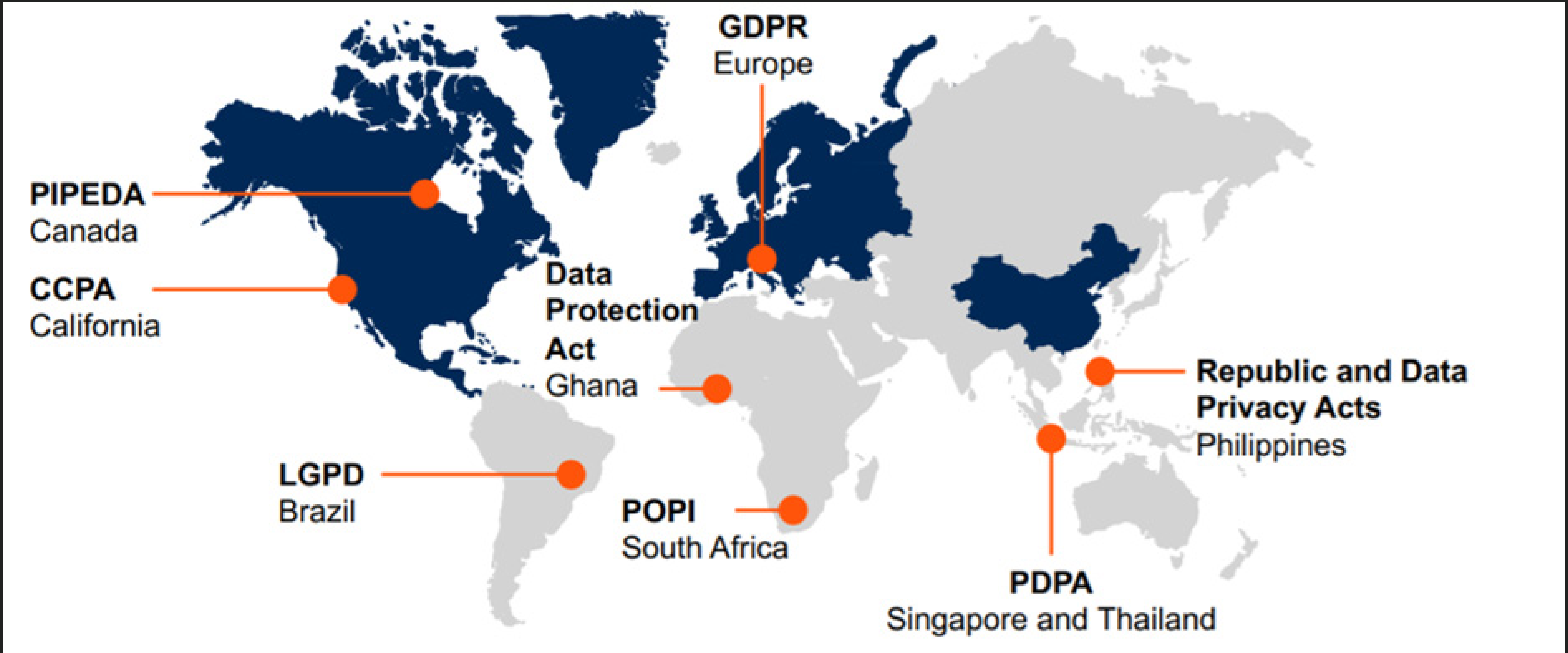
Regulatory Activities in Privacy 2021 vs 2020





The CCPA Effect

The Rise of Legislation



Gartner, 2021

GDPR under "Schrems II"

Legal safeguards:

- AWS Sarl guarantees in its contract with Doctolib, a French company, that it will challenge any general access request from a public authority.

Technical safeguards:

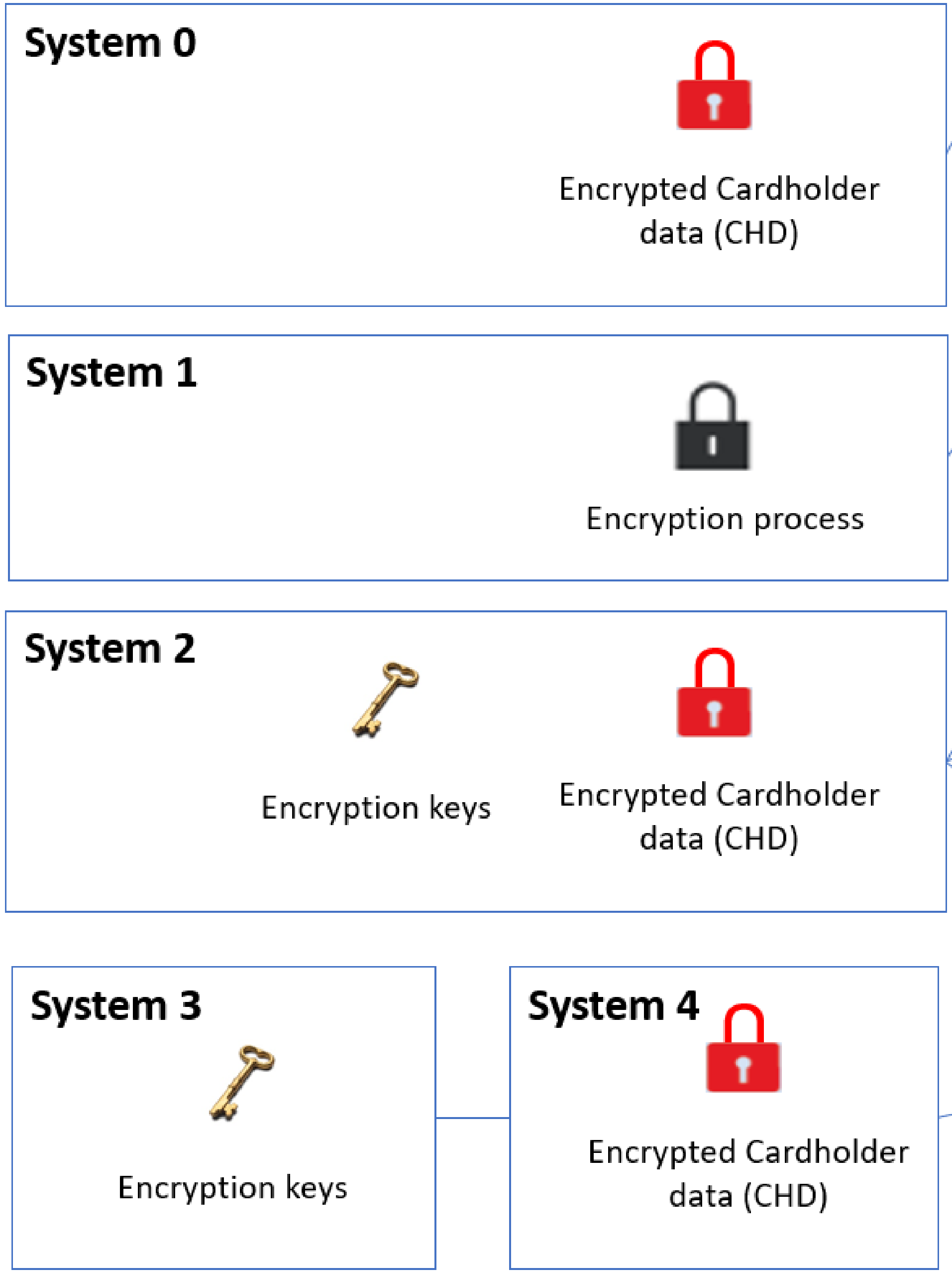
- Technically the data hosted by AWS Sarl is encrypted.
  - AWS Sarl, a Luxembourg registered company.
- The key is held by a trusted third party in France, not by AWS.

Other guarantees taken:

- No health data.
  - The data hosted relates only to the identification of individuals for the purpose of making appointments.
- Data is deleted after three months.

https://iapp.org/news/a/why-this-french-court-decision-has-far-reaching-consequences-for-many-businesses/

In scope for PCI DSS?



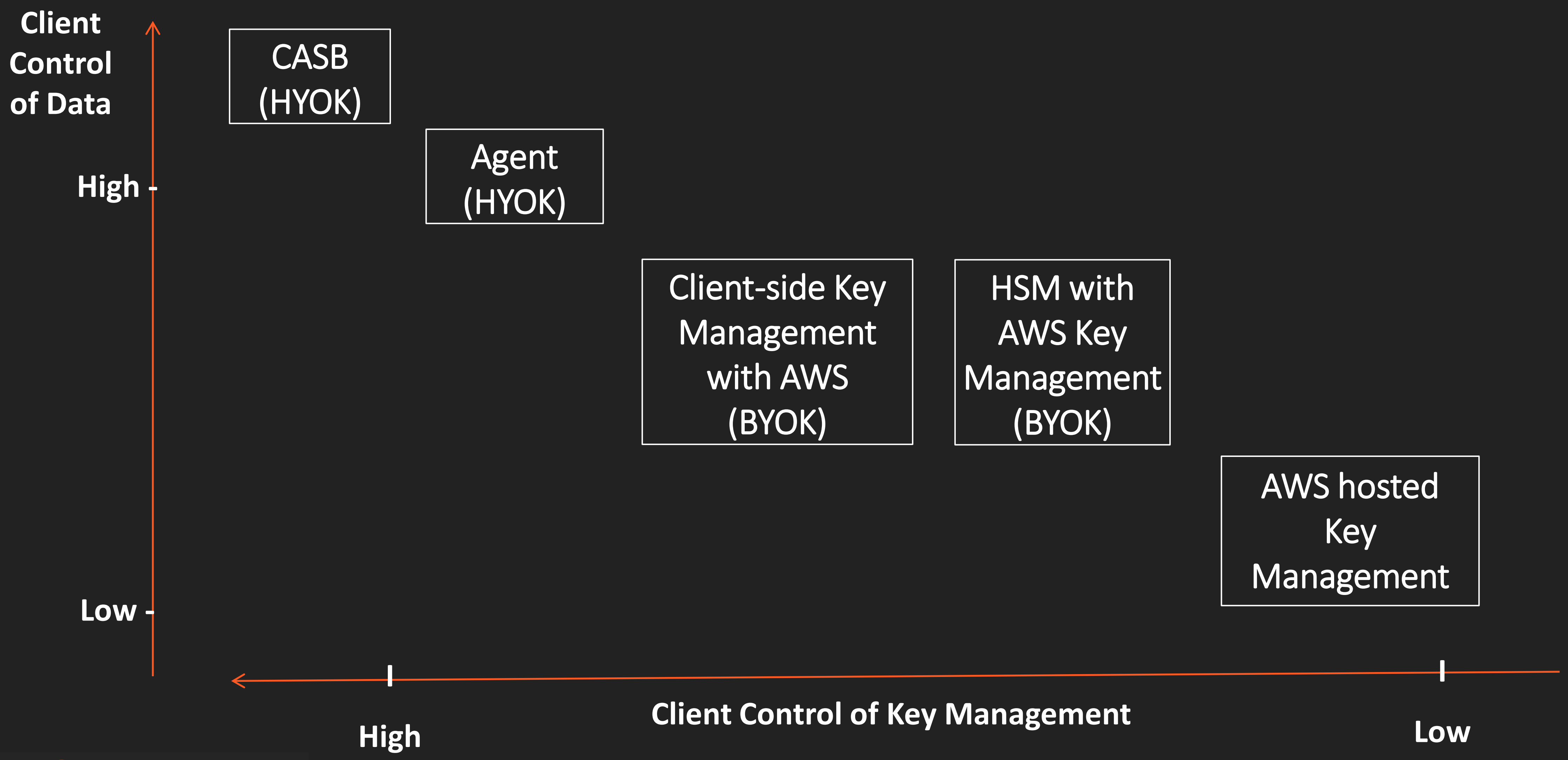
The following is NOT in scope for PCI DSS

The following are each in scope for PCI DSS:

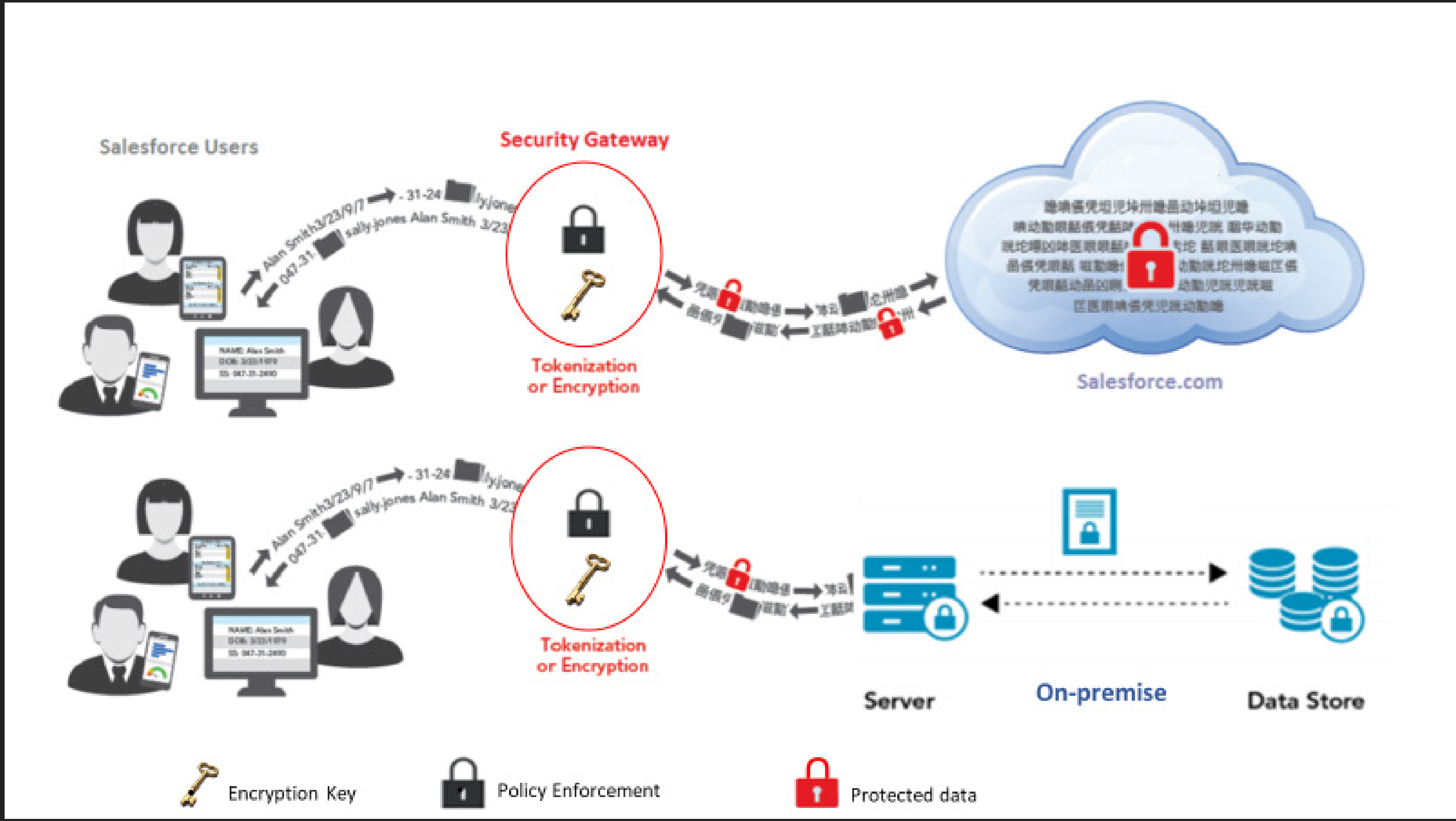
1. Systems performing **encryption** and/or decryption of cardholder data, and systems performing **key management** functions
2. **Encrypted** cardholder data that is **not isolated** from the encryption and decryption and **key** management processes
3. **Encrypted** cardholder data that is present on a system or media that also contains the decryption **key**
4. **Encrypted** cardholder data that is present in the same environment as the decryption **key**
5. **Encrypted** cardholder data that is **accessible** to an entity that also has access to the decryption **key**

https://blog.pcisecuritystandards.org

HYOK (Hold Your Own Key) vs AWS Key Management



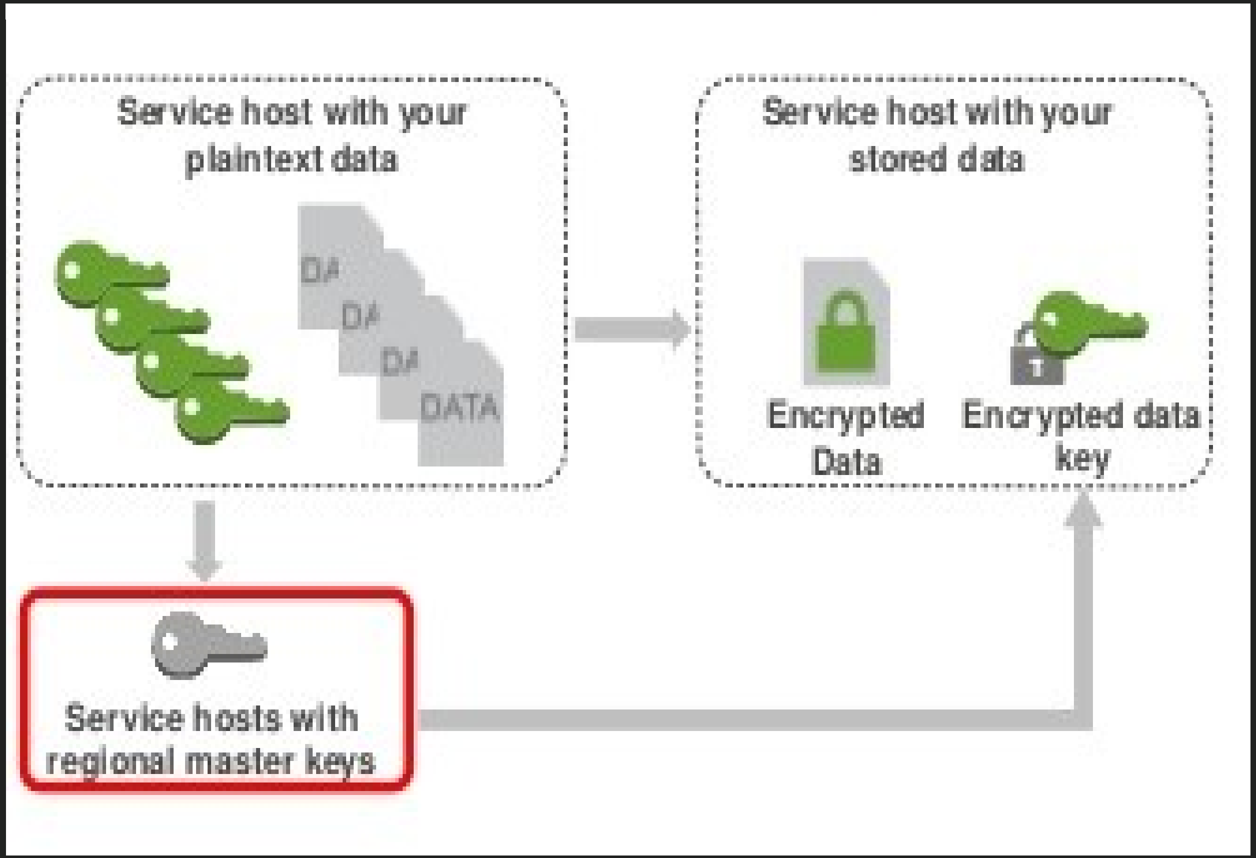
A Data Security Gateway Can Protect Sensitive Data in Cloud and On-premise



PROTEGRITY

AWS Key Management

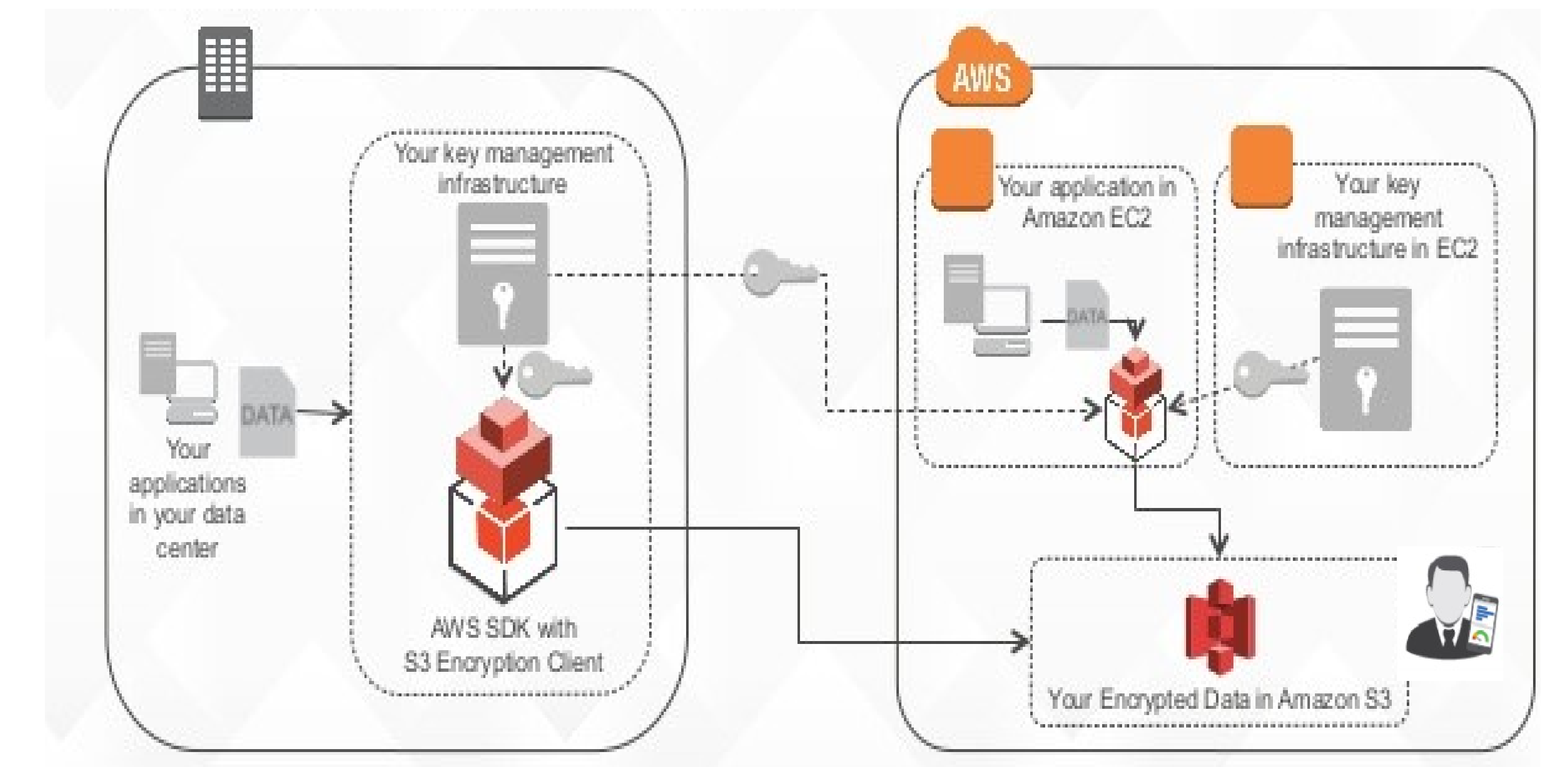
AWS hosted Key Management
(AWS Keys)



PROTEGRITY

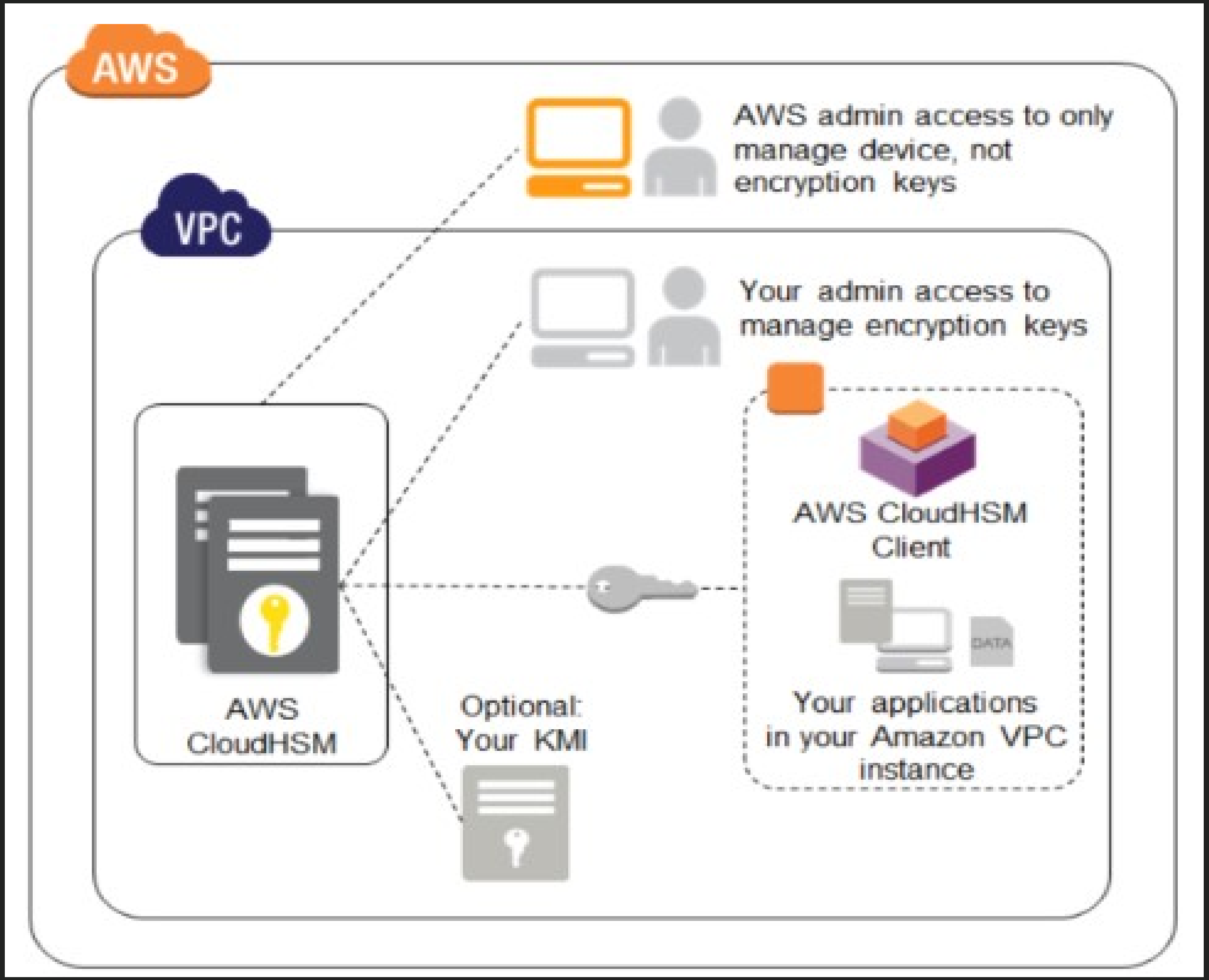
AWS Key Management

Client-side Key Management with AWS (BYOK)



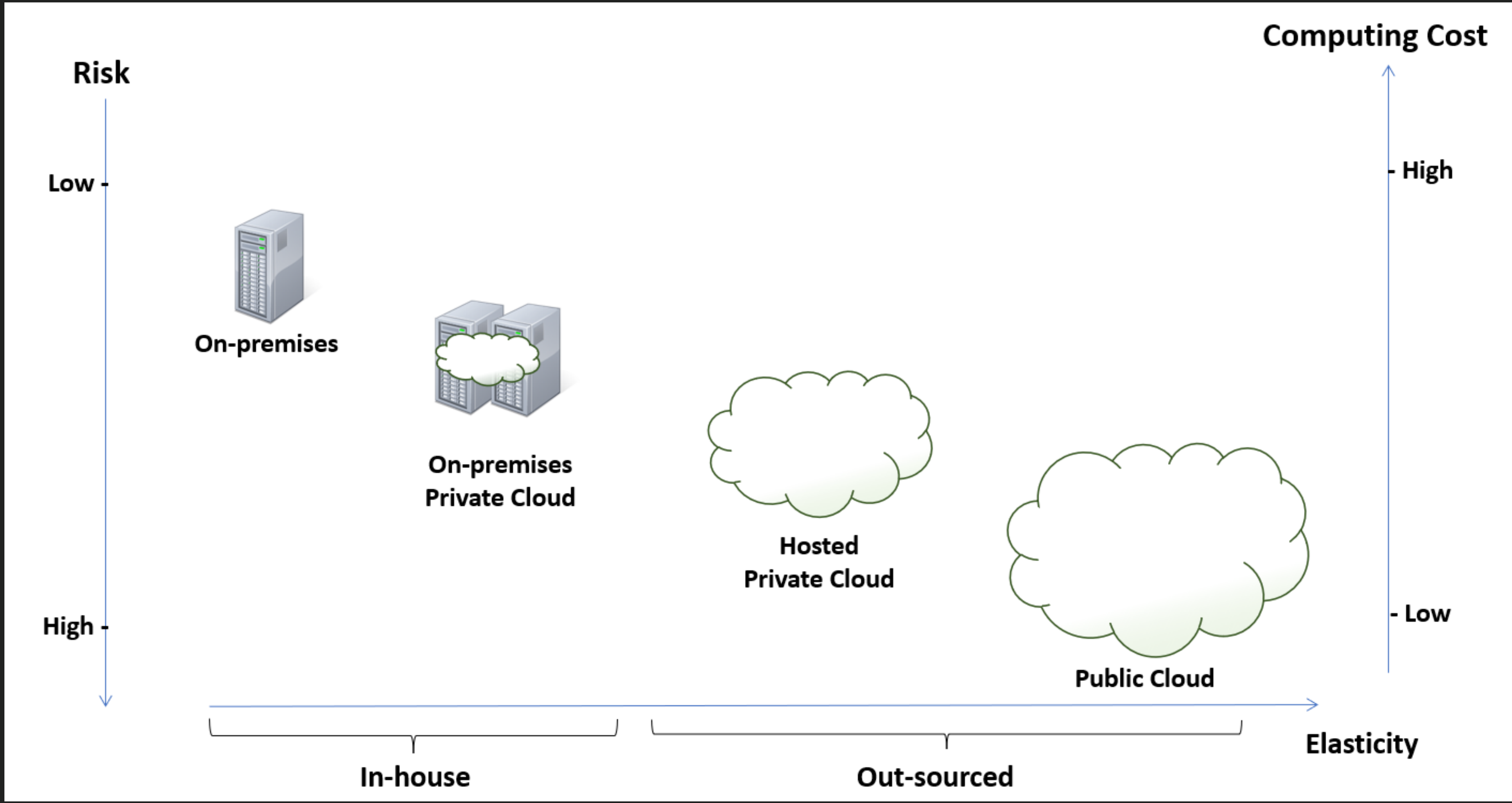
AWS Key Management

HSM with AWS Key Management (BYOK)

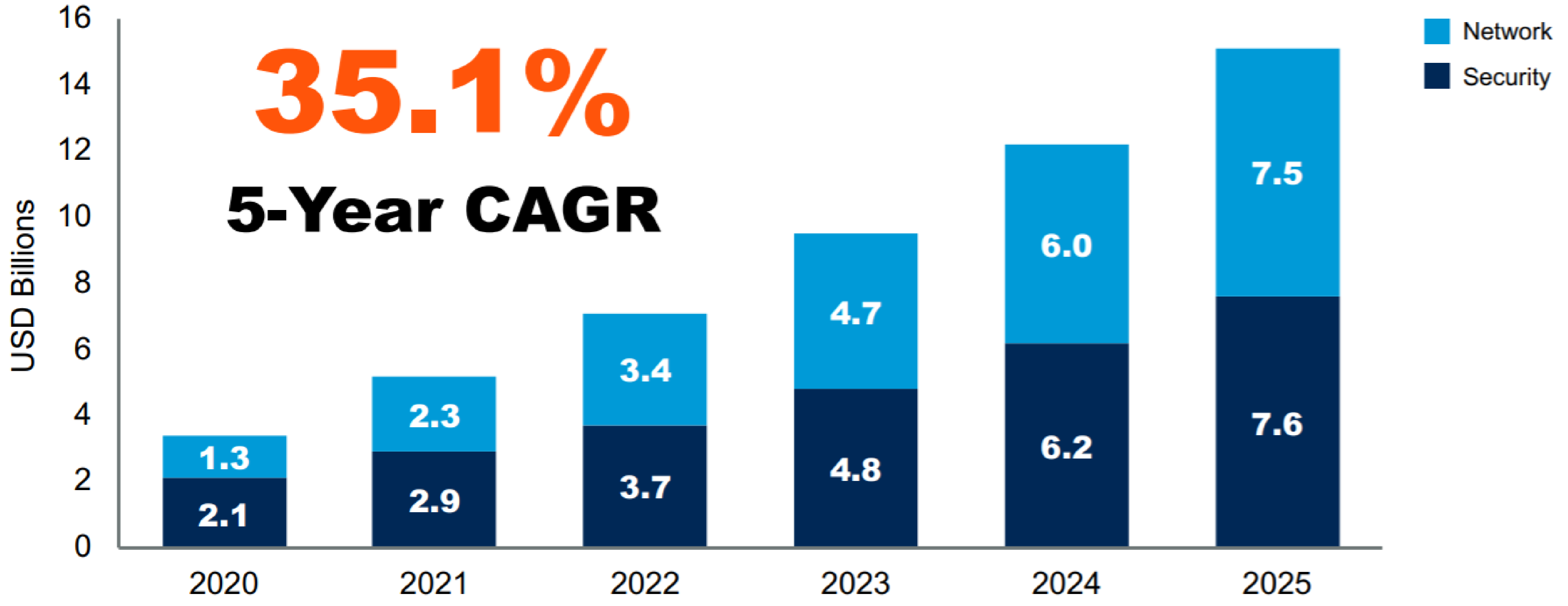


PROTEGRITY

Multi-cloud Risk

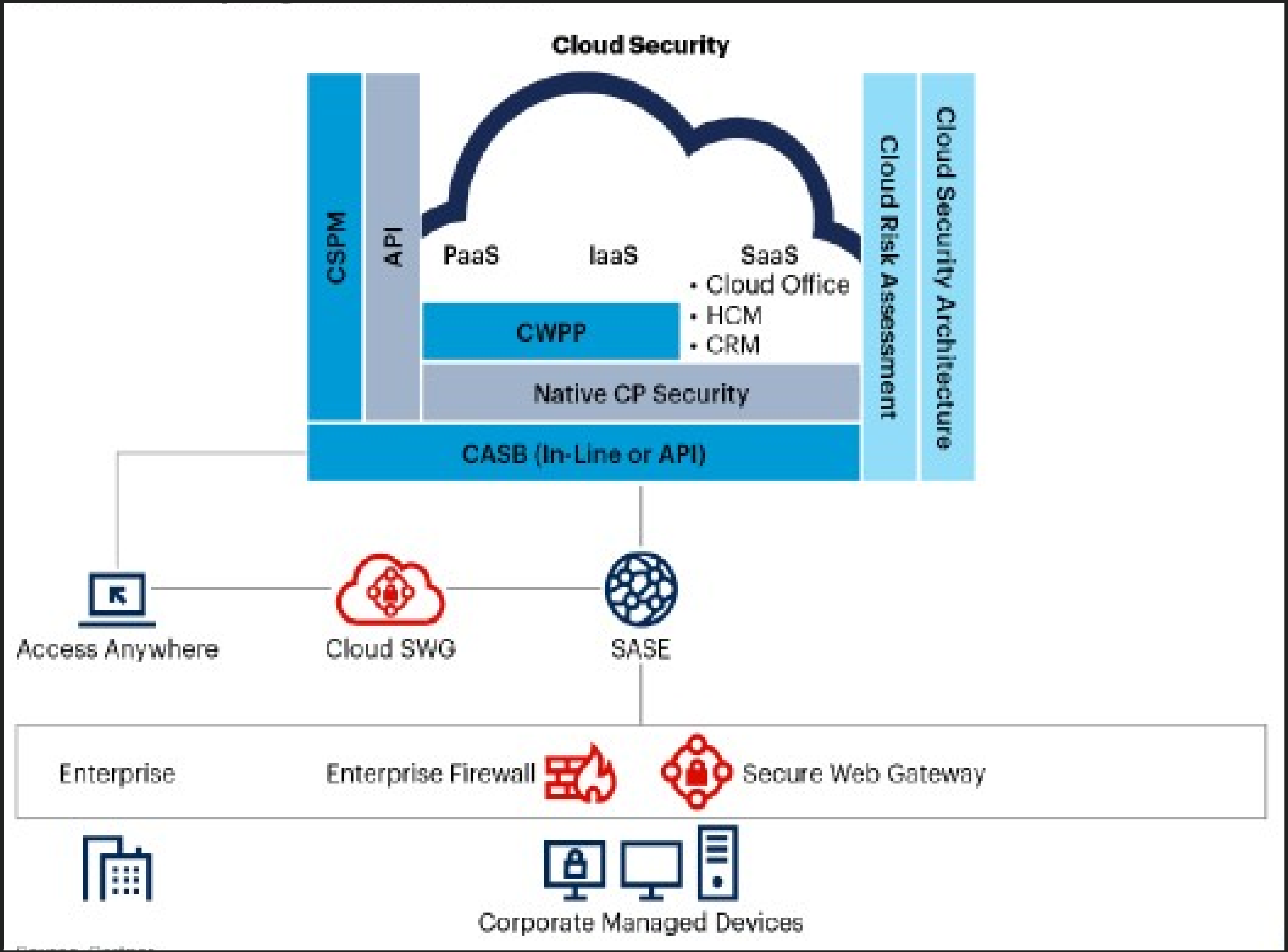


SASE Is One of the Fastest Growing Markets



Gartner

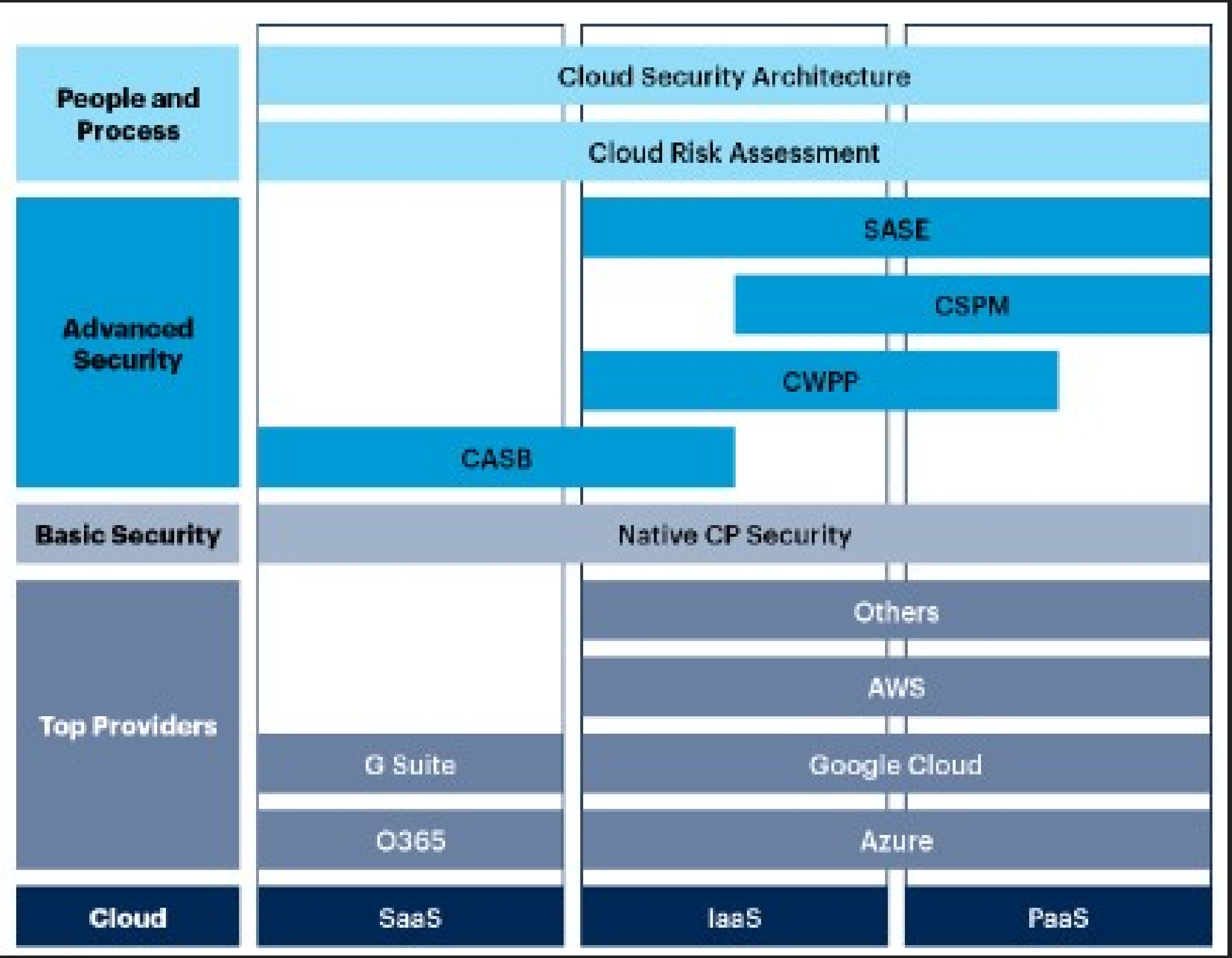
Cloud Security Logical Architecture



PROTEGRITY

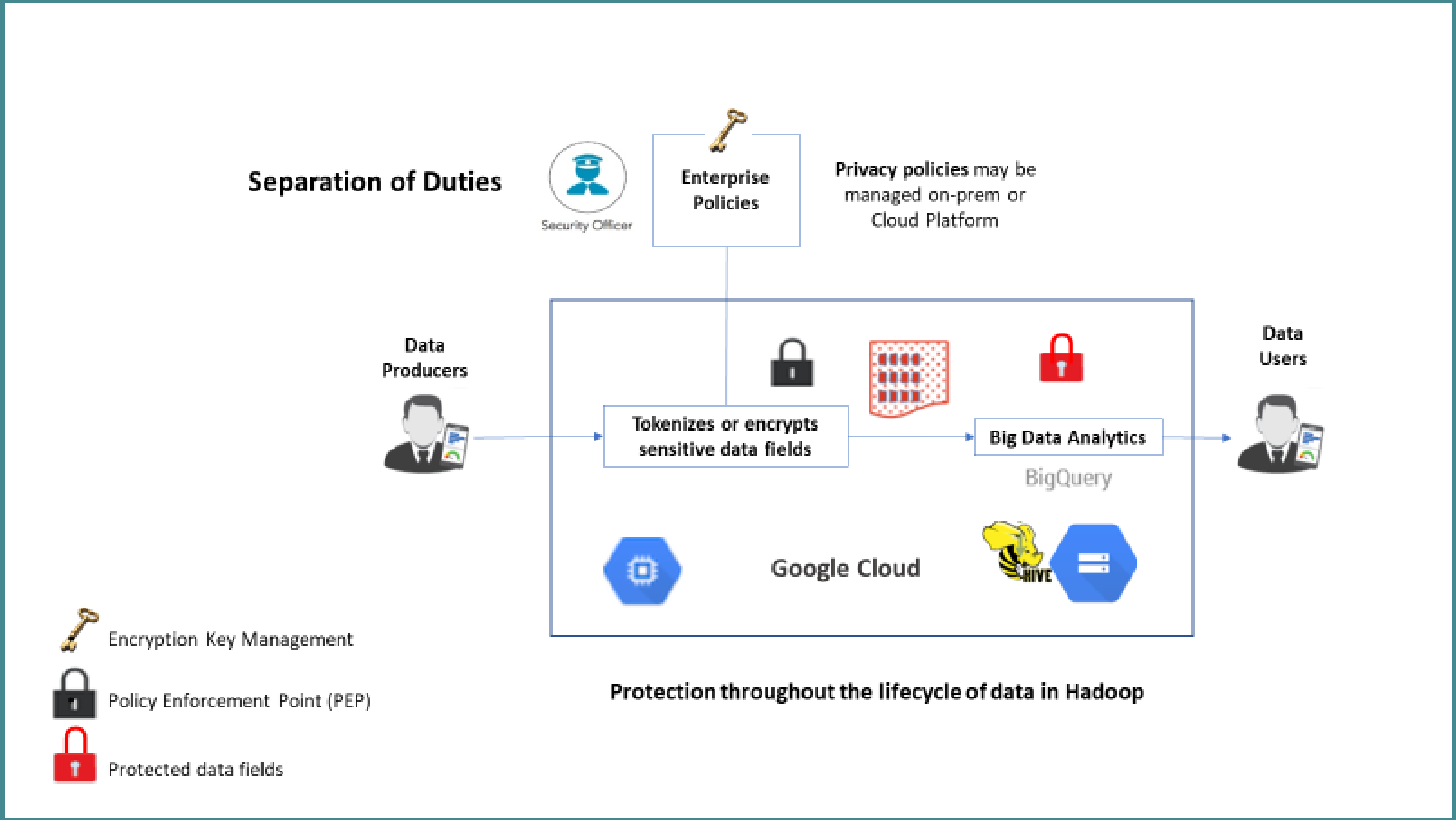
Gartner

Cloud Security Architecture

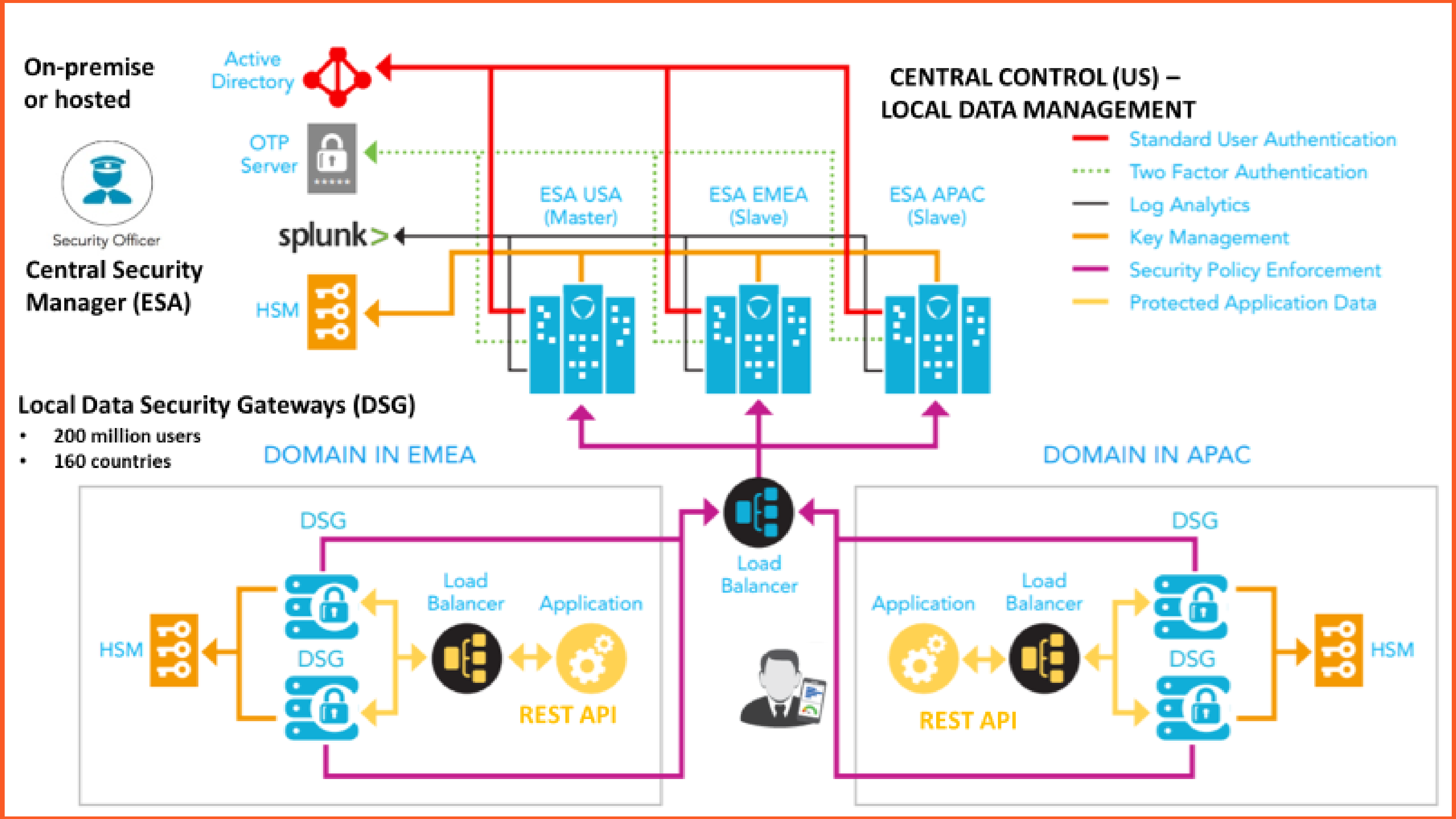


Gartner

Big Data Protection with Granular Field Level Protection for Google Cloud



Use Case (Financial Services) - Compliance with Cross-Border and Other Privacy Restrictions

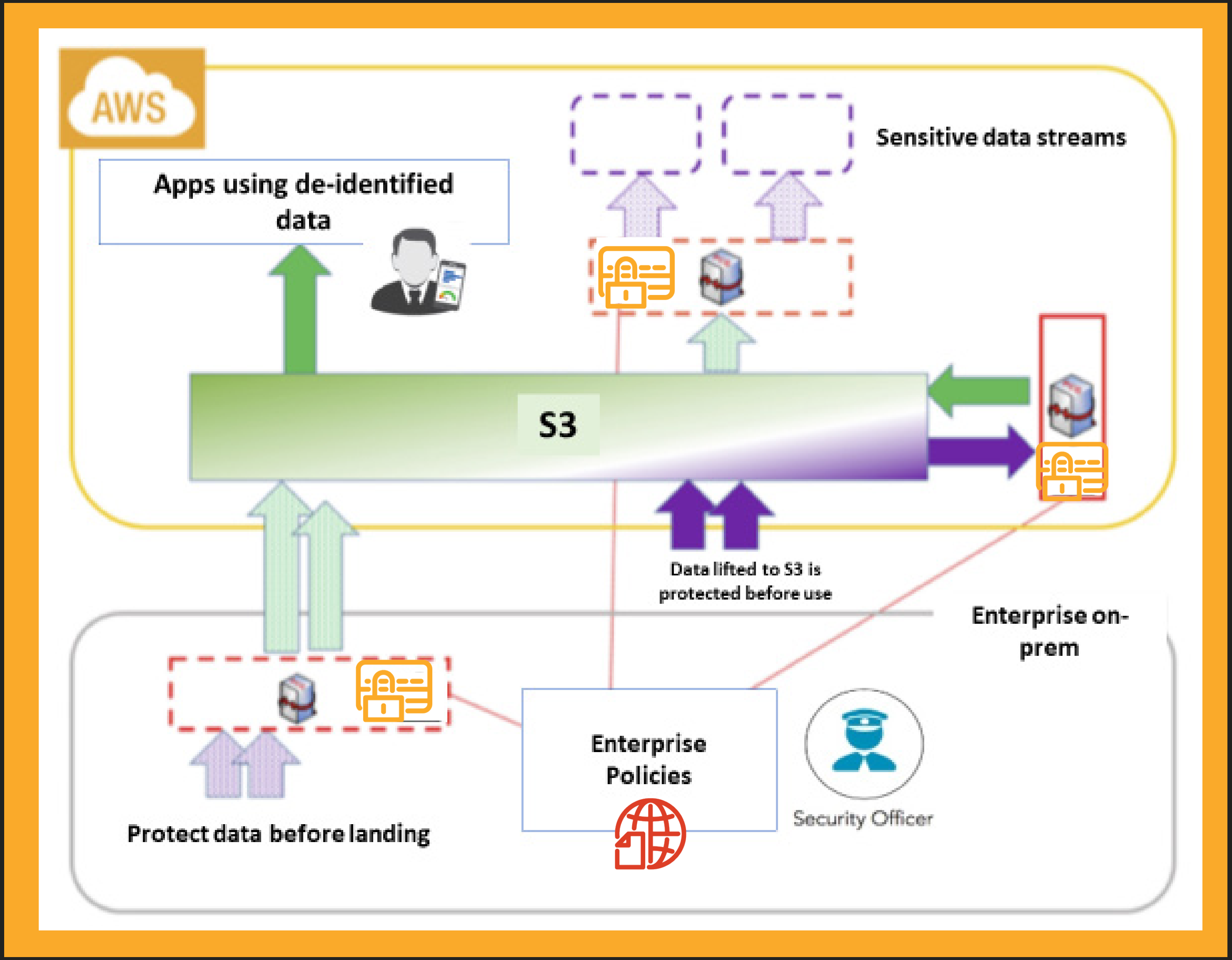


Protection of data in AWS S3 with Separation of Duties

- Applications can use de-identified data or data in the clear based on policies
- Protection of data in AWS S3 before landing in a S3 bucket

Separation of Duties

- Encryption Key Management
- Policy Enforcement Point (PEP)



PROTEGRITY

Data Security Management for Hybrid Cloud

Consistency

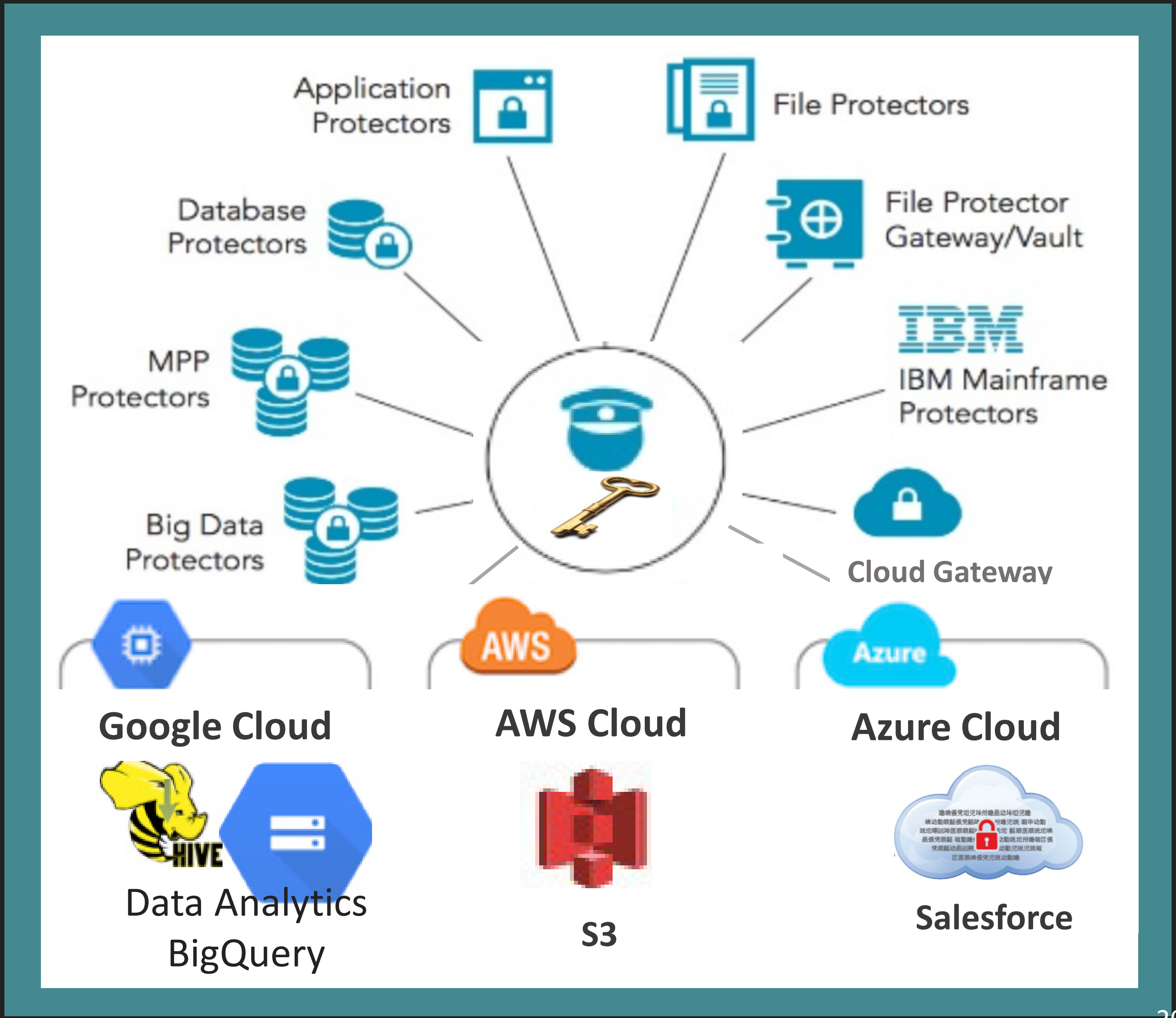
- Most firms are quite **familiar with their on-premises encryption and key management** systems, so they often prefer to leverage the **same tool and skills across multiple clouds**.
- Firms often adopt a **“best of breed” cloud** approach.

Trust

- Some customers simply do not trust their vendors.

Vendor Lock-in and Migration

- A common concern is **vendor lock-in**, and an inability to migrate to another cloud service provider.



Securosis, 2019

Some Major Cloud Databases

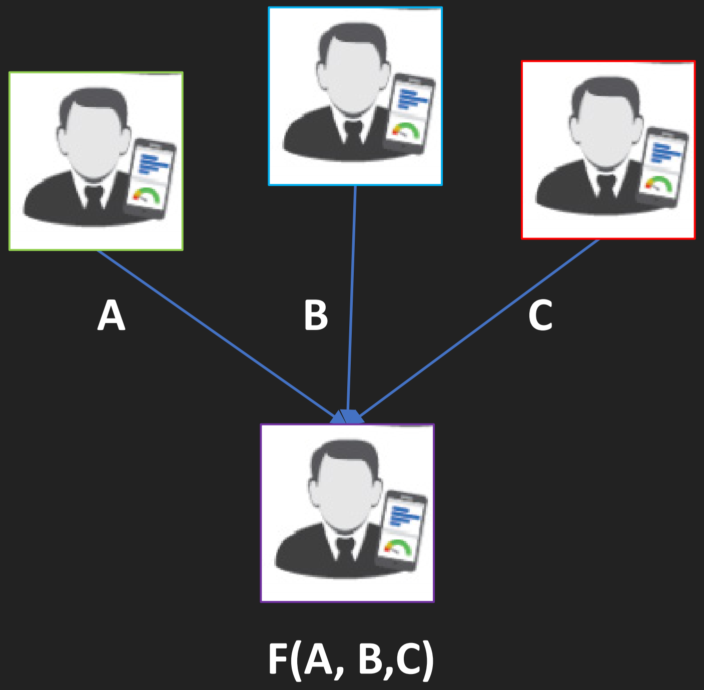
DB	Snowflake	Amazon Redshift	Azure Synapse Analytics	Google BigQuery
Sharing	Easy between different accounts.	Multiple data output formats		
Flexibility	Flexible	Less flexible		
Scale	Instant scaling. Easy to maintain	Scale up and scaledown manually		
Management	Easy to set up. Automated maintenance	Periodic vacuuming tables. Min administration		Automatic management. Intuitive
Data	Rows. Support for JSON	Column-oriented		columnar
Roll back		Roll-back on transactions		Cannot roll back on transactions
Integration	Ingesting fast. On-premise doesn't integrate	Largest cloud ecosystem. Much latency		
Speed		2 times faster than other		
Easy		User-friendly. Datalakes, easy		
Price starts	Storage costs separate. $2.01 per hour.	Pay as you use model. $0.25 per hour	Not cost-effective as others	Complicated. Cost separate for storage. Quereries cost $5/TB, add quickly
Storage		$306 per TB per month		$20 per TB per month
Security	Secure views and user-defined functions	Rich cloud services		B2B identity management with Oauth
		Encryption for client and server		
		Column-level access control.		
Support		Needs improvement.	Great support	
Stability		Robust. Some issues with stability		
Max columns		1,600 columns in a single table		10,000 columns
Workloads		Analytical not transactional		Great at big chunks in a small time. Data scientists and ML

Secure Multi-Party Computation

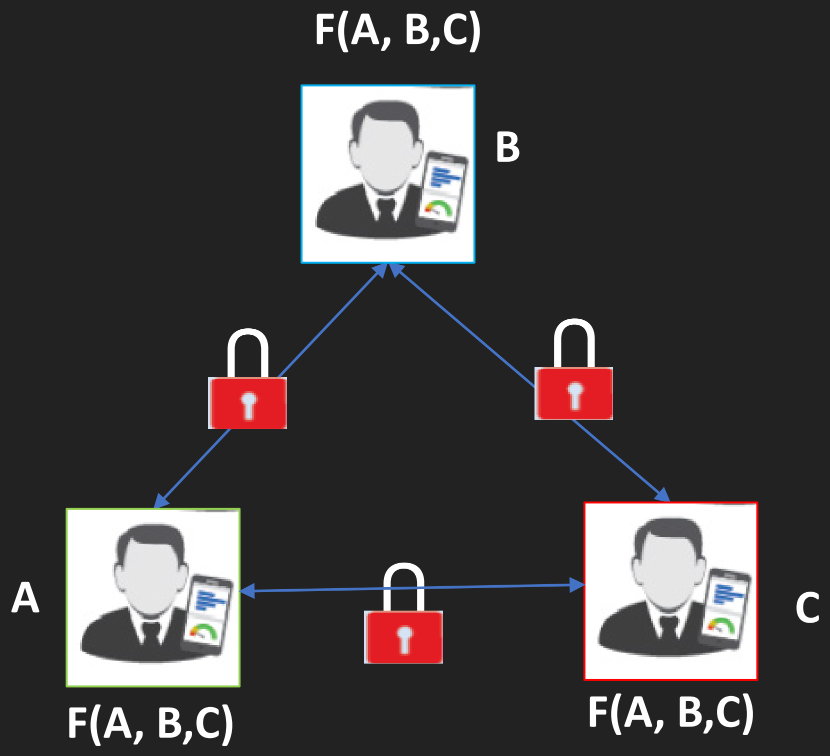
Secure Multi-Party Computation (MPC)

Private multi-party machine learning with MPC

Using MPC, **different parties send encrypted messages to each other**, and obtain the model F(A,B,C) they wanted to **compute without revealing their own private input**, and **without the need for a trusted central** authority.



Central trusted authority



Secure Multi-Party machine learning

https://royalsociety.org


Protected data fields

Case Study – HE and Securely sharing sensitive information

An example from the healthcare domain.

The recent ability to fully map the human genome has opened **endless possibilities for advances in healthcare.**

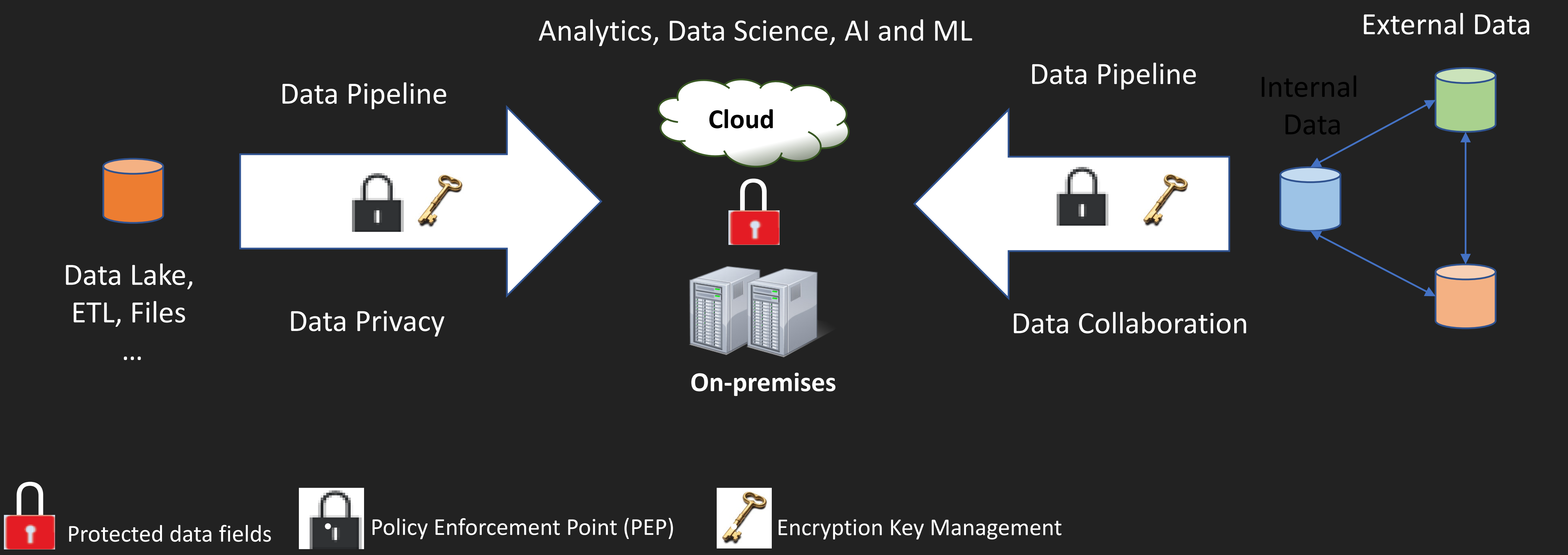
1. Data from **DNA analysis** can test for genetic abnormalities, empower **disease-risk analysis**, discover family history, and the presence of an Alzheimer's allele.
    - But these studies require **very large DNA sample sizes to detect accurate patterns**.
2. However, **sharing personal** DNA data is a particularly problematic domain.
    - Many citizens hesitate to **share such personal information with third-party providers**, uncertain of if, **how and to whom the information might be shared downstream**.
3. Moreover, **legal limitations designed to protect privacy** restrict providers from sharing this data as well.
4. **HE** techniques enable citizens to share their genome data and **retain key privacy concerns** without the traditional all-or-nothing trust threshold with third-party providers.

Analytics and AI

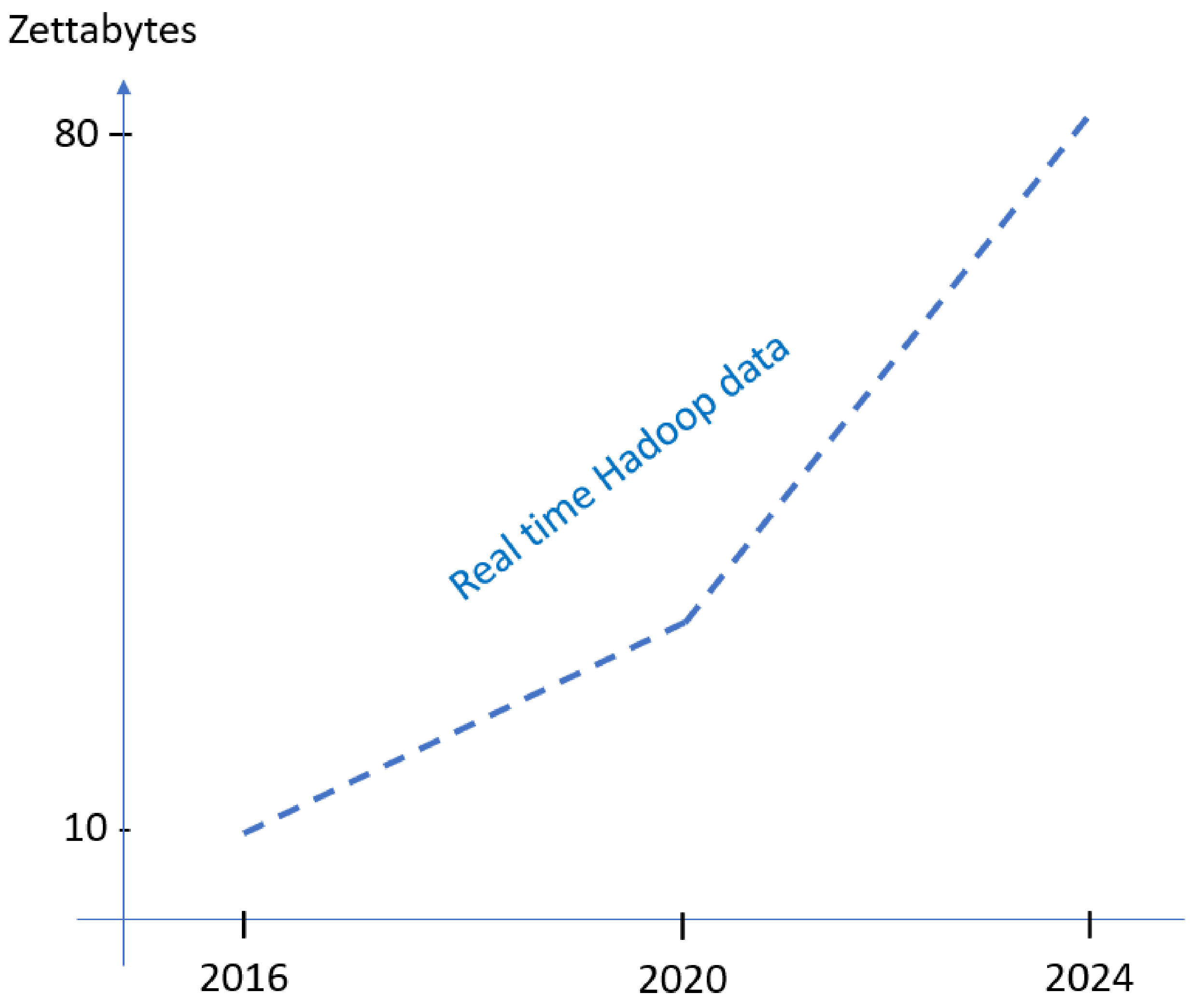
Increased need for data analytics drives requirements.

Internal and Individual Third-Party Data Sharing

Secure Multi Party Computation



Global Hadoop Big Data Analytics Market



Real-time data is significant in global datasphere

Between 2018 and 2025 the size of real-time data in the global datasphere is expected to expand tenfold, from five zettabytes to 51 zettabytes.

Source: Adapted from Maximize Market Research

Analytics and AI

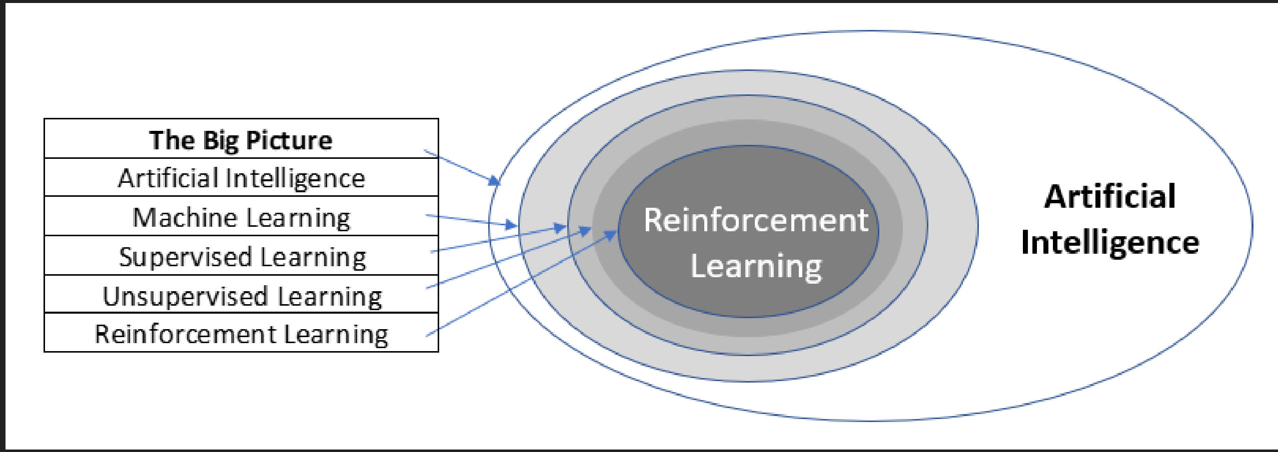
Feelings about Impact of New Technologies

	%			
Area	**Positive**	**Negative**	**Don't know**	**Equal**
Personalized medicine	50	10	25	15
Driverless cars	44	20	26	11
AI	44	20	26	11
Gene editing	41	14	27	19
Blockchain	35	18	25	21

Source: Adapted from Edelman Trust Barometer

What did we do before Machine Learning?

Artificial Intelligence and Machine Learning



Increased need for Data Analytics

Reduce Risk

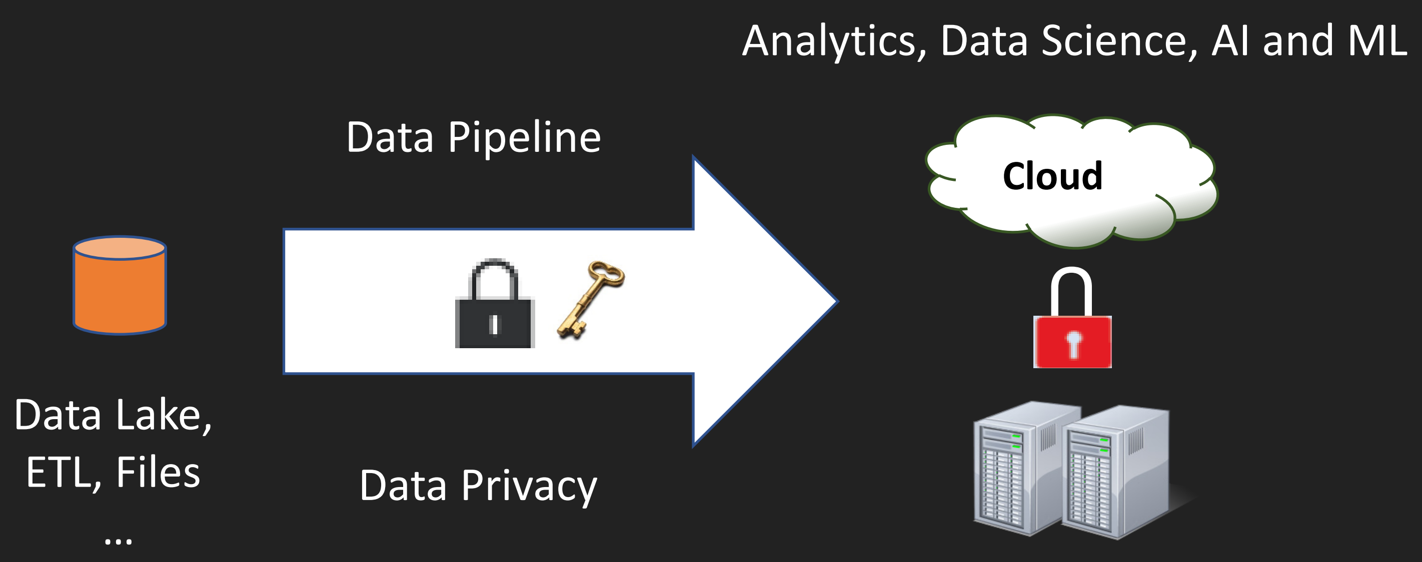
- **Secure AI & ML**

Use-cases

- **Analysis**
- **Insight**
- **Dashboarding**
- **Reporting**
- **Predictions**
- **Forecasts**
- **Simulation**
- **Optimization**

Values

- **Savings**
- **Revenue add**



Anonymization to minimize the risk of identification

Examples in Banking Credit Card Approval,

- **Reducing the risk from 26% down to 8%**
- **98% accuracy compared to the Initial Model**

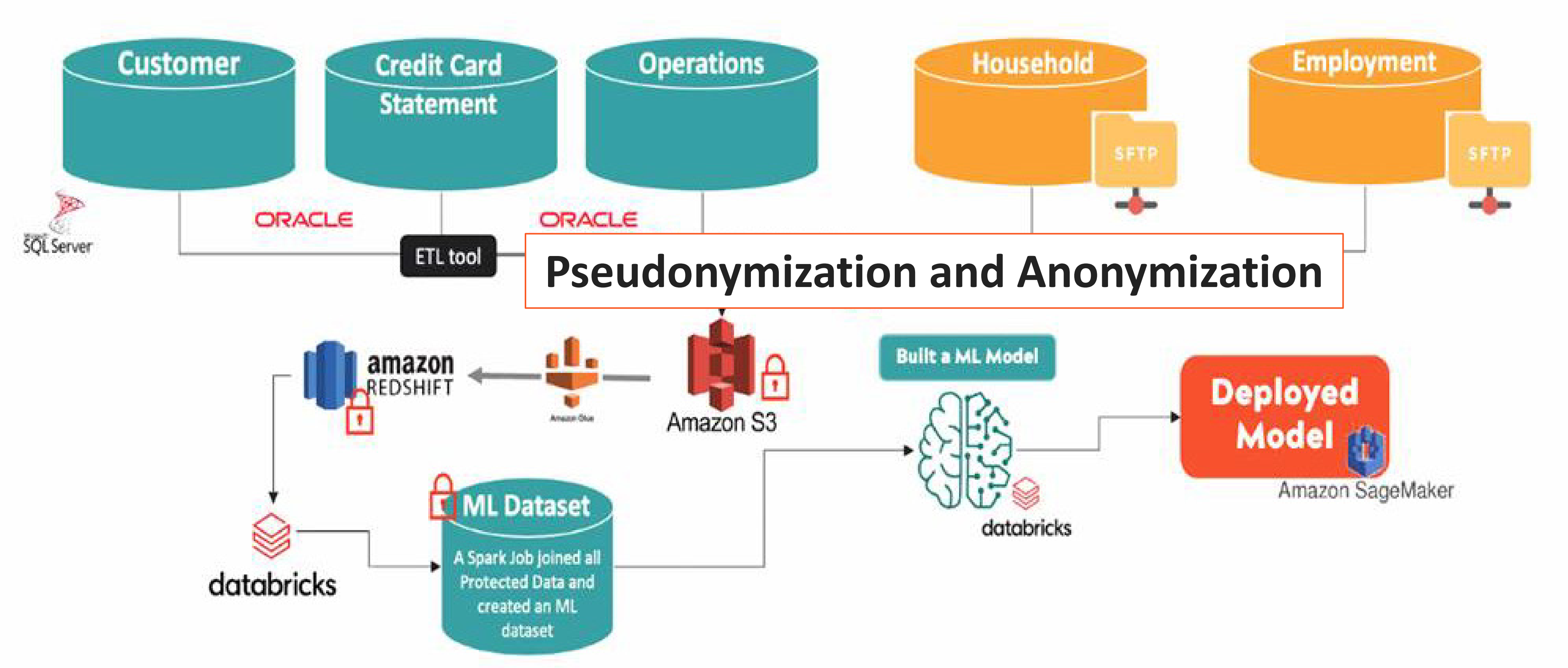
Secure AI & ML

Use-cases

- Analysis
- Insight
- Dashboarding
- Reporting
- Predictions
- Forecasts
- Simulation
- Optimization

Values

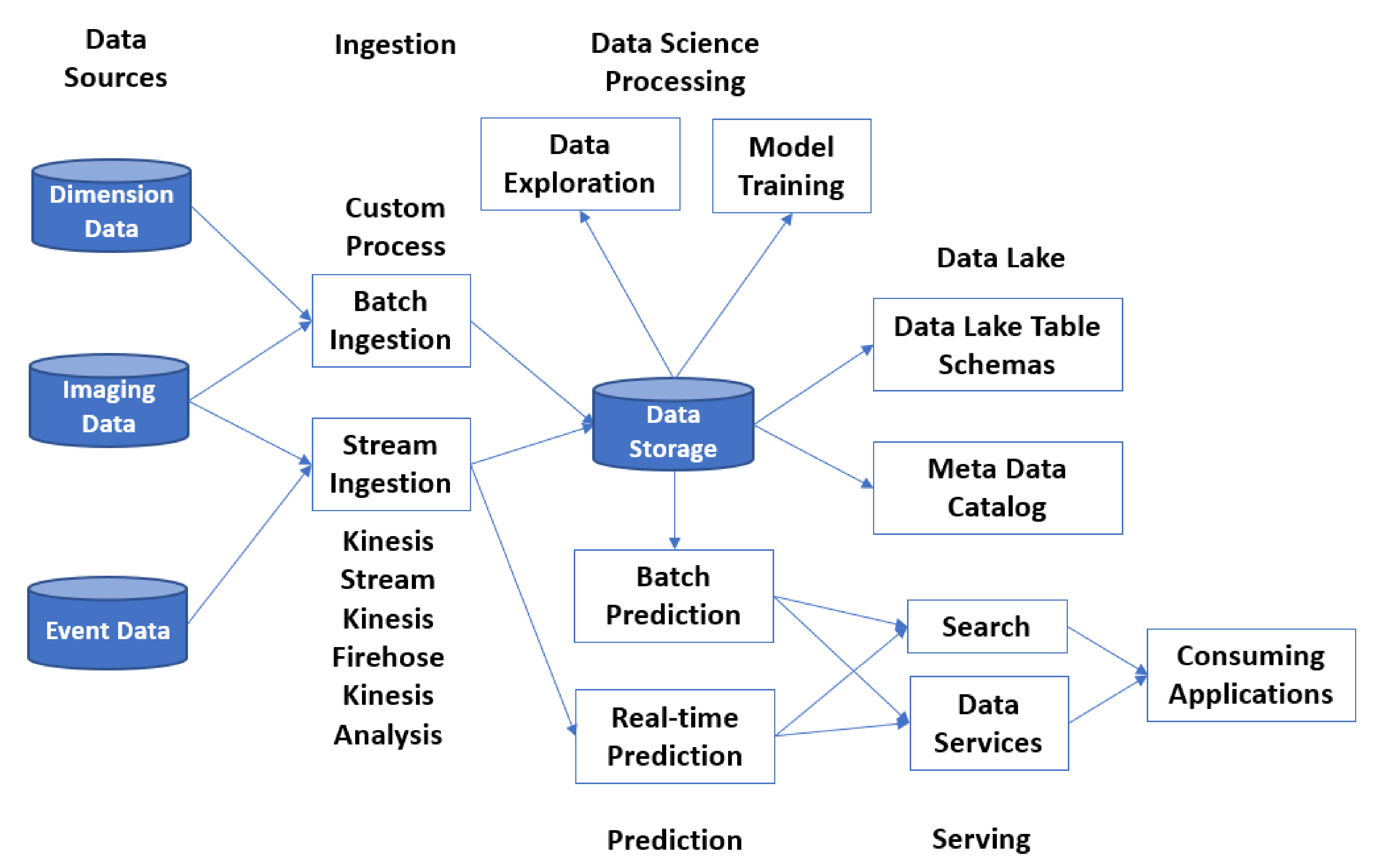
- Savings
- Revenue add



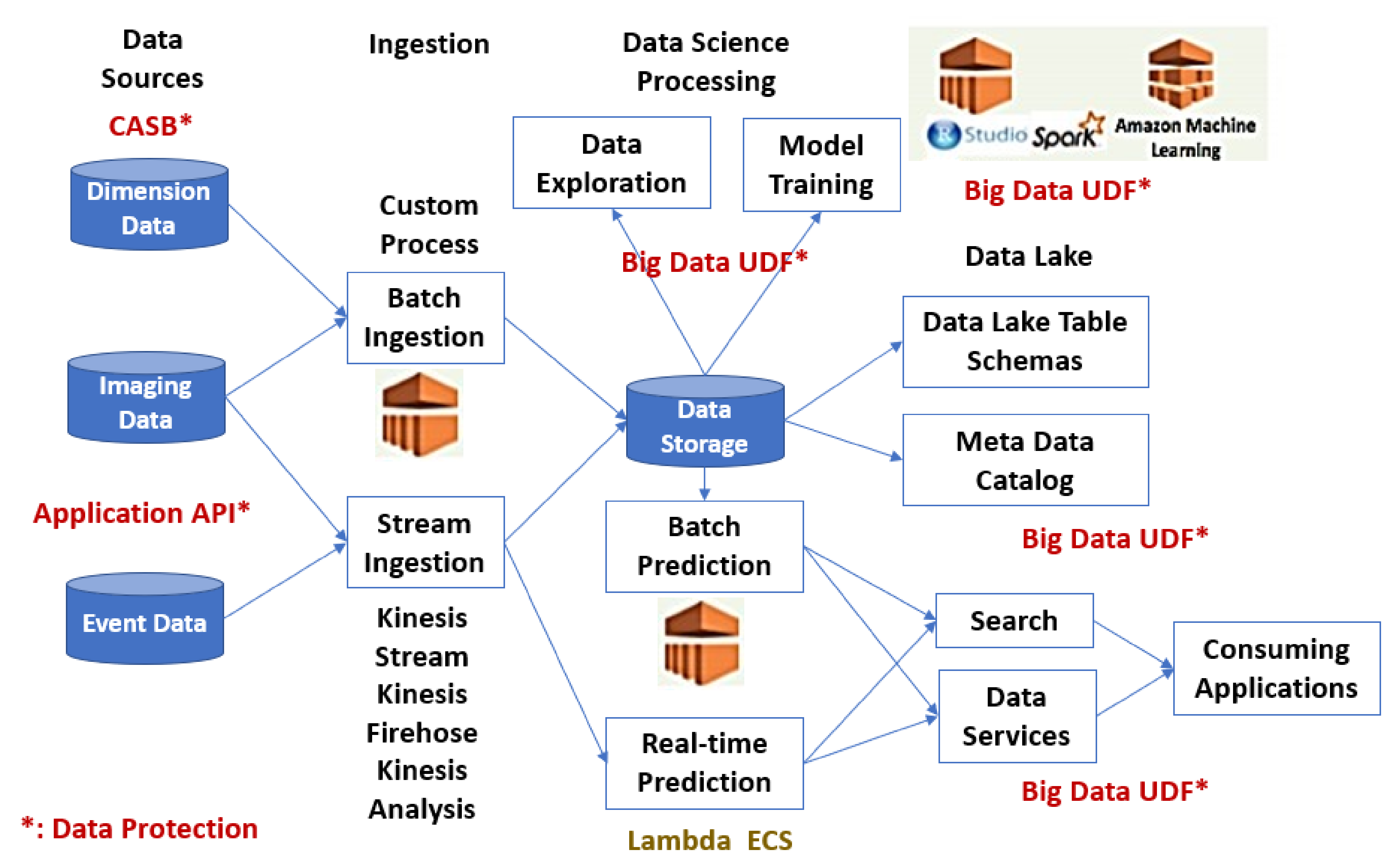
Anonymization to minimize the risk of identification

- Examples in Banking Credit Card Approval,
- Reducing the risk from 26% down to 8%
- 98% accuracy compared to the Initial Model

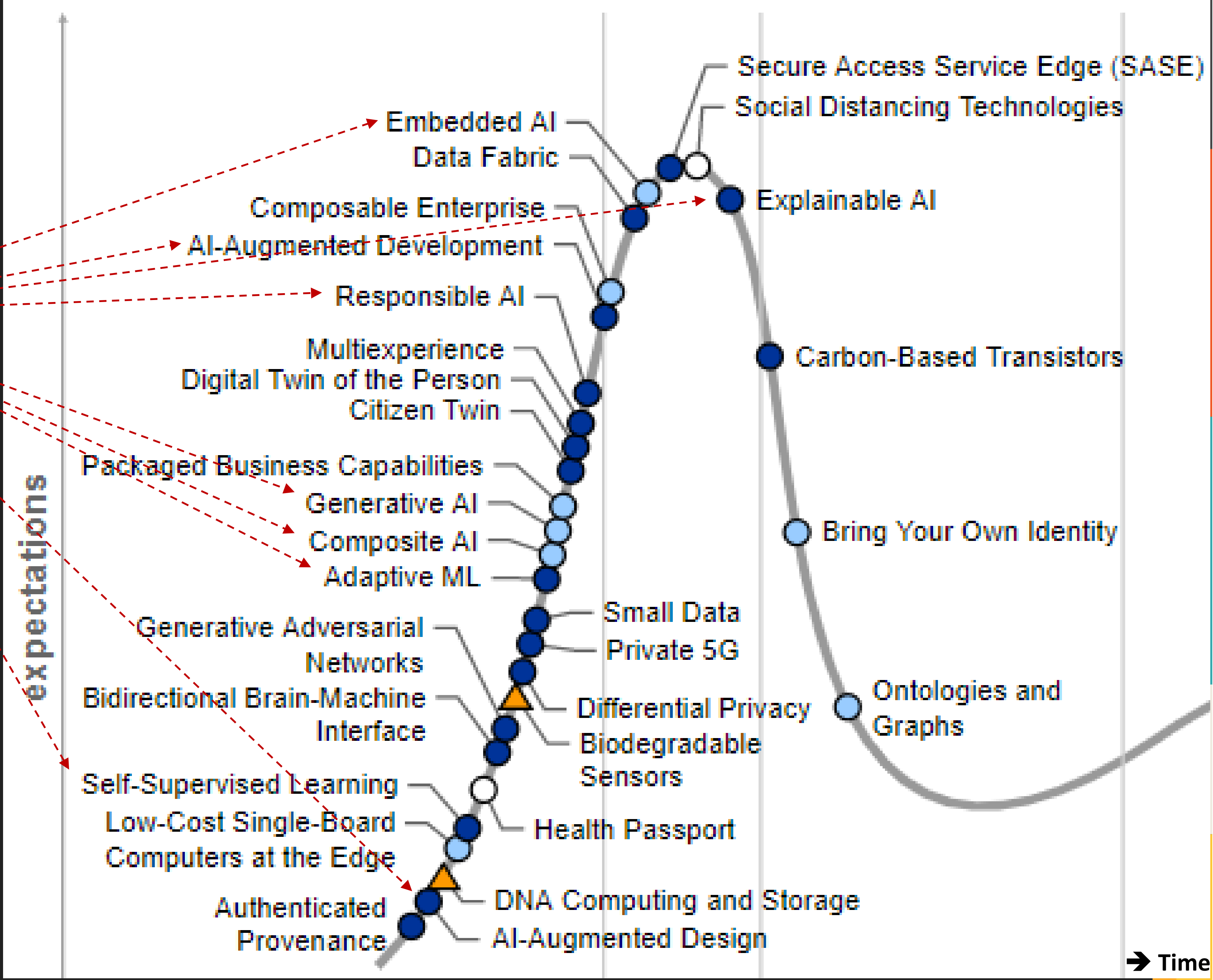
Conceptual Reference Architecture for Machine Learning



Architecture for Machine Learning - Data Protection



AI & ML

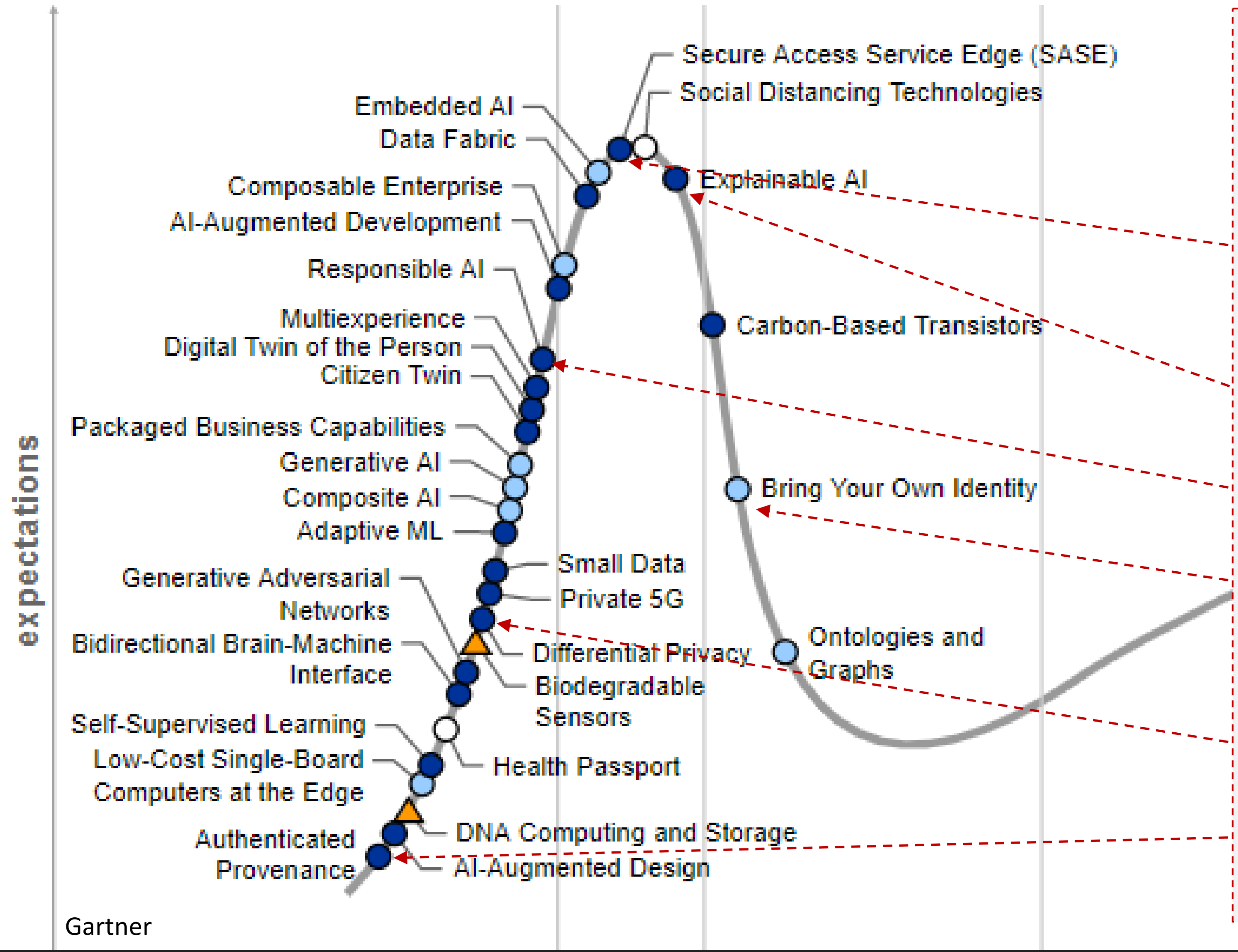


Gartner Hype Cycle for Emerging Technologies, 2020

Gartner

PROTEGRITY

Gartner Hype Cycle for Emerging Technologies, 2020



Emerging technologies tied to **algorithmic trust** include

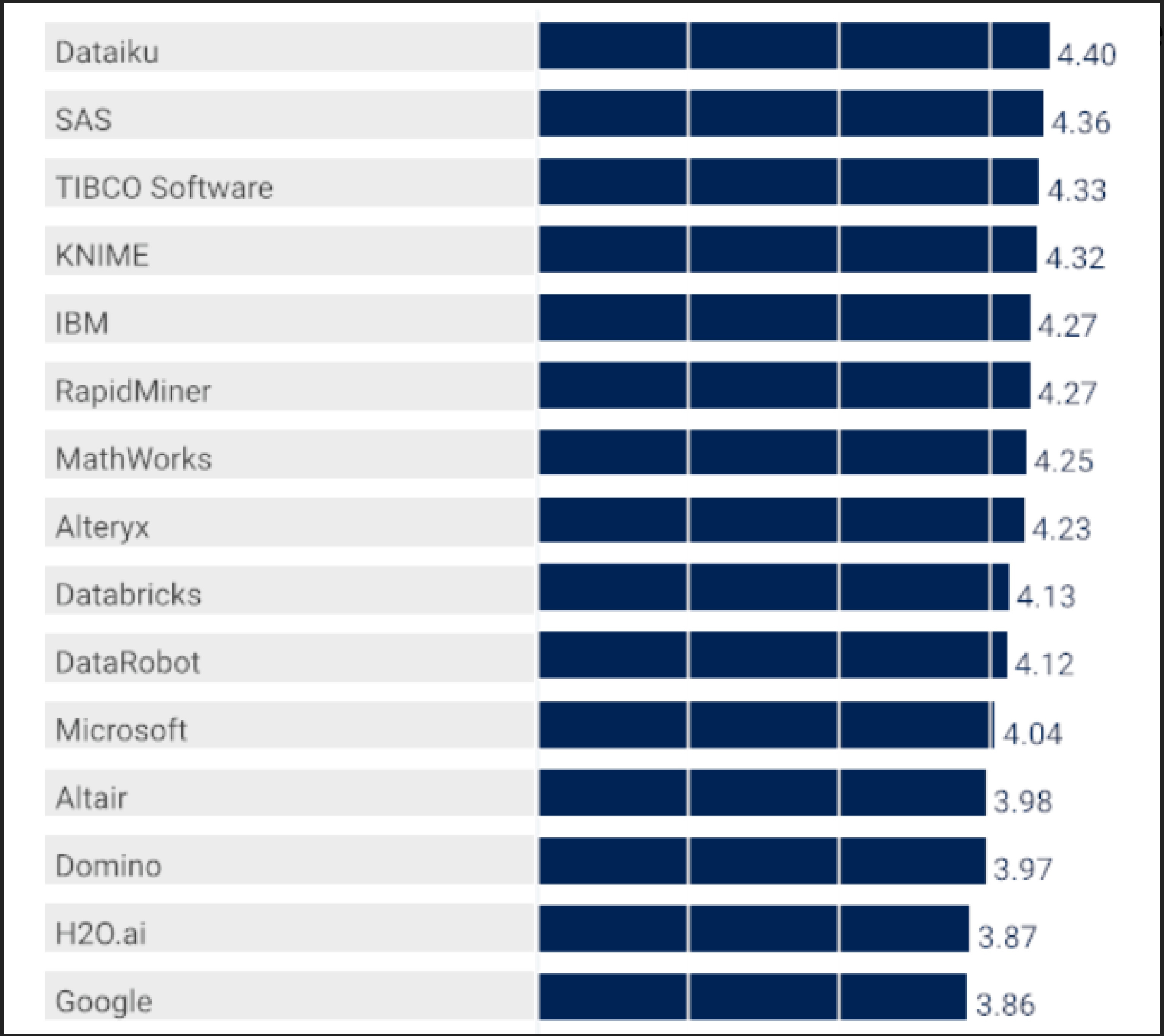
1. **Secure access service edge (SASE)**
2. **Explainable AI**
3. **Responsible AI**
4. **Bring your own identity**
5. **Differential privacy**
6. **Authenticated provenance**

Algorithmic Trust Models Can Help Ensure Data Privacy

cmswire.com

Products and Services Scores for Business and Data Exploration,

Gartner



PROTEGRITY

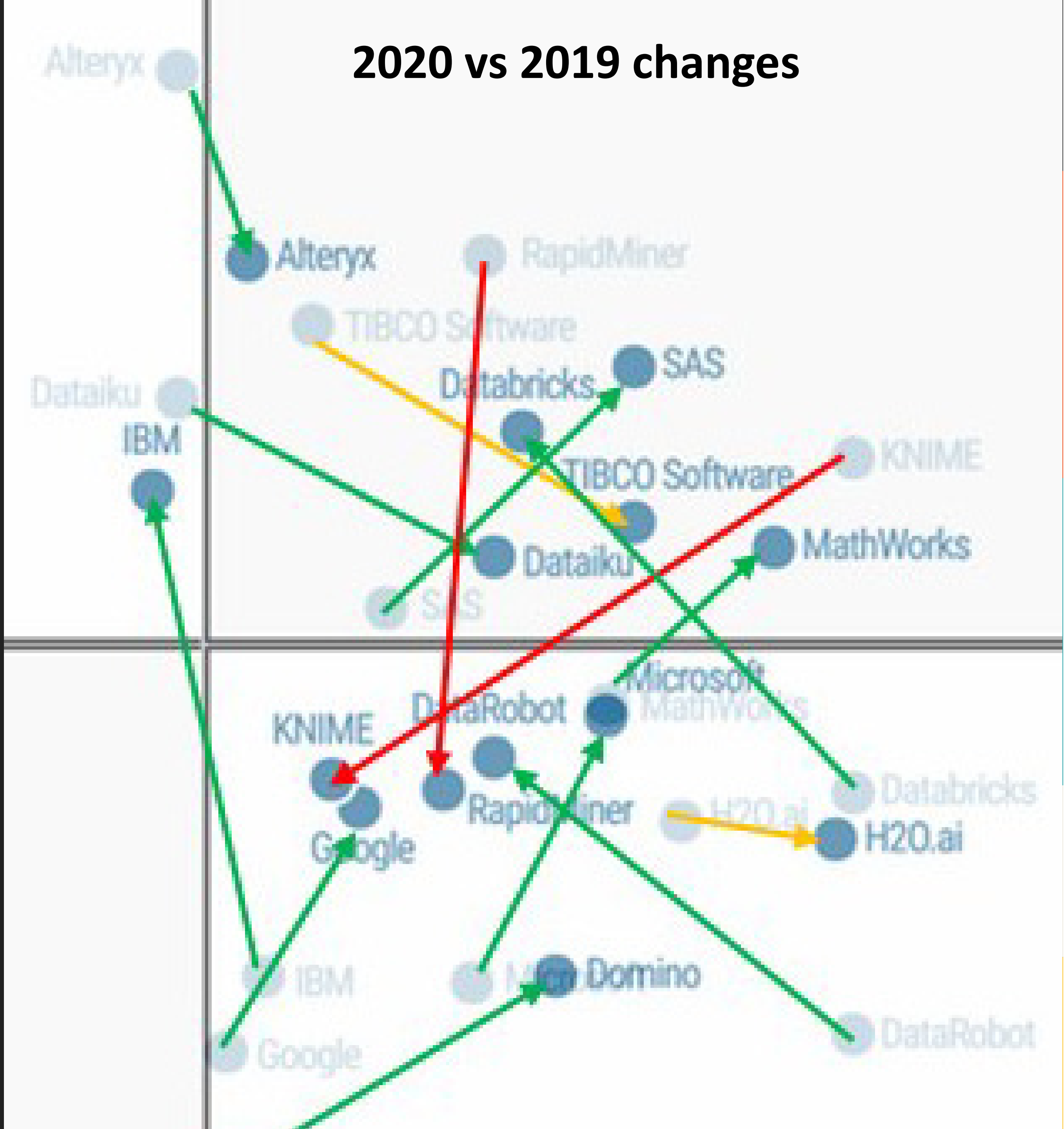
Gartner MQ for Data Science and Machine Learning Platforms

Data and analytics pipeline, including all the following areas:

1. Data ingestion
2. Data preparation
3. Data exploration
4. Feature engineering
5. Model creation and training
6. Model testing
7. Deployment
8. Monitoring
9. Maintenance
10. Collaboration

https://www.kdnuggets.com/2020/02/gartner-mq-2020-data-science-machine-learning.html

2020 vs 2019 changes



Gartner MQ for Data Science and Machine Learning Platforms, 2021

Data and analytics pipeline, including all the following areas:

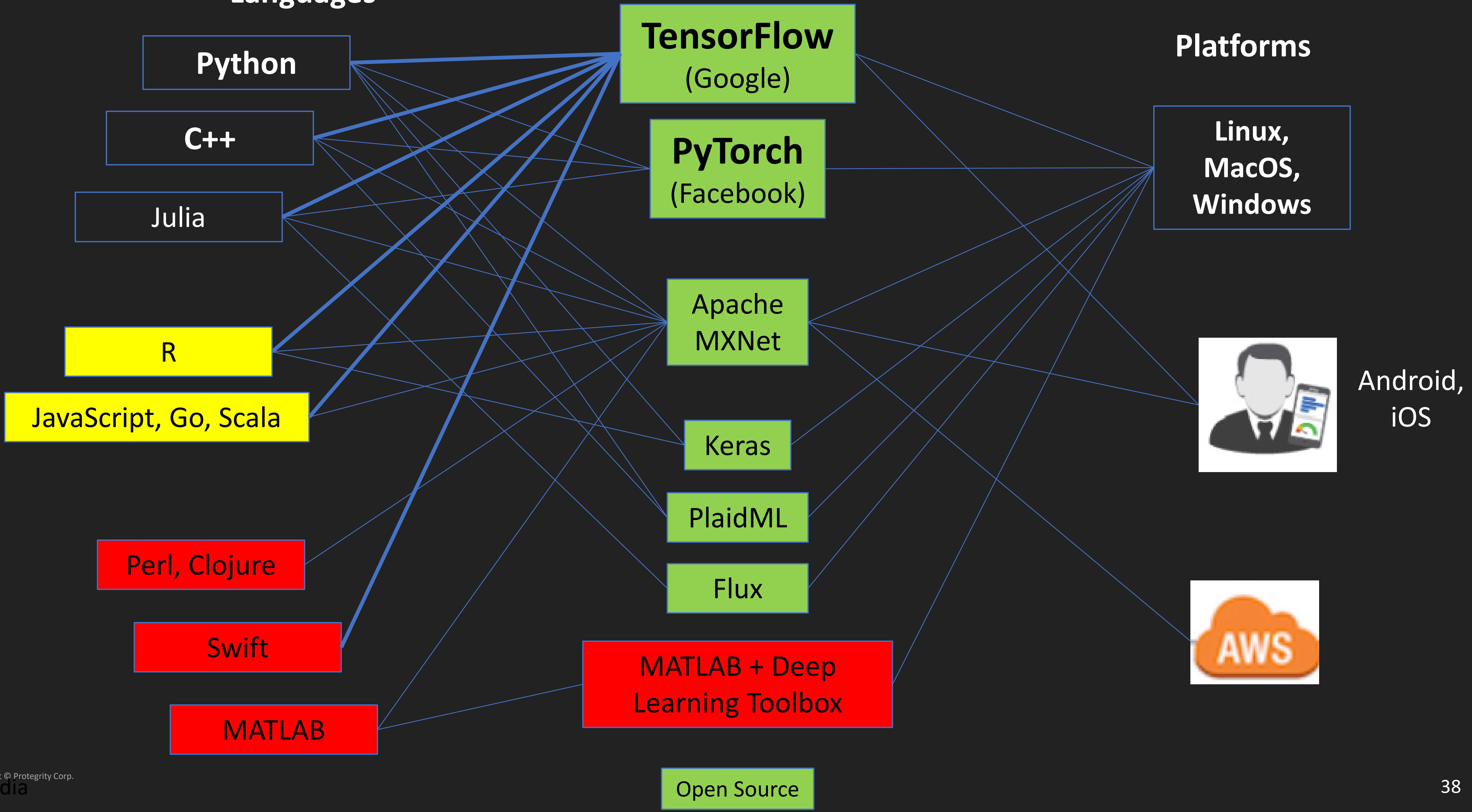
1. Data ingestion
2. Data preparation
3. Data exploration
4. Feature engineering
5. Model creation and training
6. Model testing
7. Deployment
8. Monitoring
9. Maintenance
10. Collaboration

Machine Learning
Deep-learning

Frameworks & libraries

Languages

Platforms

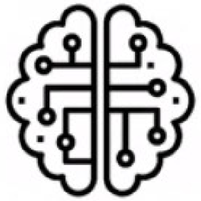


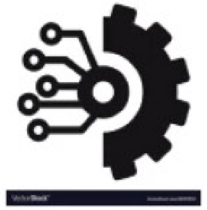
wikipedia

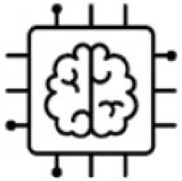
PROTEGRITY

Responsible AI

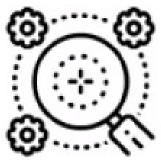
Aims to address the Trust and Ethics related to decisions made by AI models


Ethics & Fairness


Interpretability

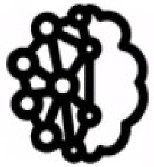

Privacy

Data Generalization



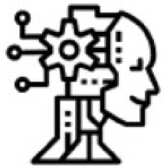
- Differential Privacy
- k-anonymity l-diversity t-closeness

Synthetic Data



- GANs
- VAE
- GANs with DP

Private AI Models

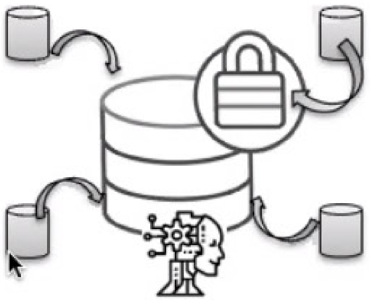


- Torch with DP
- Scikit-learn DP

Confidential AI

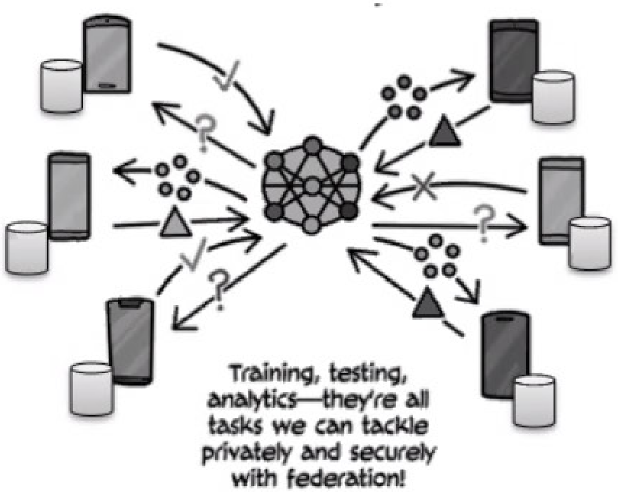
Aims to enable centralized and decentralized AI consortiums safely

ML on Encrypted Data



- Queries as-a-service
- AI Models as-a-service

Federated Learning

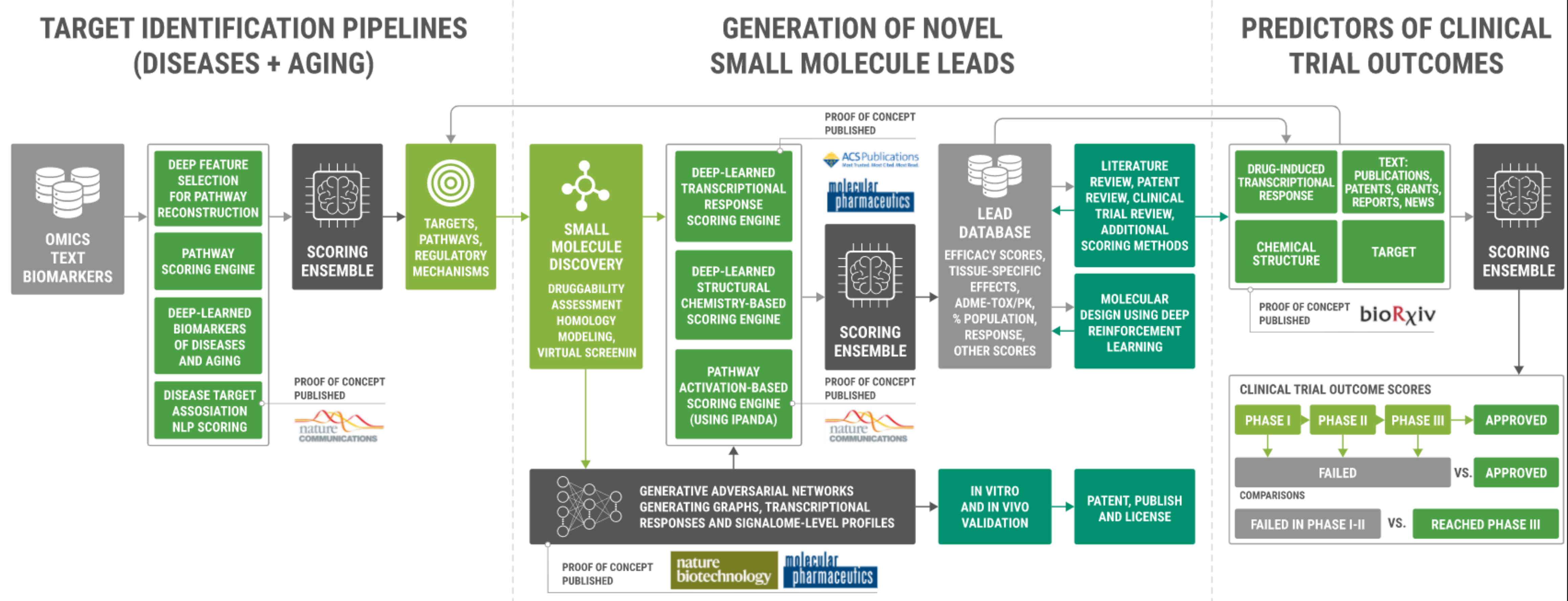


Use Case: Insilico Medicine

An alternative to animal testing for research and development programs in the pharmaceutical industry.

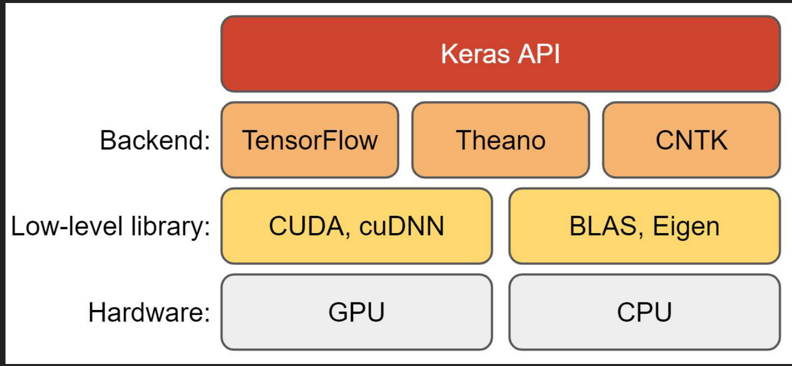
- By using **artificial intelligence** and deep-learning techniques, Insilico is able to analyze how a compound will affect cells and what drugs can be used to treat the cells in addition to possible side effects

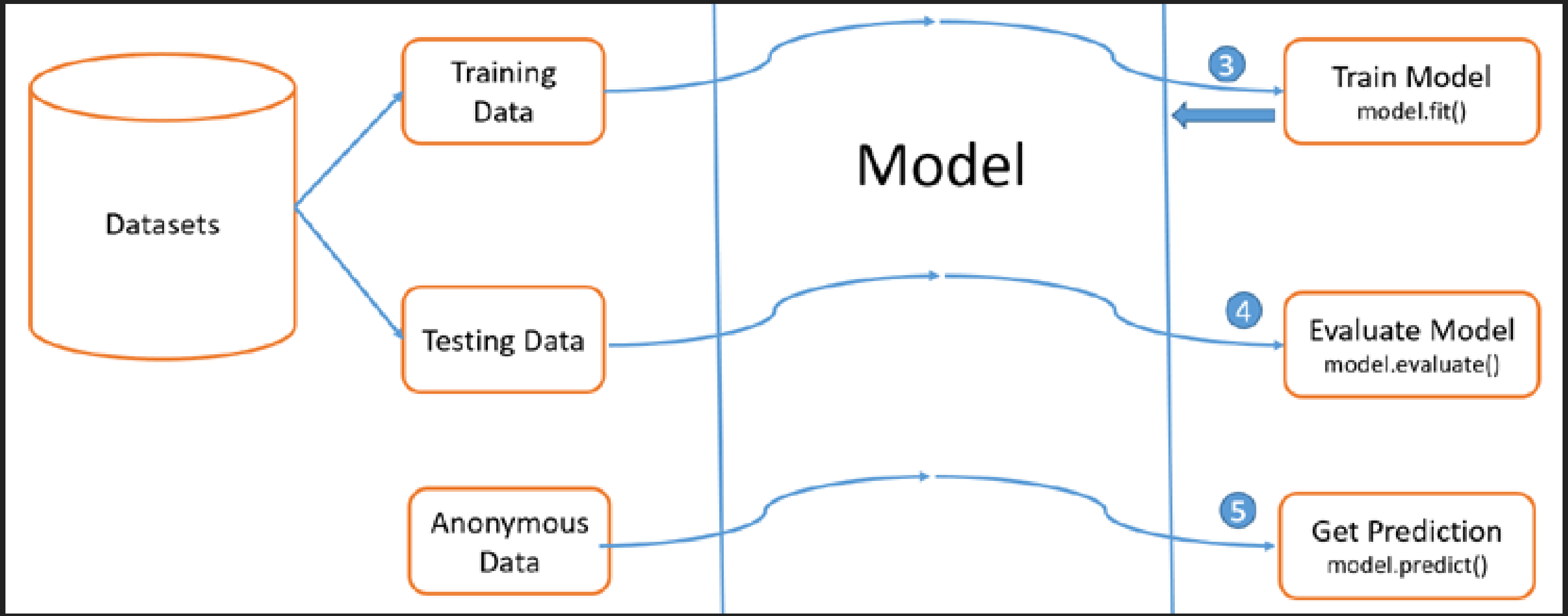
A comprehensive drug discovery engine, which utilizes **millions of samples and multiple data types** to discover signatures of disease and identify the most promising targets for billions of molecules that already exist or can be generated de novo with the desired set of parameters.

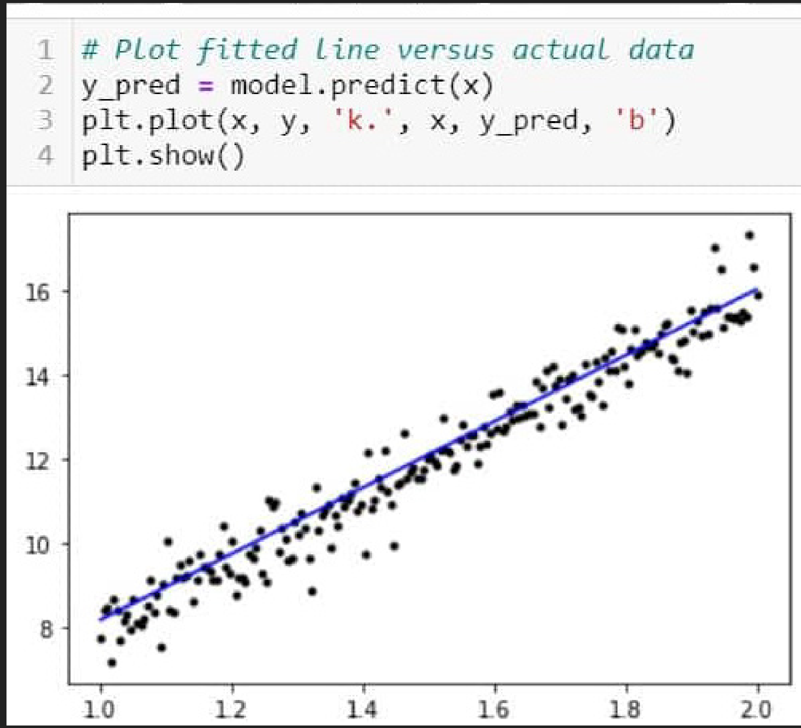


Machine Learning Model Lifecycle - Example

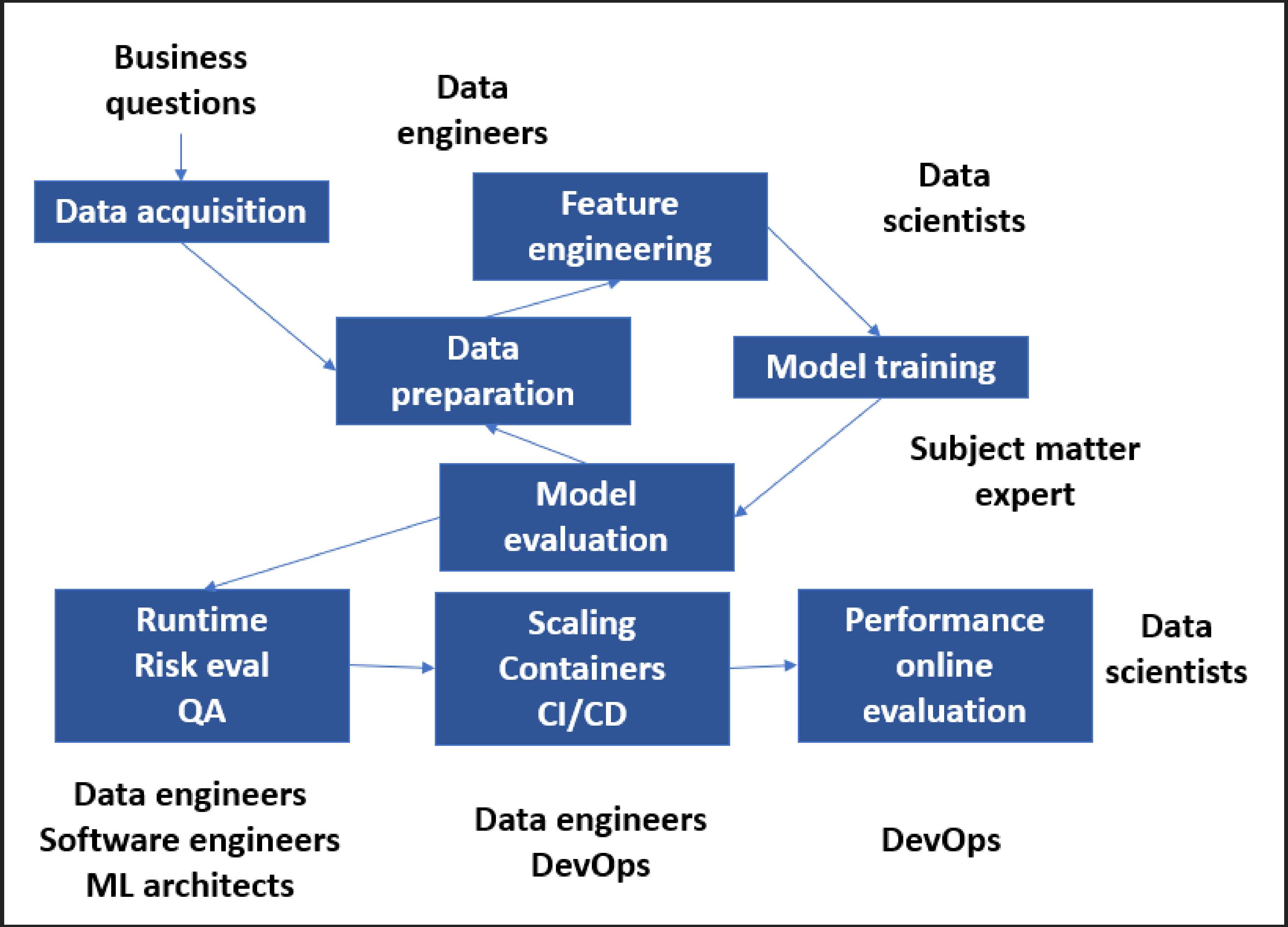
1. **Define** the model: using the Sequential or Model class and add the layers
2. **Compile** the model: call compile method and specify the loss, optimizer and metrics
3. **Train** the model: call fit method and use training data
4. **Evaluate** the model: call evaluate method and use testing data to evaluate trained model
5. Get predictions: use predict method on **new data for predictions**







Digikey, techbrij



Business questions
Data engineers
Data acquisition
Feature engineering
Data scientists
Data preparation
Model training
Subject matter expert
Model evaluation
Runtime
Risk eval
QA
Scaling
Containers
CI/CD
Performance
online
evaluation
Data scientists
Data engineers
Software engineers
ML architects
Data engineers
DevOps
DevOps

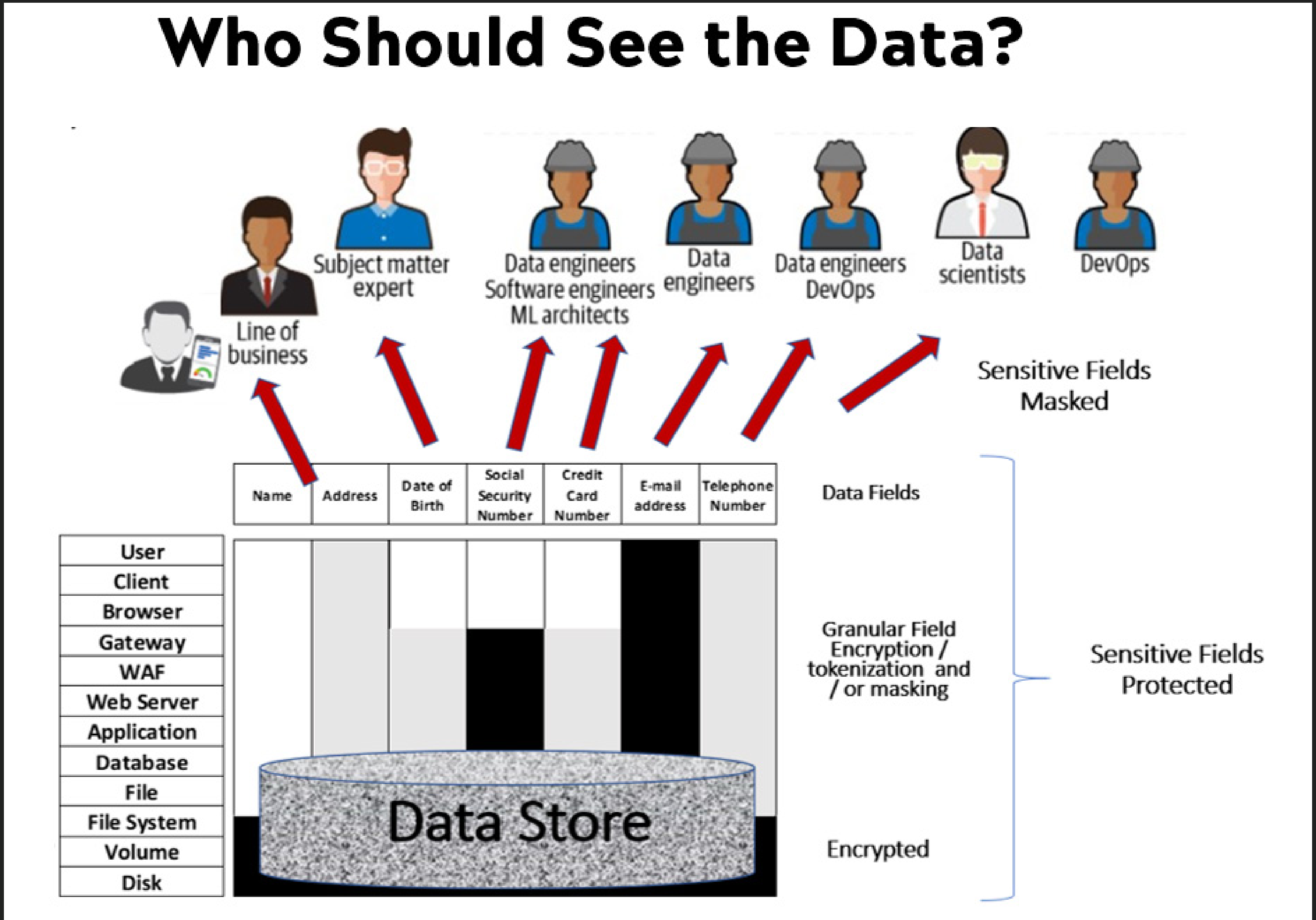
PROTEGRITY

Specify Access Control and Data Protection to Use

Review Use Cases and Types of Data

Implement

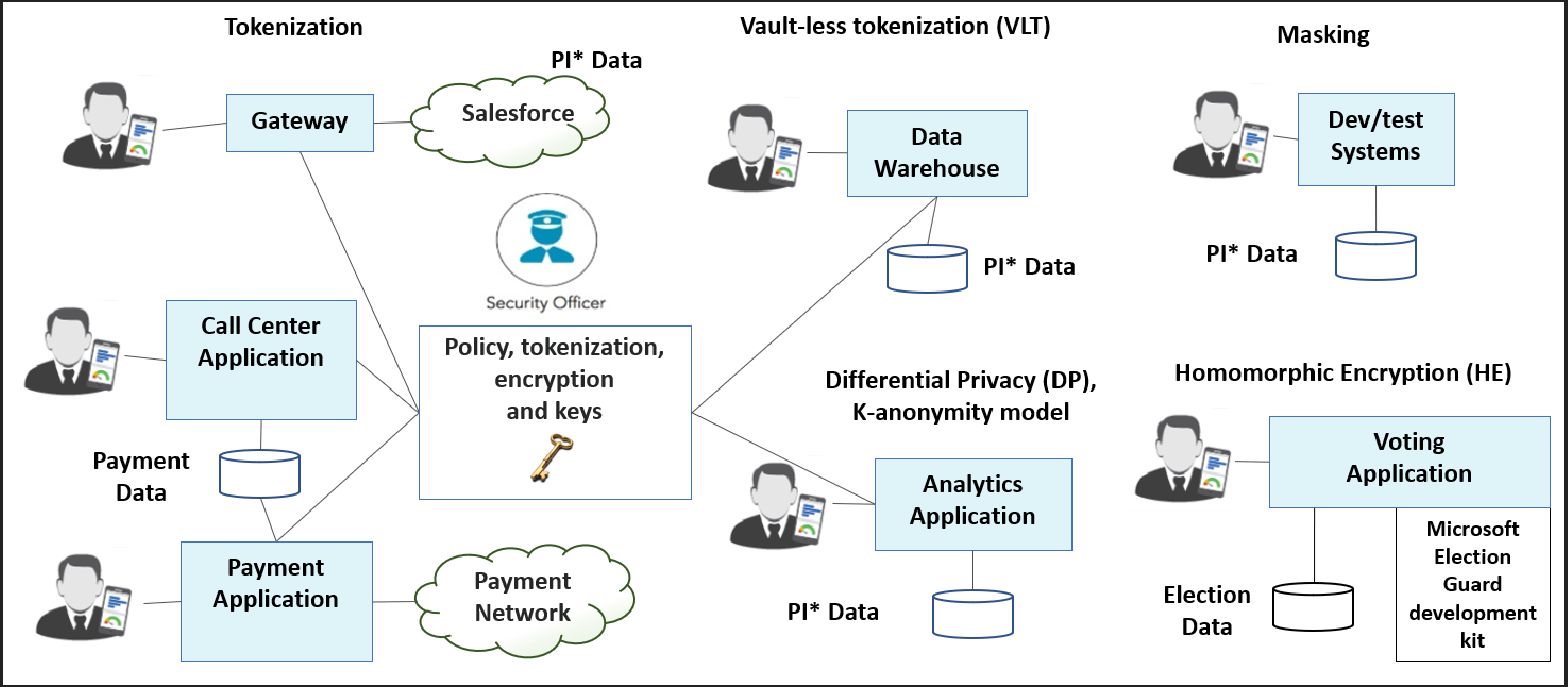
1. Dynamic Masking
2. Tokenization
3. Encryption



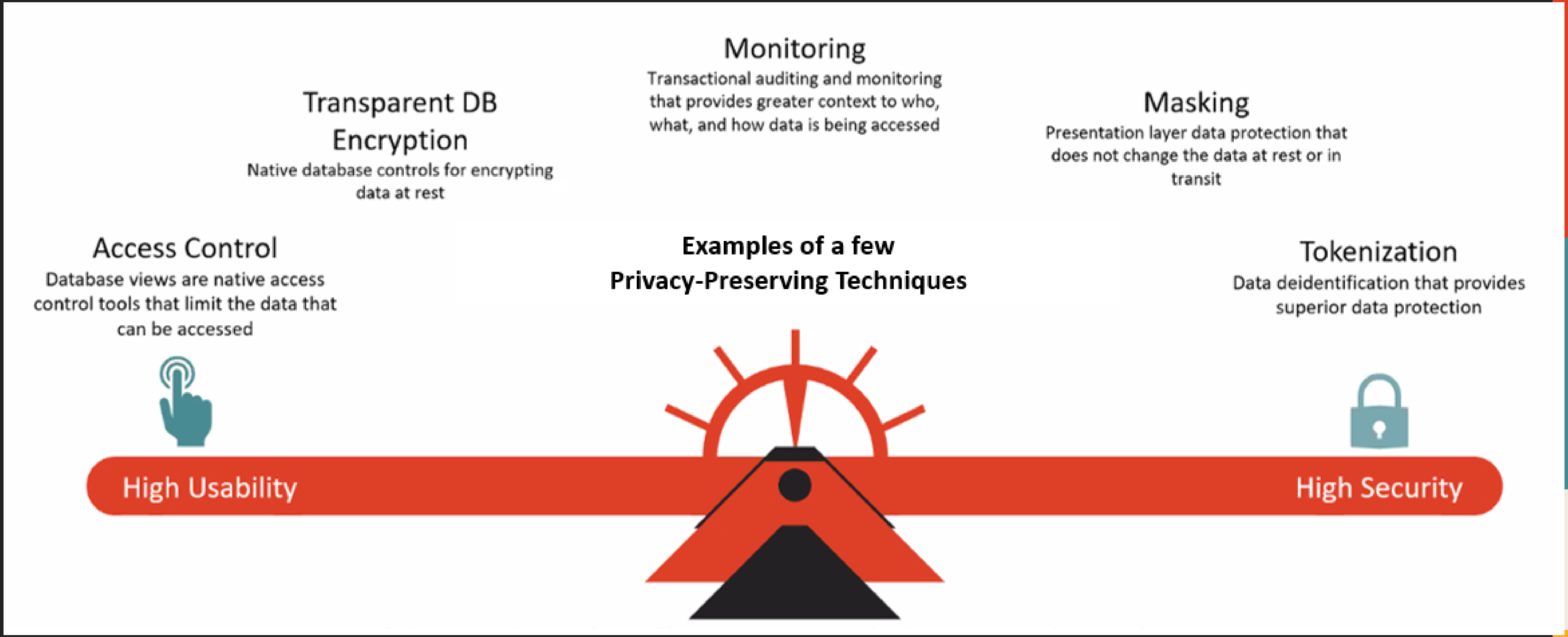
PROTEGRITY

Data Protection Techniques

Use Cases

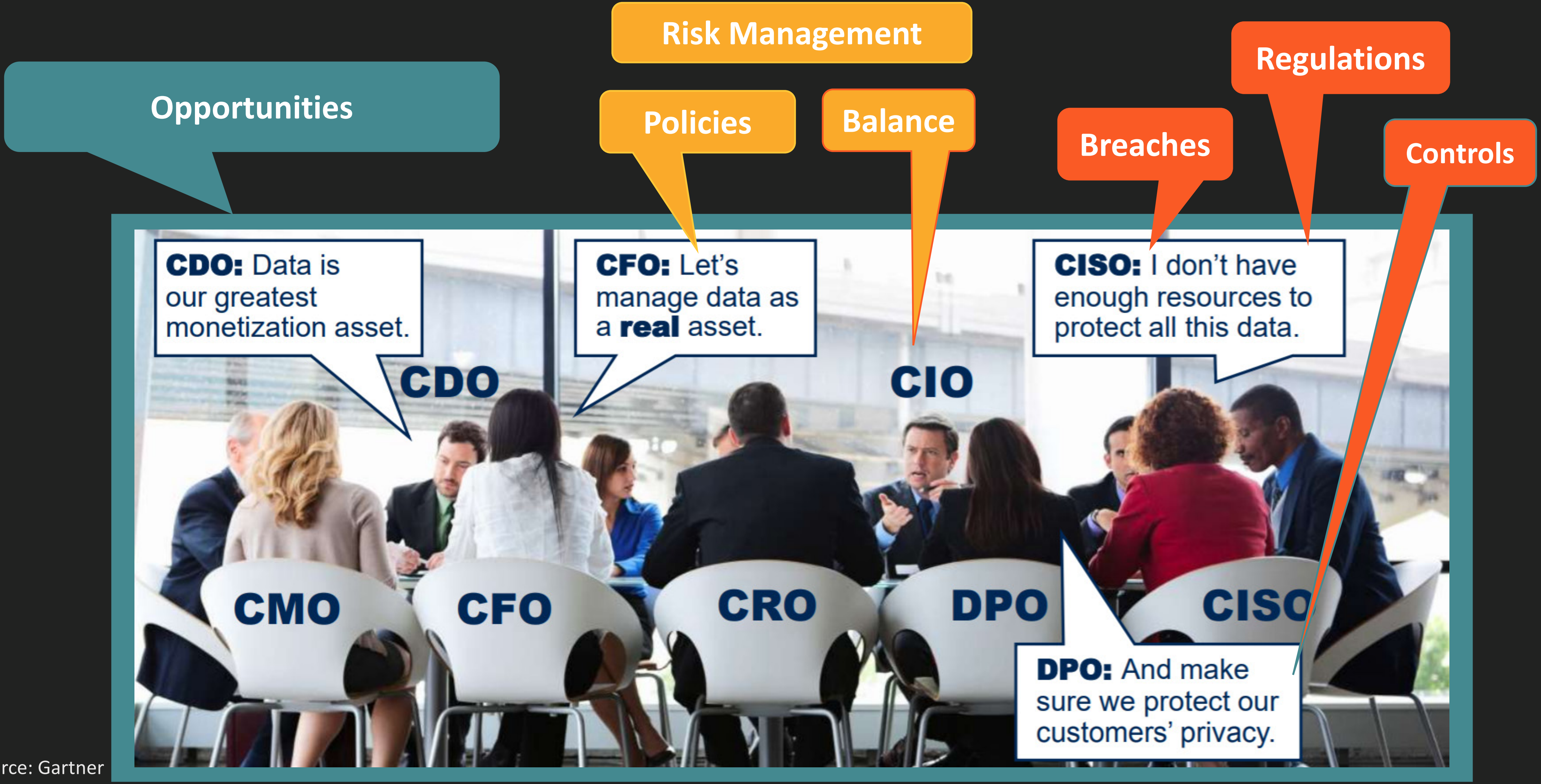

Tokenization
PI* Data
Gateway
Salesforce
Security Officer
Call Center Application
Policy, tokenization, encryption and keys
Payment Data
Payment Application
Payment Network
Vault-less tokenization (VLT)
Data Warehouse
PI* Data
Differential Privacy (DP), K-anonymity model
Analytics Application
PI* Data
Masking
Dev/test Systems
PI* Data
Homomorphic Encryption (HE)
Voting Application
Microsoft Election Guard development kit
Election Data

What Data Protection Technique do I need?



Unlock the Potential of Data Security
- Data Security Governance Stakeholders

Source: Gartner



Source: Gartner

PROTEGRITY

Protect data in ways that are transparent to business processes and compliant to regulations

What are the Drivers for these People?

Opportunities

Risk Management

Policies

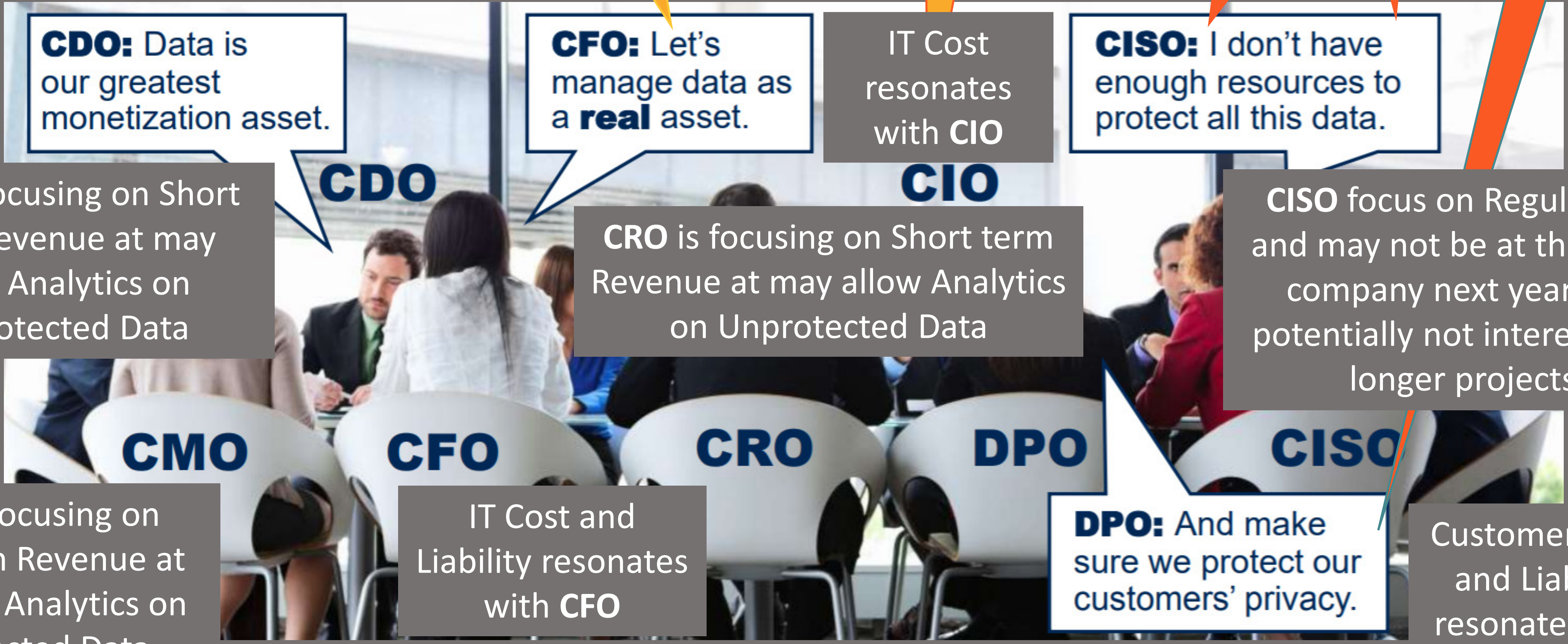
Balance

Regulations

Breaches

Controls

CEO is focusing on Short term Revenue



IT Cost resonates with **CIO**

CDO is focusing on Short term Revenue at may allow Analytics on Unprotected Data

CRO is focusing on Short term Revenue at may allow Analytics on Unprotected Data

CISO focus on Regulations and may not be at the same company next year and potentially not interested in longer projects

CMO is focusing on Short term Revenue at may allow Analytics on Unprotected Data

IT Cost and Liability resonates with **CFO**

Customer Trust and Liability resonates with **DPO**

Source: Adapted from Gartner

PROTEGRITY

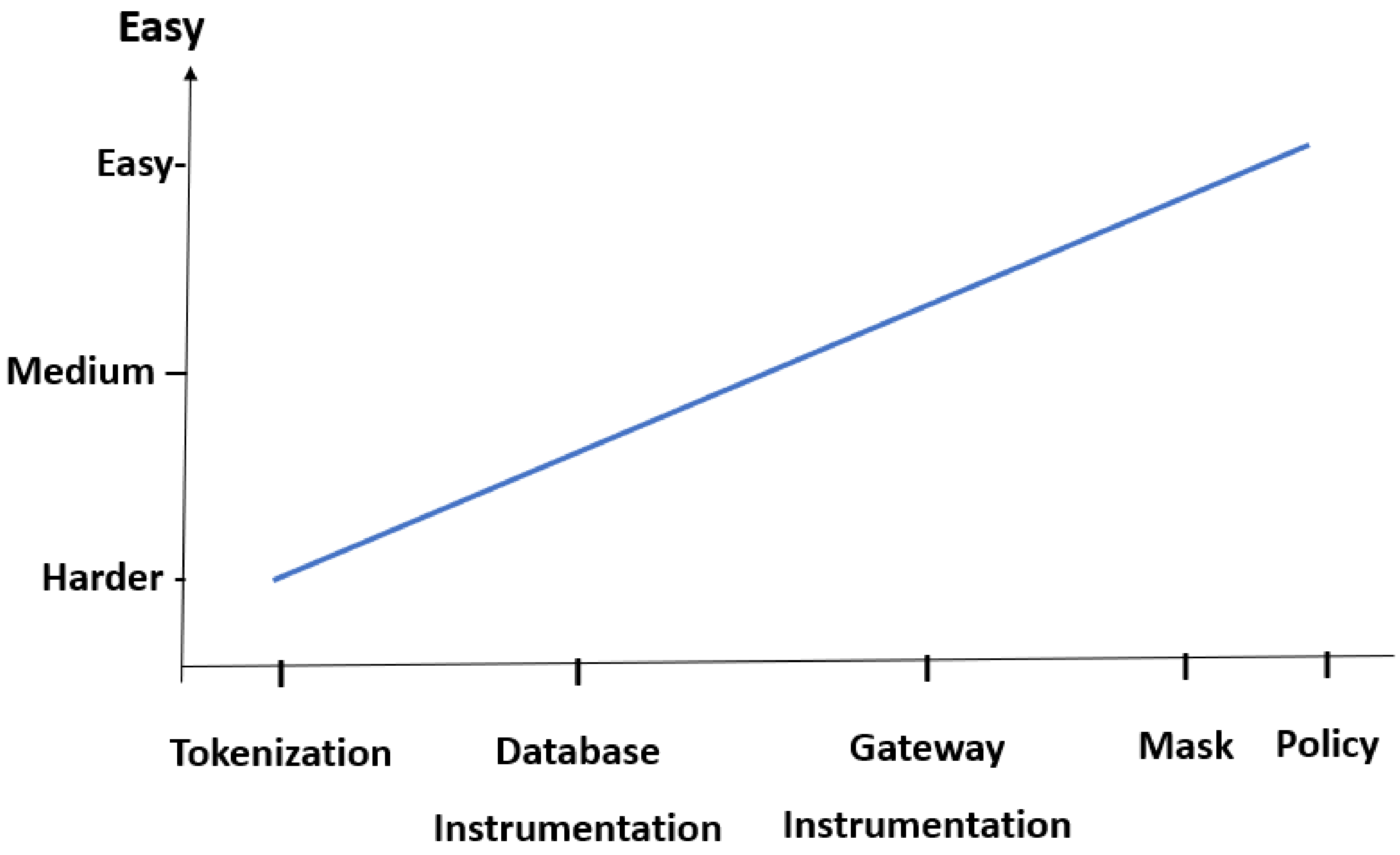
What is the Cost of Implementing different Data Protection?

Implementation Effort

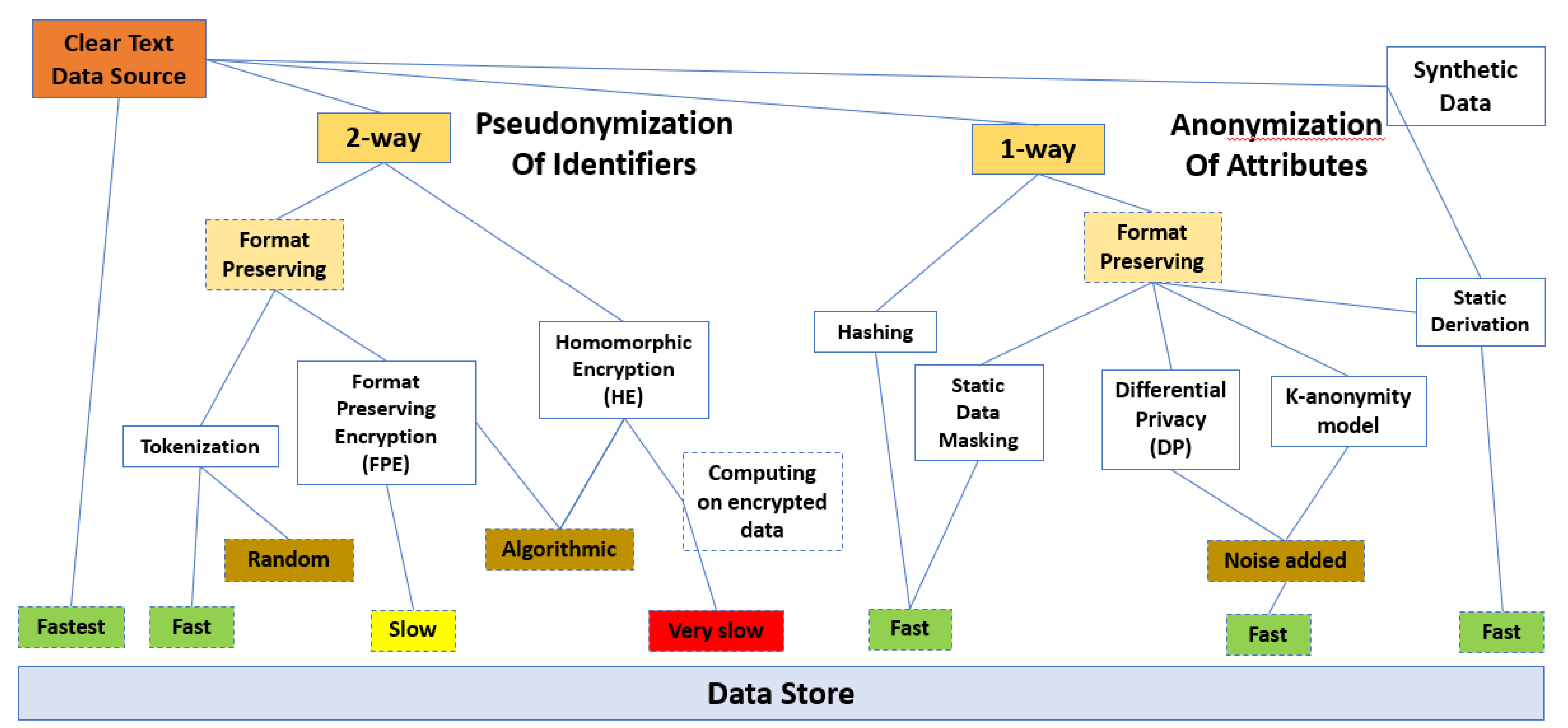
What you need to do	Effort
Create Data Element	⌛
Create Roles	⌛
Create Policy	⌛
Create Masks	⌛
Instrument DSG	⌛ ⌛
Instrument Databases	⌛ ⌛
Tokenize Data at Rest	⌛ ⌛ ⌛

⌛ Some Effort Required

⌛ No Effort Required



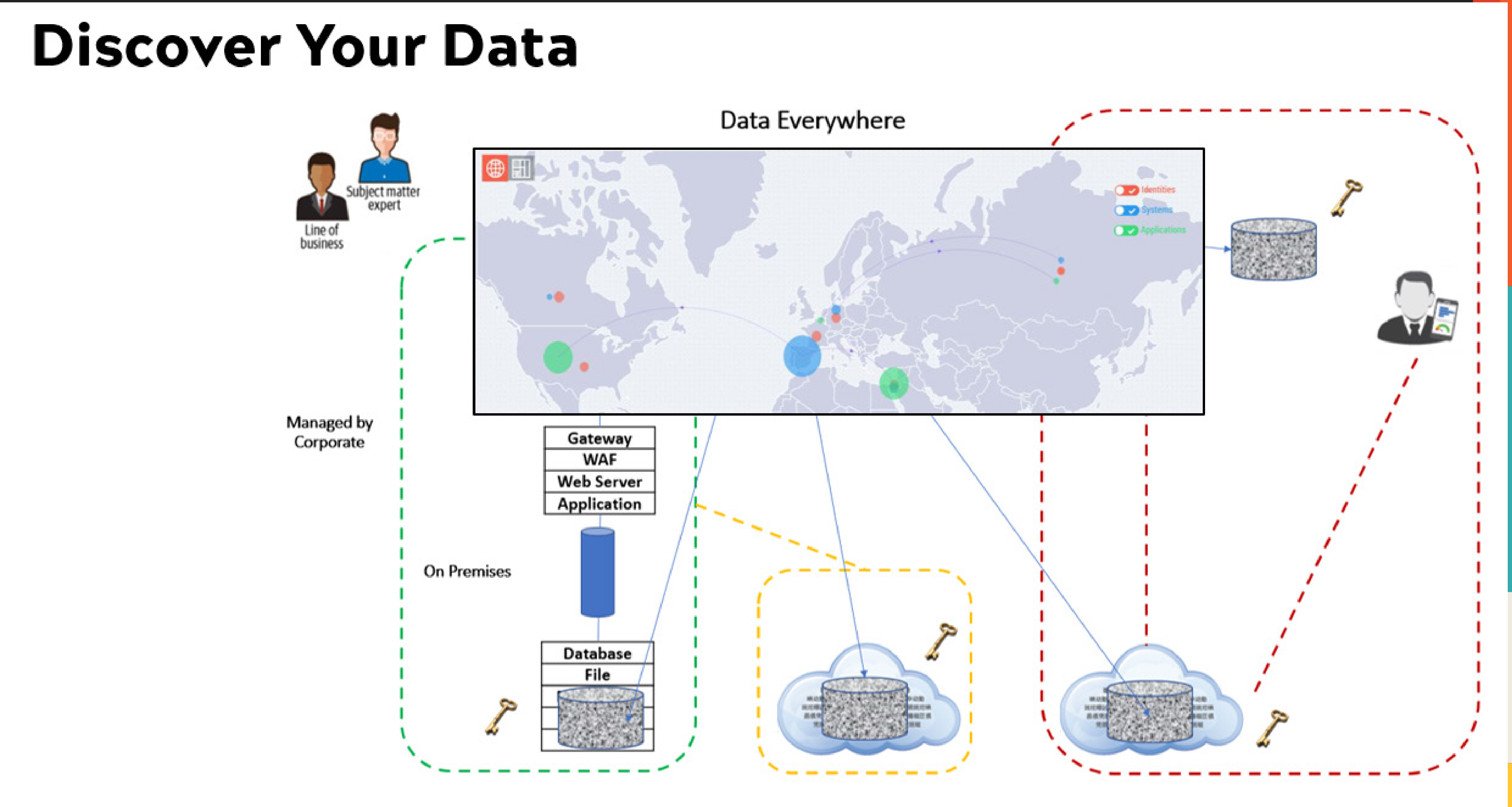
How are Data Protection Techniques different?



What Data is Sensitive?

What Data is Regulated?

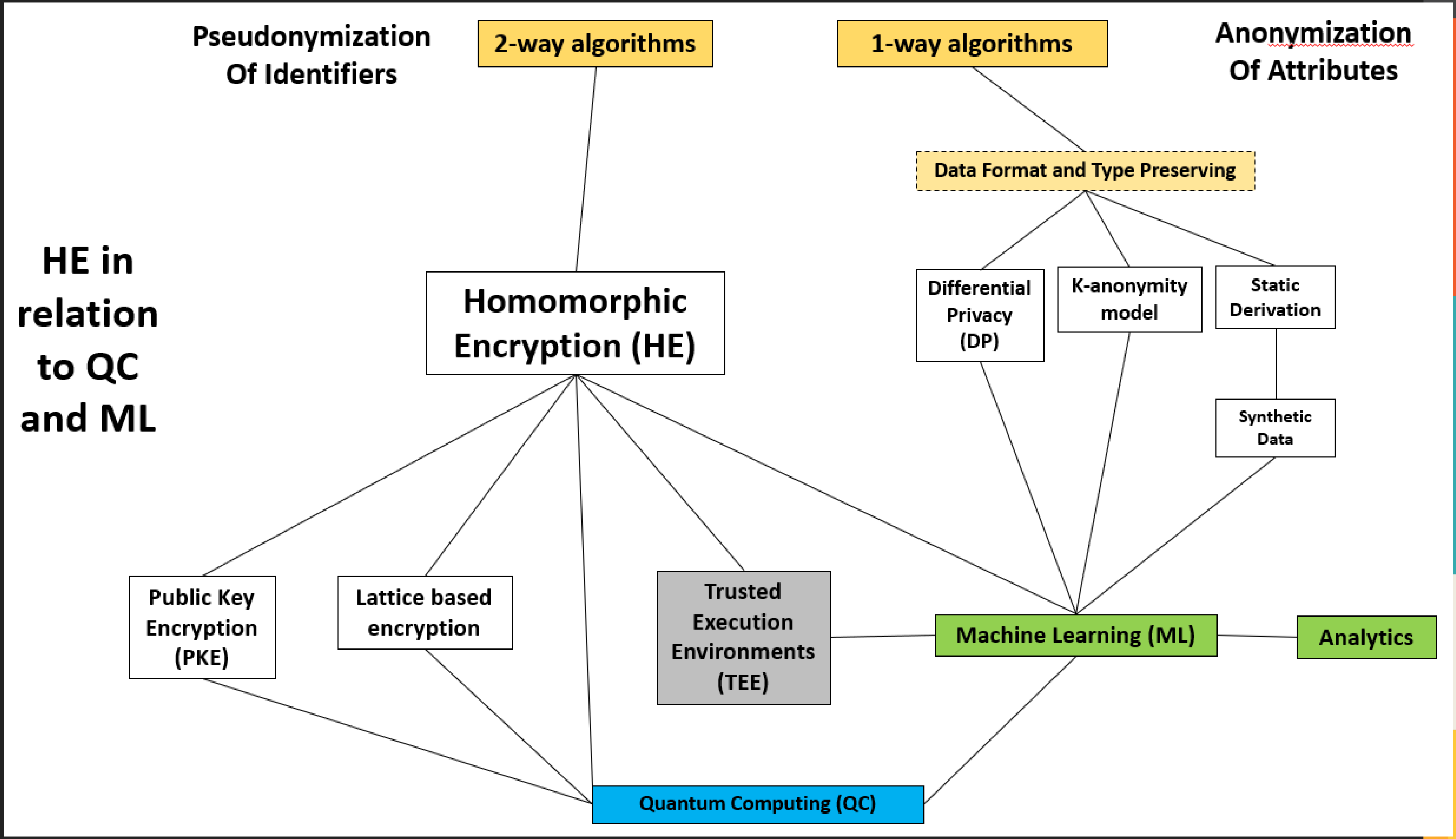
- EU GDPR
- US California CCPA / CPRA
- PCI DSS
- US HIPAA



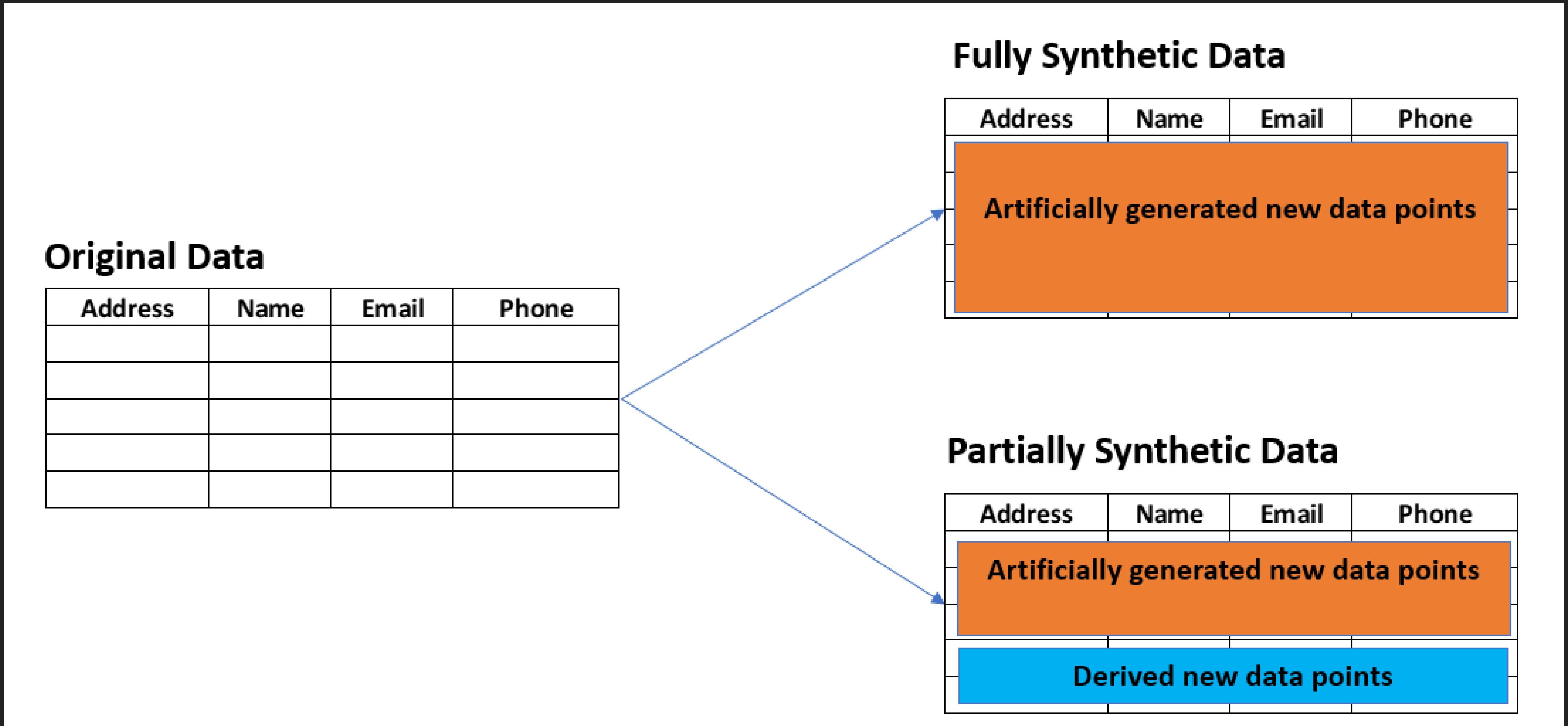
Use Cases involving Analytics

Are my Use Cases involving Multi-party Computing and Homomorphic Encryption?

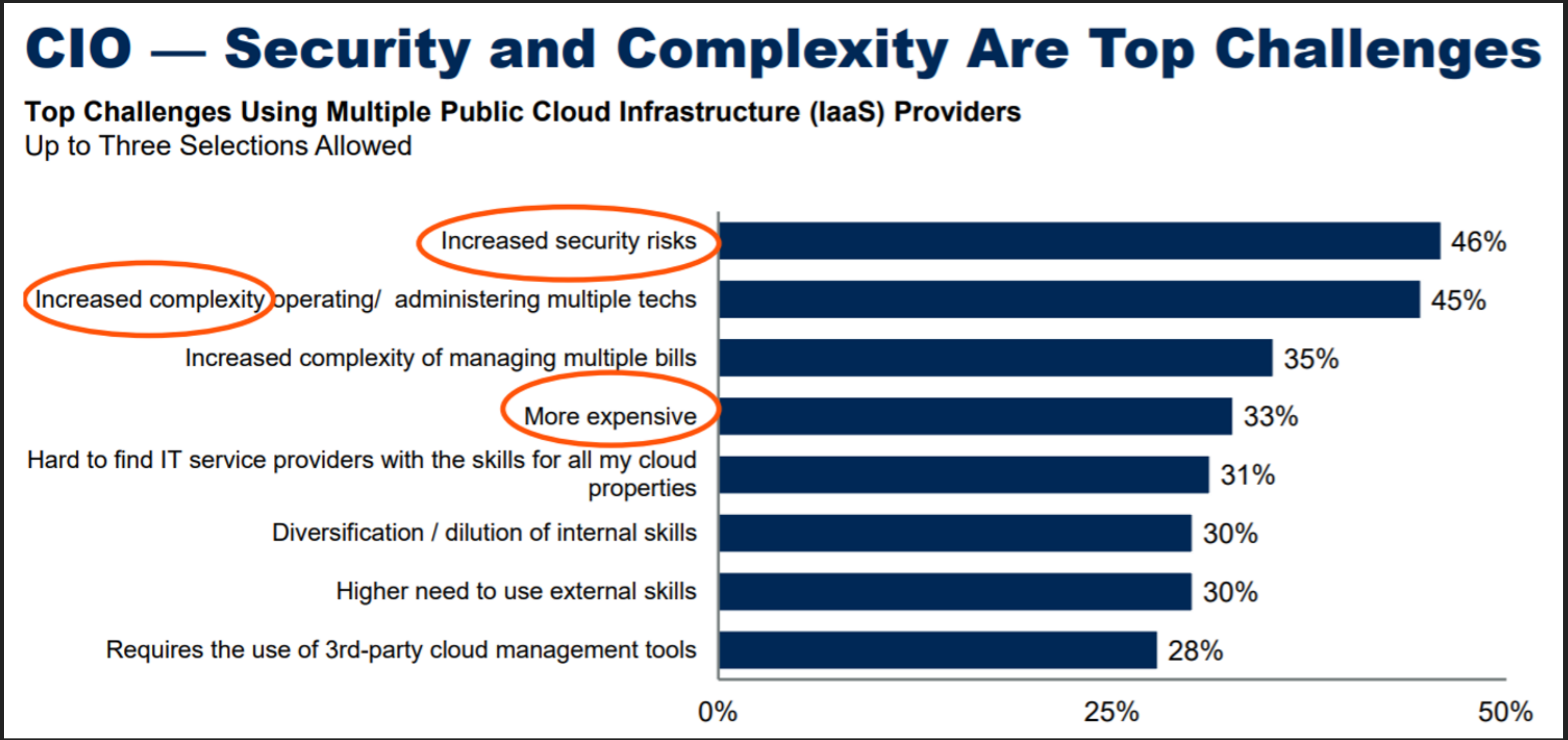
Do I need to plan for Quantum Computing?



Secure AI – Use Case with Synthetic Data



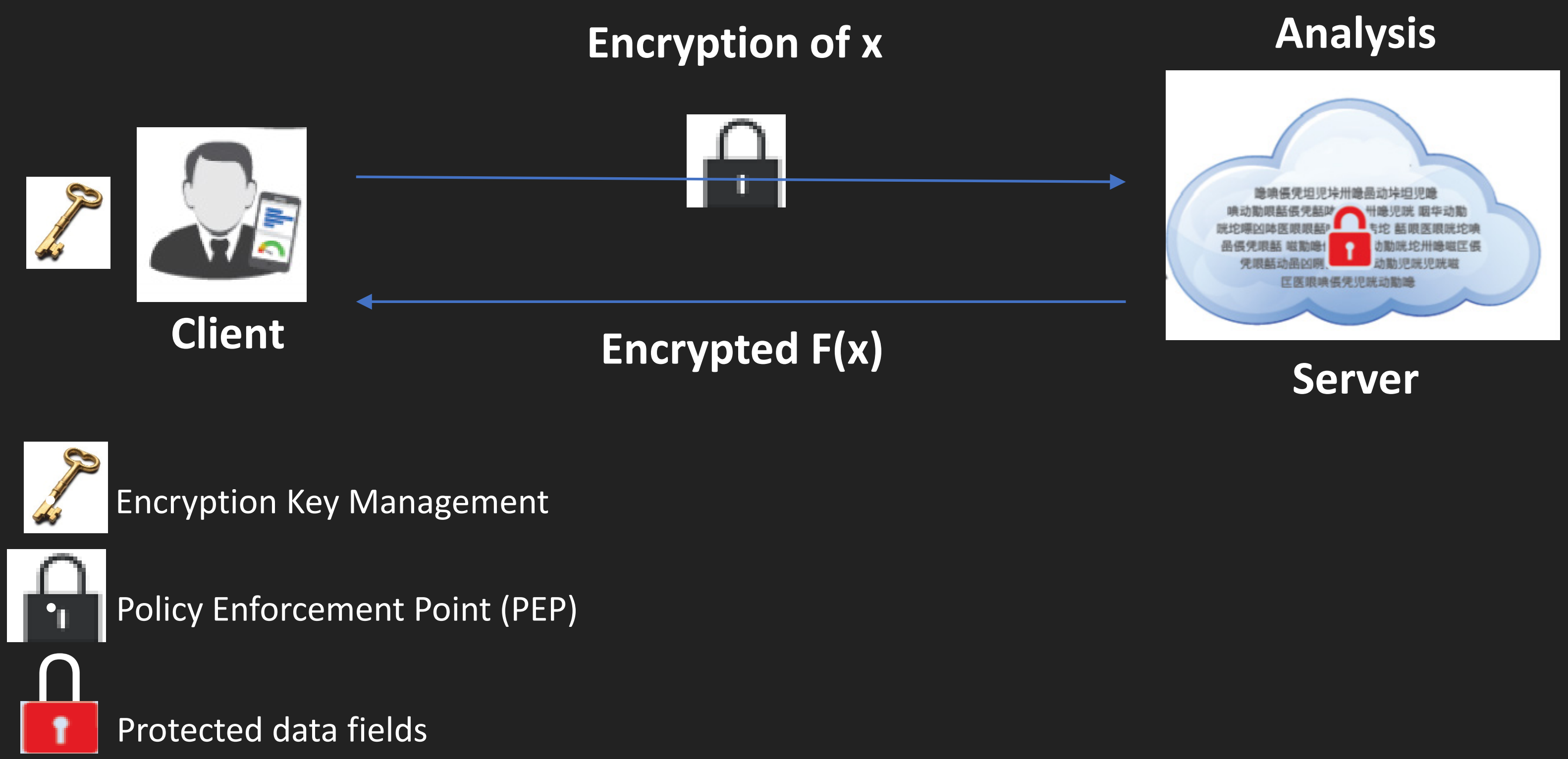
Public Cloud (Iaas) Risk, Complexity & Cost



Homomorphic encryption (HE)

HE depicted in a client-server model

- The client sends encrypted data to a server, where a **specific analysis is performed on the encrypted** data, without decrypting that data.
- The encrypted result is then sent to **the client, who can decrypt it to obtain the result** of the analysis they wished to outsource.



https://royalsociety.org

Use Cases for Secure Multi Party Computation & Homomorphic Encryption (HE)

Business models and application domains:

Domain	Genomics	Health	National Security	Education	Social Security	Business Analytics	Cloud
Sample Topics	*GWAS*	*billing and reporting*	*smart grid*	*school dropouts*	*credit history*	*prediction*	*storage, sharing*
Data Owner	*medical institutions*	*clinics and hospitals*	*nodes and network*	*schools, welfare*	*government*	*business owners*	*clients*
Why HE?	*HIPAA*	*cyber insurance*	*privacy*	*FERPA*	*cyber crimes*	*data are valuable*	*untrusted server*
Who pays?	*health insurance*	*hospital*	*energy company*	*DoE*	*government*	*business owners*	*clients*

http://homomorphicencryption.org

Use case – Retail - Data for Secondary Purposes

Large **aggregator of credit card transaction data.**

Open a new revenue stream

- Using its data with its **business partners**: **retailers**, **banks** and **advertising** companies.
- They could help their partners achieve **better ad conversion rate**, improved **customer satisfaction**, and more **timely offerings**.
- Needed to respect user privacy and specific **regulations**. In this specific case, they wanted to work with a retailer.
- Allow the retailer to **gain insights** while **protecting user privacy**, and the credit card organization's IP.
- An **analyst** at **each** organization's **office** first used the software to **link the data without exchanging any of the underlying data**.

Data used to **train the machine learning and statistical models**.

- A logistic and **linear regression model** was trained using **secure multi-party computation** (SMC).
- In the simplest form SMC **splits a dataset into secret shares** and enables you to train a model without needing to put together the pieces.
- The information that is communicated between the peers is **encrypted at all times** and cannot be reverse engineered.

With the augmented dataset, the retailer was able to get a **better picture of its customers buying habits**.

PROTEGRITY

Use case - Financial services industry

Confidential financial datasets which are vital for gaining significant **insights**.

- The use of this data requires navigating a minefield of **private client information** as well as **sharing data** between independent financial institutions, to create a **statistically significant dataset**.

- Data privacy regulations such as **CCPA, GDPR and other emerging regulations** around the world
- **Data residency** controls as well as enable data **sharing in a secure and private** fashion.

Reduce and remove the legal, **risk and compliance** processes

- Collaboration **across divisions**, **other organizations** and **across jurisdictions** where data **cannot be relocated or shared**
- Generating **privacy** respectful datasets with **higher analytical value** for Data Science and Analytics applications.

Use case: Bank - Internal Data Usage by Other Units

A large bank wanted to broaden access to its **data lake** without compromising data **privacy**, preserving the data's **analytical value**, and at reasonable infrastructure costs.

- Current approaches to **de-identify data** did **not** fulfill the **compliance** requirements and business needs, which had led to several bank **projects being stopped**.
- The issue with these techniques, like **masking, tokenization, and aggregation, was that they did not sufficiently protect** the data **without overly degrading data quality**.

This approach allows creating privacy protected datasets that retain their analytical value for Data Science and business applications.

A plug-in to the organization's analytical pipeline **to enforce the compliance policies before the data was consumed by data science** and business teams from the data lake.

- The analytical quality of the data was preserved for machine learning purposes by-using AI and leveraging privacy models like **differential privacy and k-anonymity**.

Improved data access for teams **increased the business' bottom line** without adding excessive infrastructure costs, while **reducing the risk** of-consumer information exposure.

Trusted execution environments

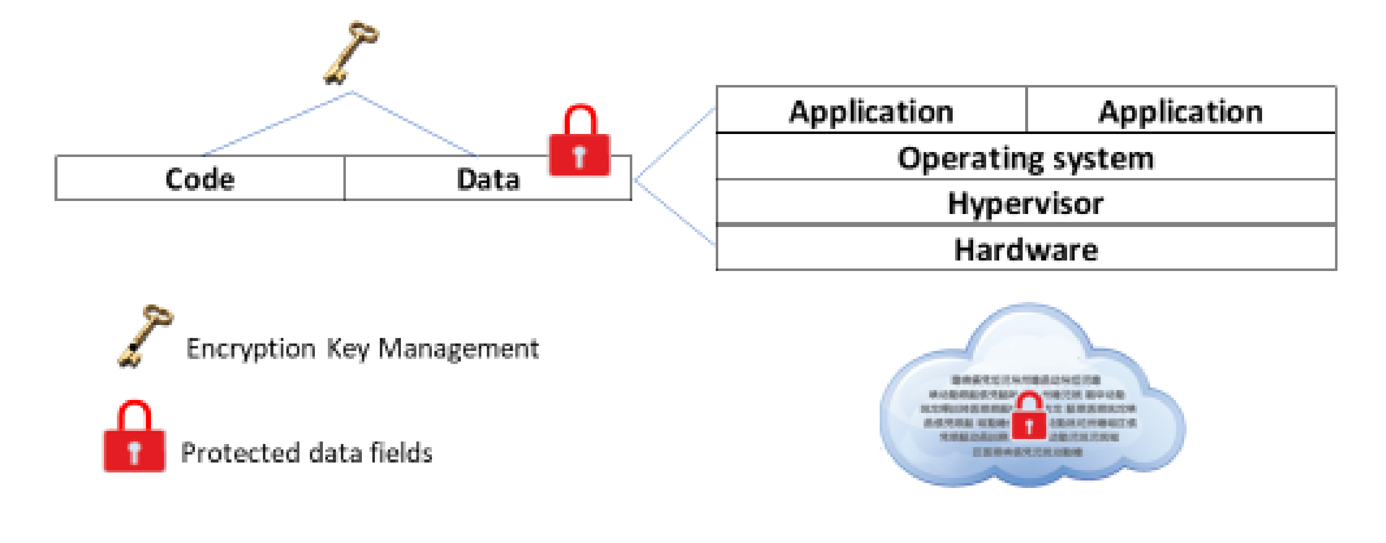
Trusted Execution Environments (TEEs) provide secure computation capability through a combination of **special-purpose hardware** in modern processors and software built to use those hardware features.

The special-purpose hardware provides a mechanism by which a process can run on a processor **without its memory or execution state being visible to any other process** on the processor,

- not even the operating system or other privileged code.

Computation in a TEE is **not** performed on data while it remains **encrypted**.

- Typically, the **memory** space of each TEE (enclave) application is **protected from access**
  - **AES-encrypted when and if it is stored off-chip**.

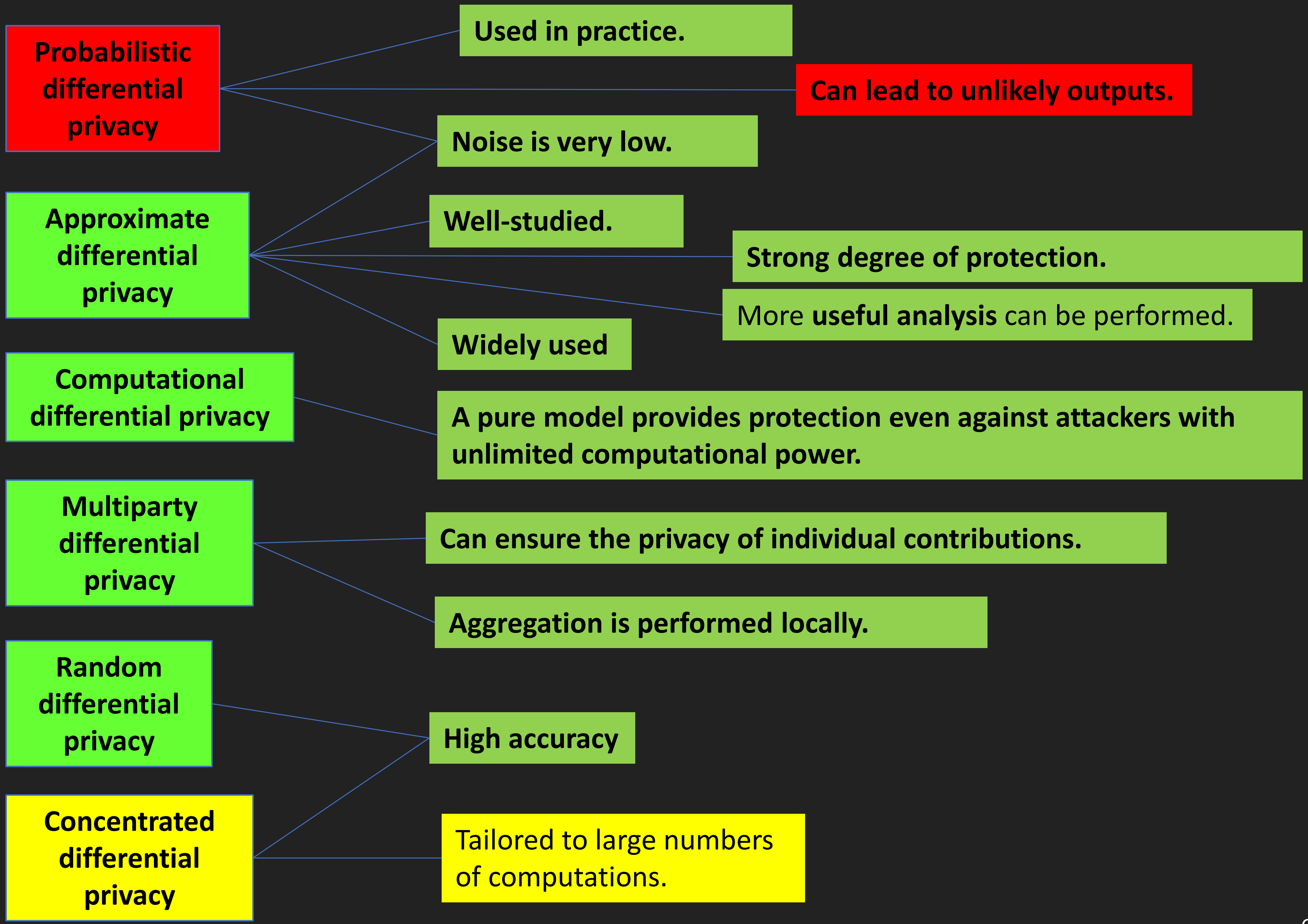


Usability is low and products/services are emerging in MS Azure, IBM's cloud service Amazon AWS (late 2020)*

*: Source: http://publications.officialstatistics.org

6 Differential Privacy Models

In differential privacy, the concern is about privacy as the relative **difference in the result whether a specific individual or entity is included** in the input or excluded



PROTEGRITY

k-Anonymity

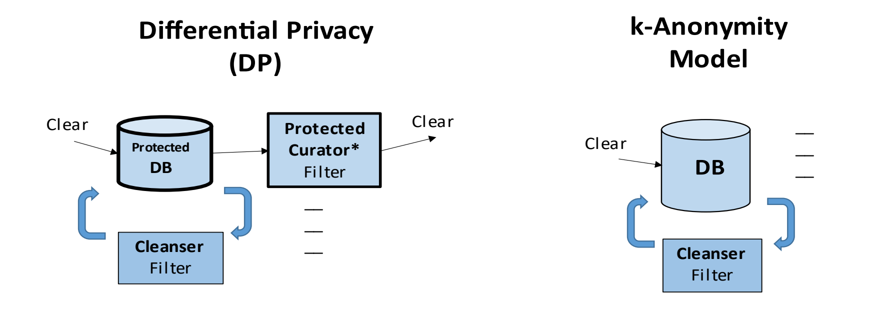
Name	Postcode	Age	Gender	Disease
Patrick	SW1 4YB	22	Male	Cardiovascular
Sebastian	SW1 4ZE	23	Male	Respiratory
Reece	SW1 2HY	18	Male	No Illness
Tilly	NW10 8FN	47	Female	Cancer
Abby	NW10 4AB	42	Female	No Illness
Elise	NW10 0FW	56	Female	Cardiovascular
Morgan	E17 9QY	23	Male	Respiratory
George	E17 3SF	29	Male	Liver
Sienna	E17 5WD	18	Female	Cancer

Name	Postcode	Age	Gender	Disease
*	SW1 *	22	Male	Cardiovascular
*	SW1 *	23	Male	Respiratory
*	SW1 *	18	Male	No Illness
*	NW10 *	47	Female	Cancer
*	NW10 *	42	Female	No Illness
*	NW10 *	56	Female	Cardiovascular
*	E17 *	23	*	Respiratory
*	E17 *	29	*	Liver
*	E17 *	18	*	Cancer

For k-anonymity to be achieved, there need to be at least k individuals in the dataset who share the set of attributes that might become identifying for each individual.

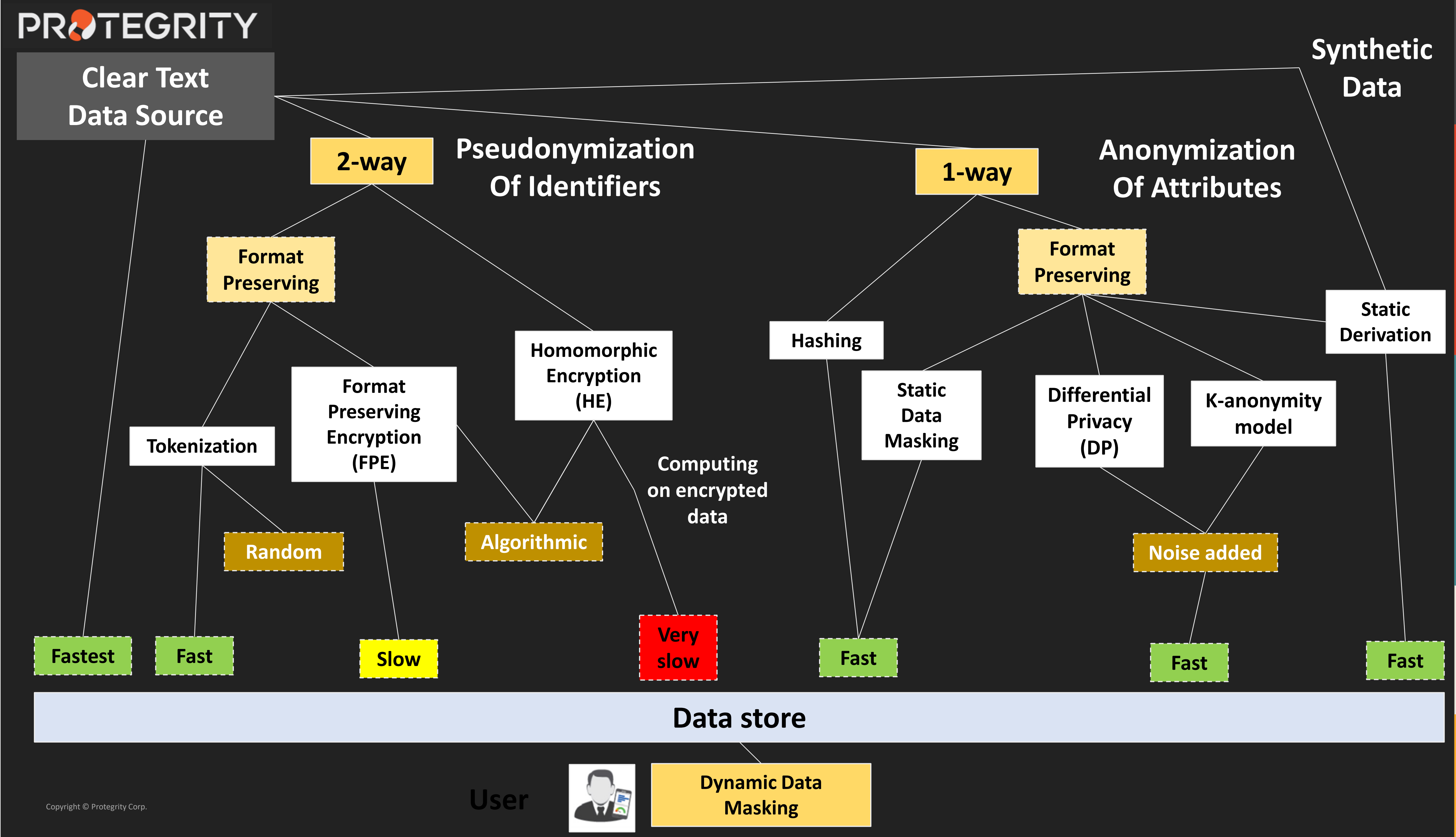
K-anonymity might be described as a ‘hiding in the crowd’ guarantee: if each individual is part of a larger group, then any of the records in this group could correspond to a single person.

This second table shows the data anonymised to achieve k-anonymity of k = 3, as you can see this was achieved by generalising some quasi-identifier attributes and redacting some others.

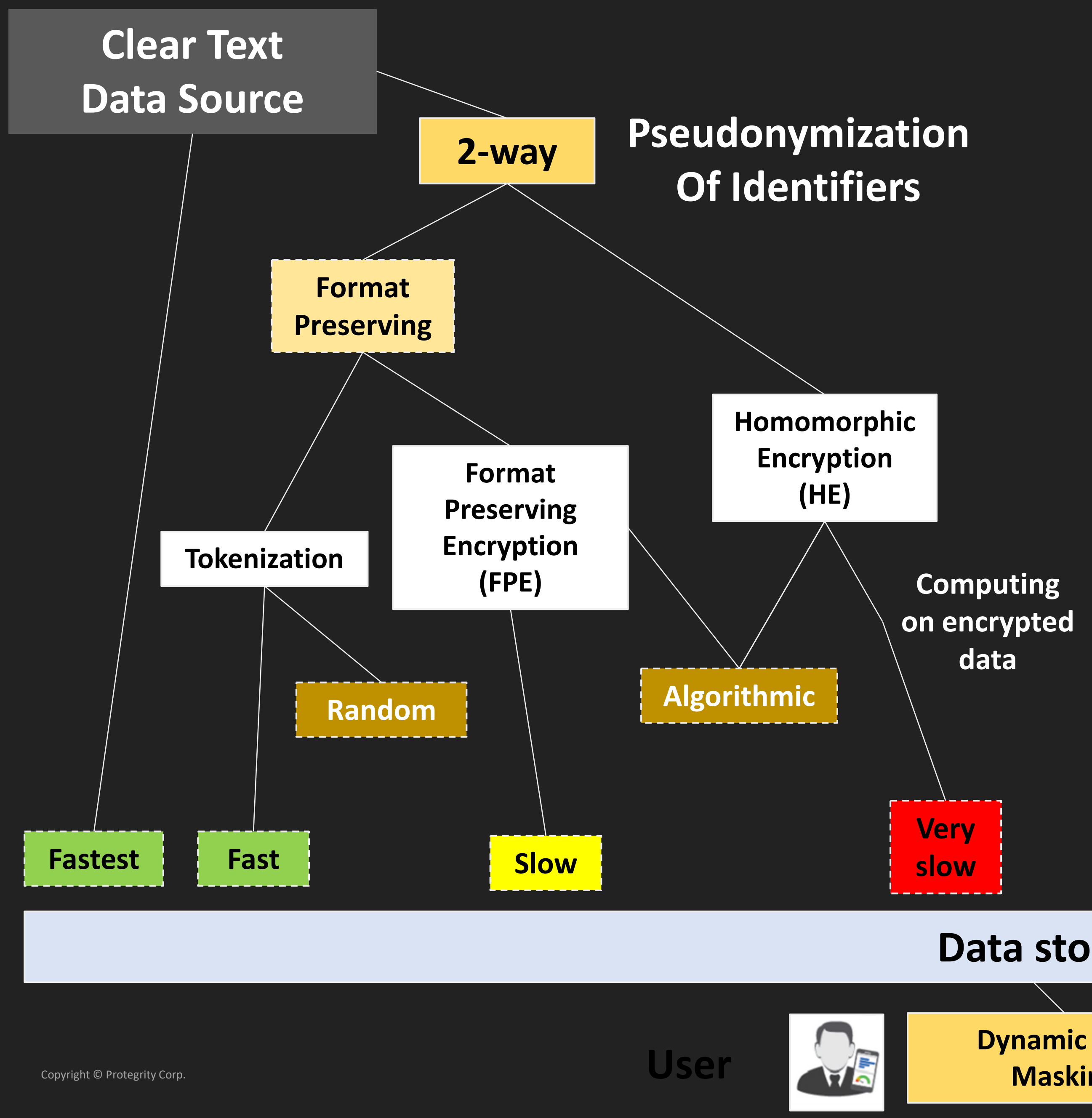


Data protection techniques: Deployment on-premises, and clouds

Privacy enhancing data de-identification terminology and classification of techniques			Data Warehouse	Centralized	Distributed	On-premises	Public Cloud	Private Cloud
De-identification techniques	Tokenization	Vault-based tokenization		y				y
		Vault-less tokenization	y	y	y	y	y	y
	Cryptographic tools	Format preserving encryption		y	y	y	y	y
		Homomorphic encryption			y		y	
	Suppression techniques	Masking	y	y	y	y	y	y
		Hashing	y	y	y	y	y	y
Formal privacy measurement models	Differential Privacy	Server model	y	y	y	y	y	y
		Local model	y	y	y	y	y	y
	K-anonymity model	L-diversity	y	y	y	y	y	y
		T-closeness	y	y	y	y	y	y


PROTEGRITY
Clear Text Data Source
Synthetic Data
2-way
Pseudonymization Of Identifiers
1-way
Anonymization Of Attributes
Format Preserving
Format Preserving
Static Derivation
Hashing
Homomorphic Encryption (HE)
Format Preserving Encryption (FPE)
Tokenization
Static Data Masking
Differential Privacy (DP)
K-anonymity model
Computing on encrypted data
Random
Algorithmic
Noise added
Fastest
Fast
Slow
Very slow
Fast
Fast
Fast
Data store
User
Dynamic Data Masking

PROTEGRITY



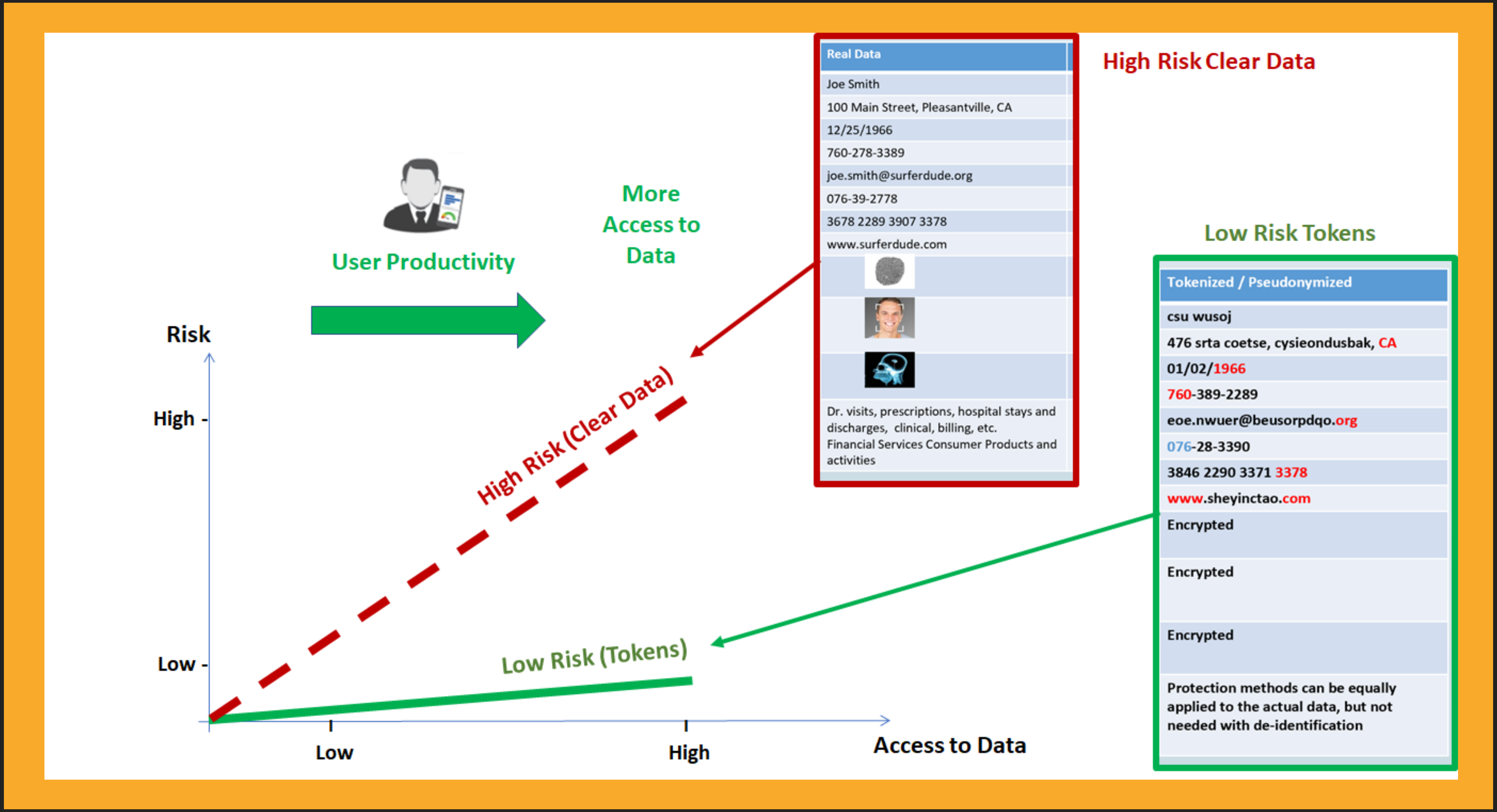
Quantum Computers?

- **Quantum** computers and other strong computers can **break algorithms** and patterns in encrypted data.
- **We can instead use random** numbers to secure sensitive data.
- **Random** numbers are **not** based on an algorithm or pattern that computers can **break.**

Tech giants are building their own machines and speeding to make them available to the world as a cloud computing service. In the competition: IBM, Google, Microsoft, Intel, Amazon, IonQ, Quantum Circuits, Rigetti Computing

PROTEGRITY

Lower Risk and Higher Productivity with More Access to More Data



Non-reversable Data Transformations

Example of Data Generalization

Source data:

Patient	Age	Gender	Region	Disease
173965429	57	Female	Hamburg	Gastric ulcer

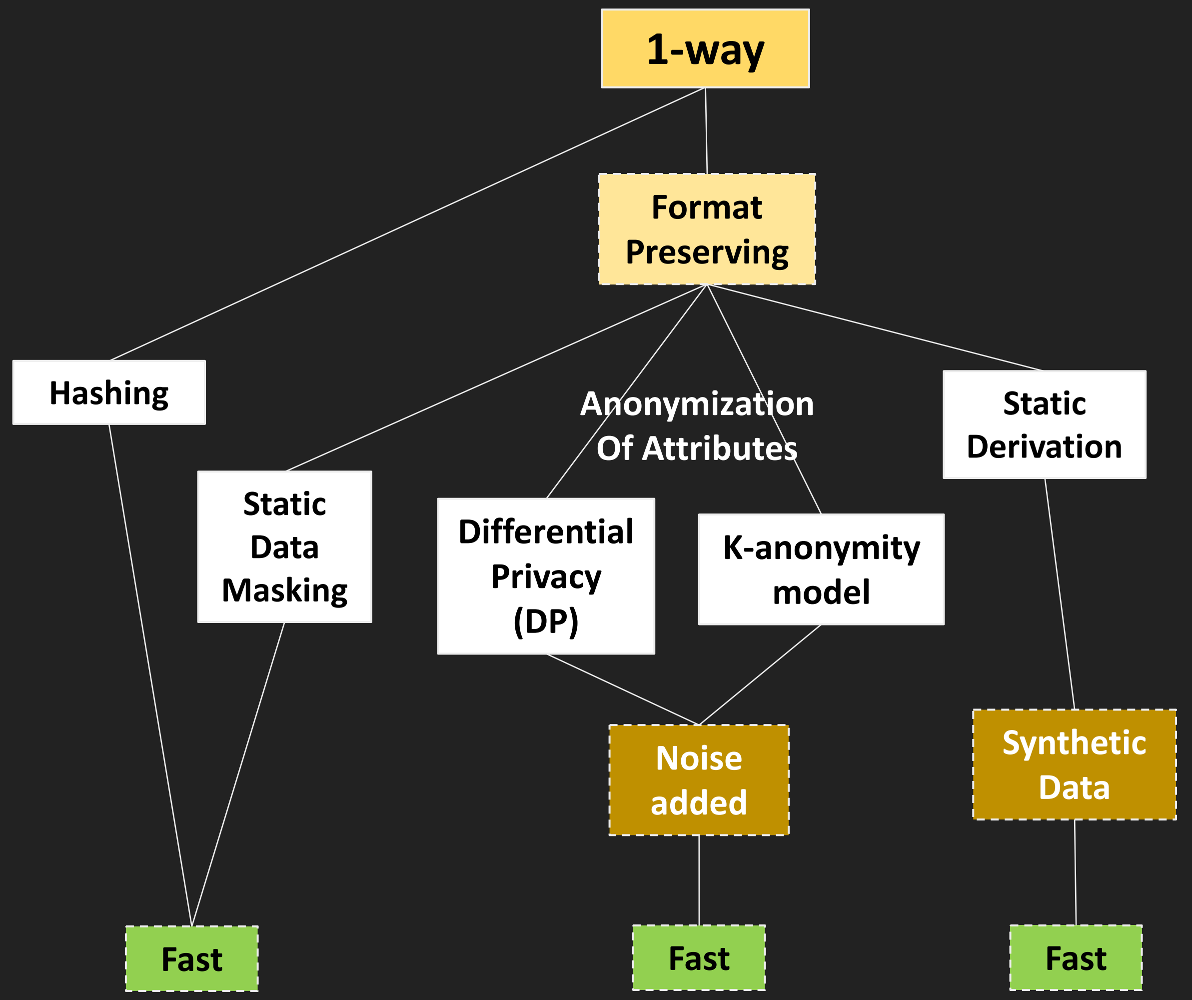
Generalization

Generalization

Output data:

Patient	Age	Gender	Region	Disease
173965429	>50	Female	Germany	Gastric ulcer

Clear Text Data Source



Data store

User

Dynamic Data Masking

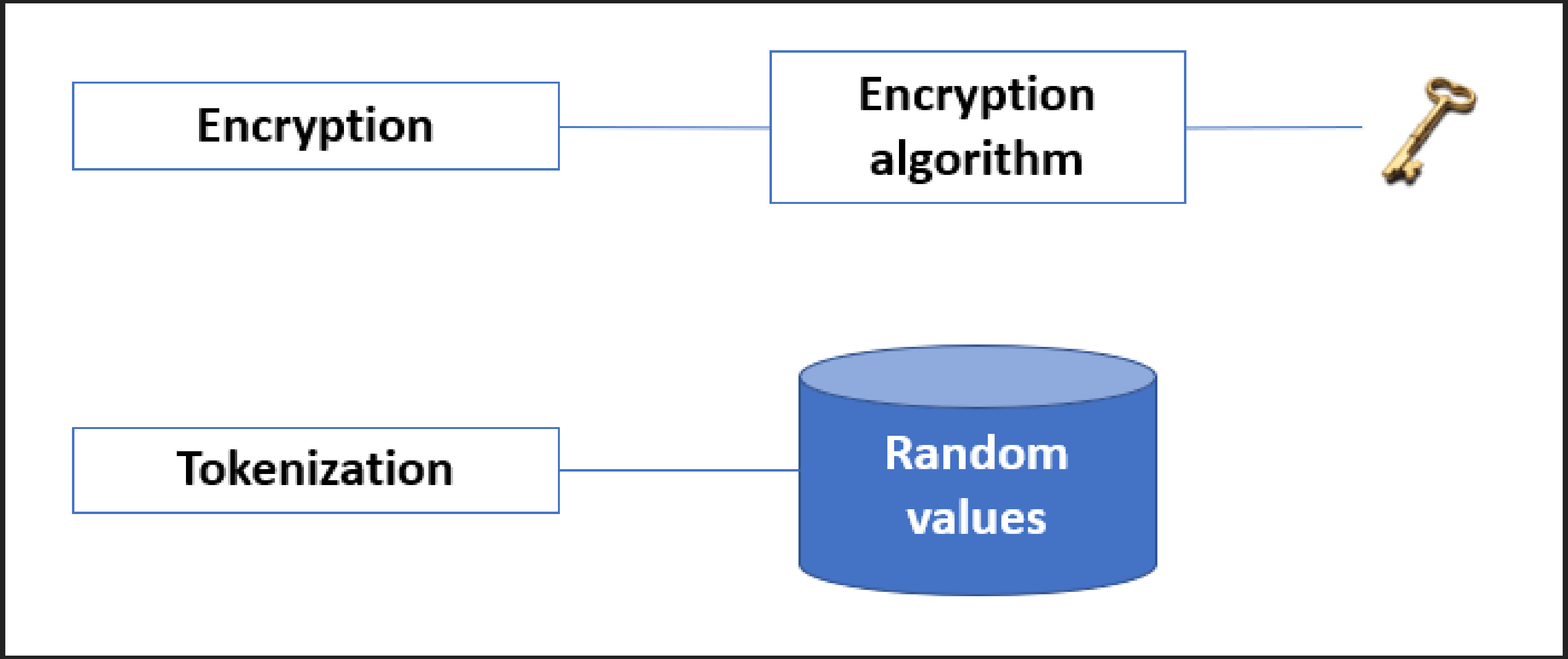
PROTEGRITY

Risk Reduction

Technique name		Use Case / User Story	Data protected in			Data truthfulness at record level	Applicable to types of attributes	Reduces the risk of		
			Transit	Use	Storage			Singling out	Linking	Inference
Pseudonymization	Tokenization	Protects the data flow from attacks	Yes	Yes	Yes	Yes	Direct identifiers	No	Partially	No
Cryptographic tools	Deterministic encryption	Protects the data when not used in processing operations	Yes	No	Yes	Yes	All attributes	No	Partially	No
	Order-preserving encryption	Protects the data from attacks	Partially	Partially	Partially	Yes	All attributes	No	Partially	No
	Homomorphic encryption	Protects the data also when used in processing operations	Yes	Yes	Yes	Yes	All attributes	No	No	No
Suppression	Masking	Protects the data in dev/test and analytical applications	Yes	Yes	Yes	Yes	Local identifiers	Yes	Partially	No
	Local suppression	Protects the data in analytical applications	Yes	Yes	Yes	Yes	Identifying attributes	Partially	Partially	Partially
	Record suppression	Removes the data from the data set	Yes	Yes	Yes	Yes	All attributes	Yes	Yes	Yes
	Sampling	Exposes only a subset of the data for analytical applications	Partially	Partially	Partially	Yes	All attributes	Partially	Partially	Partially
Generalization	Generalization	Protects the data in dev/test and analytical applications	Yes	Yes	Yes	Yes	Identifying attributes	Partially	Partially	Partially
	Rounding	Protects the data in dev/test and analytical applications	Yes	Yes	Yes	Yes	Identifying attributes	No	Partially	Partially
	Top/bottom coding	Protects the data in dev/test and analytical applications	Yes	Yes	Yes	Yes	Identifying attributes	No	Partially	Partially
Randomization	Noise addition	Protects the data in dev/test and analytical applications	Yes	Yes	Yes	No	Identifying attributes	Partially	Partially	Partially
	Permutation	Protects the data in dev/test and analytical applications	Yes	Yes	Yes	No	Identifying attributes	Partially	Partially	Partially
	Micro aggregation	Protects the data in dev/test and analytical applications	Yes	Yes	Yes	No	All attributes	No	Partially	Partially
Privacy models	Differential privacy	Protects the data in analytical applications	No	Yes	Yes	No	Identifying attributes	Yes	Yes	Partially
	K-anonymity	Protects the data in analytical applications	No	Yes	Yes	Yes	Quai identifiers	Yes	Partially	No

Source: INTERNATIONAL STANDARD ISO/IEC 20889

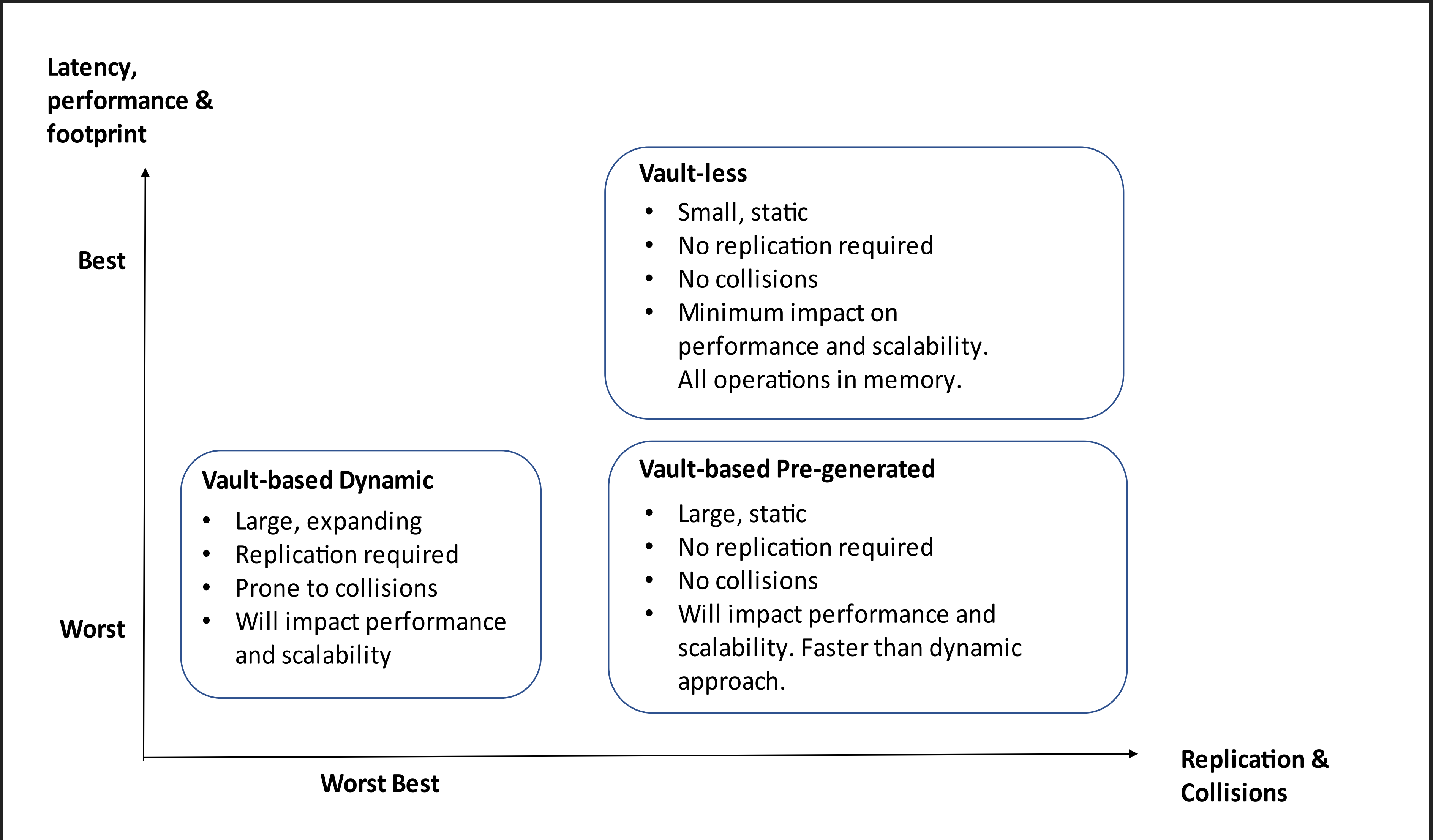
Difference between encryption and tokenization techniques



Source:
INTERNATIONAL STANDARD ISO/IEC 20889

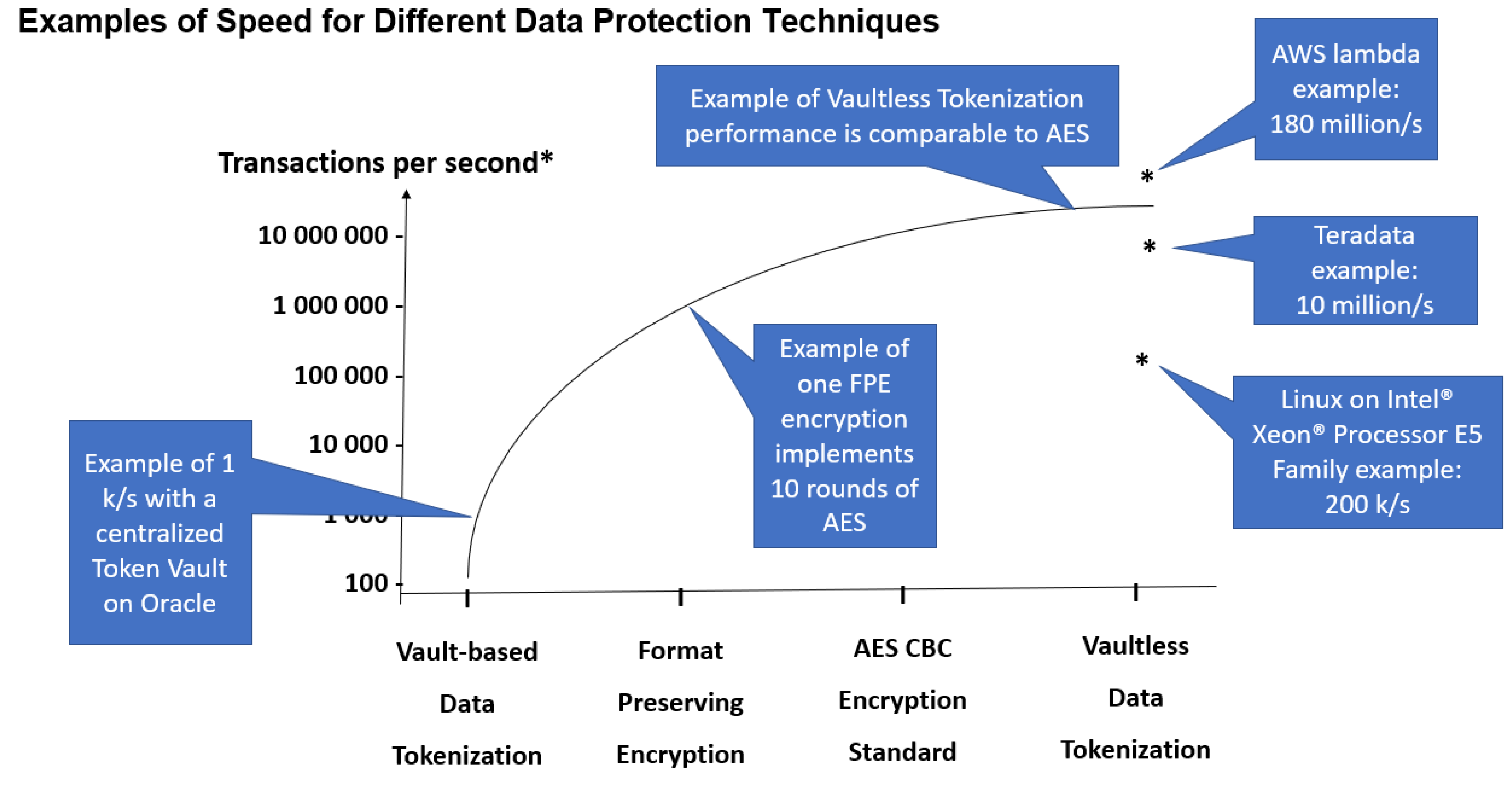
PROTEGRITY

Examples of operational aspects of different tokenization techniques



Source: INTERNATIONAL STANDARD ISO/IEC 20889

What Data Protection Technique is Fastest?



*: Speed will depend on the configuration

k-Anonymity Use Case

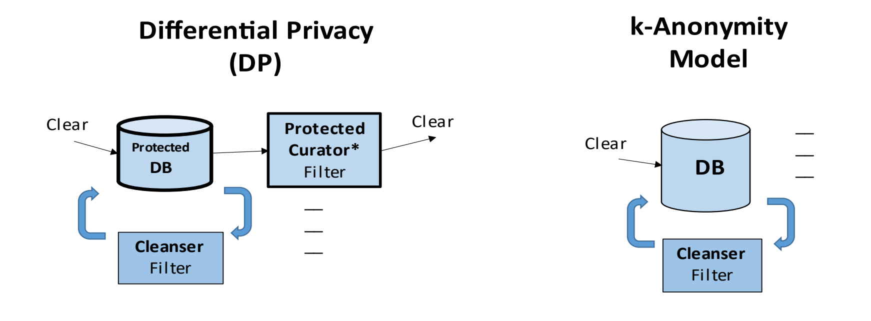
Name	Postcode	Age	Gender	Disease
Patrick	SW1 4YB	22	Male	Cardiovascular
Sebastian	SW1 4ZE	23	Male	Respiratory
Reece	SW1 2HY	18	Male	No Illness
Tilly	NW10 8FN	47	Female	Cancer
Abby	NW10 4AB	42	Female	No Illness
Elise	NW10 0FW	56	Female	Cardiovascular
Morgan	E17 9QY	23	Male	Respiratory
George	E17 3SF	29	Male	Liver
Sienna	E17 5WD	18	Female	Cancer

Name	Postcode	Age	Gender	Disease
*	SW1 *	22	Male	Cardiovascular
*	SW1 *	23	Male	Respiratory
*	SW1 *	18	Male	No Illness
*	NW10 *	47	Female	Cancer
*	NW10 *	42	Female	No Illness
*	NW10 *	56	Female	Cardiovascular
*	E17 *	23	*	Respiratory
*	E17 *	29	*	Liver
*	E17 *	18	*	Cancer

For k-anonymity to be achieved, there need to be at least k individuals in the dataset who share the set of attributes that might become identifying for each individual.

K-anonymity might be described as a ‘hiding in the crowd’ guarantee: if each individual is part of a larger group, then any of the records in this group could correspond to a single person.

This second table shows the data anonymised to achieve k-anonymity of k = 3, as you can see this was achieved by generalising some quasi-identifier attributes and redacting some others.



PROTEGRITY

11 Published International Privacy Standards

ISO Privacy Standards

- Techniques
- Management
- Cloud
- Framework
- Impact
- Requirements
- Process

20889 IS Privacy enhancing de-identification terminology and classification of techniques

27701 IS Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines

27018 IS Code of practice for protection of PII in public clouds acting as PII processors

29100 IS Privacy framework

29101 IS Privacy architecture framework

29134 IS Guidelines for Privacy impact assessment

29190 IS Privacy capability assessment model

29191 IS Requirements for partially anonymous, partially unlinkable authentication

29151 IS Code of Practice for PII Protection

19608 TS Guidance for developing security and privacy functional requirements based on 15408

27550 TR Privacy engineering for system lifecycle processes

IS: International Standard

Guidelines to help comply with ethical standards

TS: Technical Specification

TR: Technical Report

PROTEGRITY

References A:

1. C. Gentry. “A Fully Homomorphic Encryption Scheme.” Stanford University. September 2009, https://crypto.stanford.edu/craig/craig-thesis.pdf
2. Status Report on the Second Round of the NIST Post-Quantum Cryptography Standardization Process, https://csrc.nist.gov/publications/detail/nistir/8309/final
3. ISO/IEC 29101:2013 (Information technology – Security techniques – Privacy architecture framework)
4. ISO/IEC 19592-1:2016 (Information technology – Security techniques – Secret sharing – Part 1: General)
5. ISO/IEC 19592-2:2017 (Information technology – Security techniques – Secret sharing – Part 2: Fundamental mechanisms
6. Homomorphic Encryption Standardization, Academic Consortium to Advance Secure Computation, https://homomorphicencryption.org/standards-meetings/
7. Homomorphic Encryption Standardization, https://homomorphicencryption.org/
8. NIST Post-Quantum Cryptography PQC, https://csrc.nist.gov/Projects/Post-Quantum-Cryptography
9. UN Handbook on Privacy-Preserving Computation Techniques, http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf
10. ISO/IEC 29101:2013 Information technology – Security techniques – Privacy architecture framework, https://www.iso.org/standard/45124.html

11. Homomorphic encryption, https://brilliant.org/wiki/homomorphic-encryption/

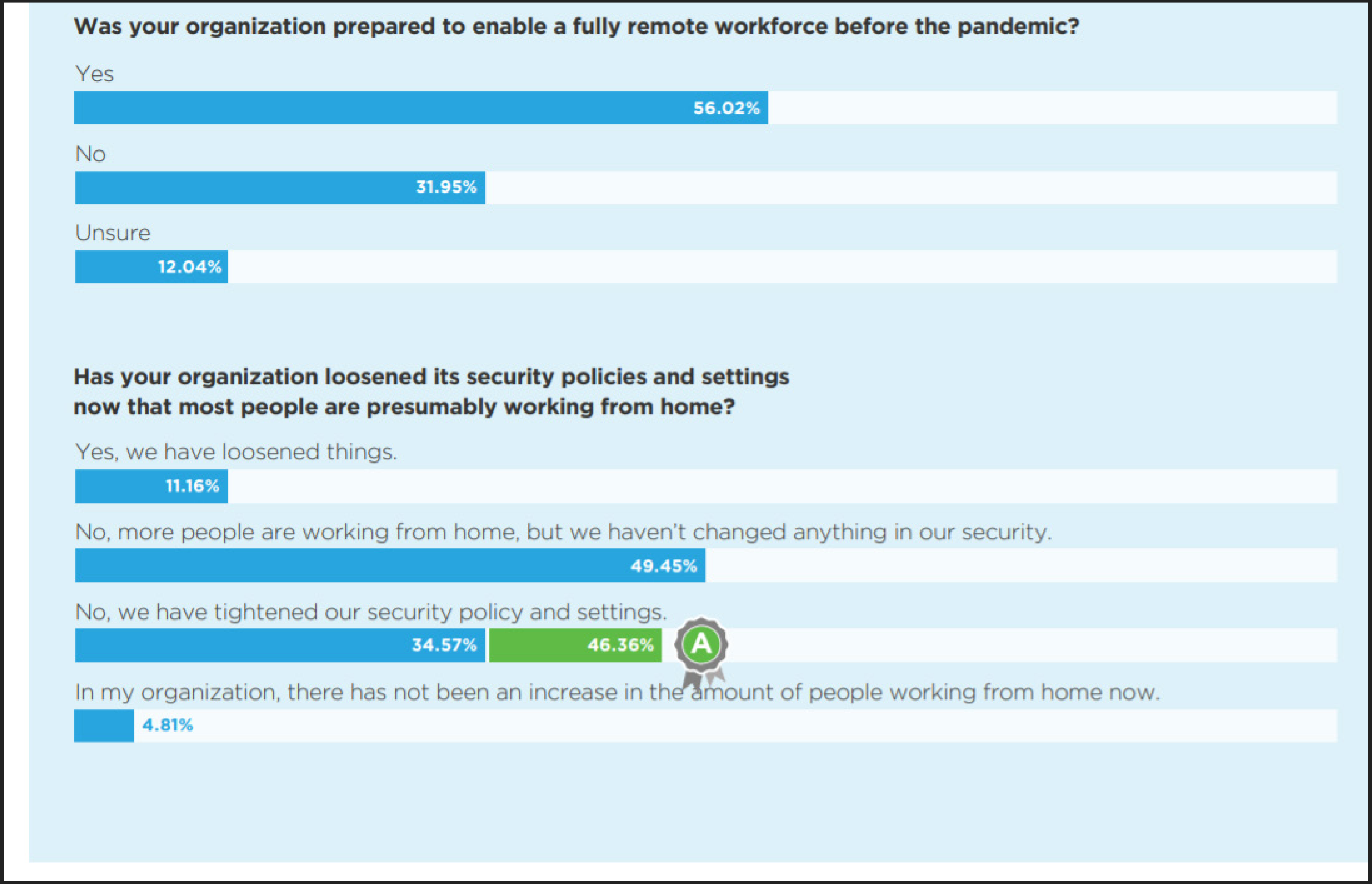
Thank You!

Ulf Mattsson

Chief Security Strategist

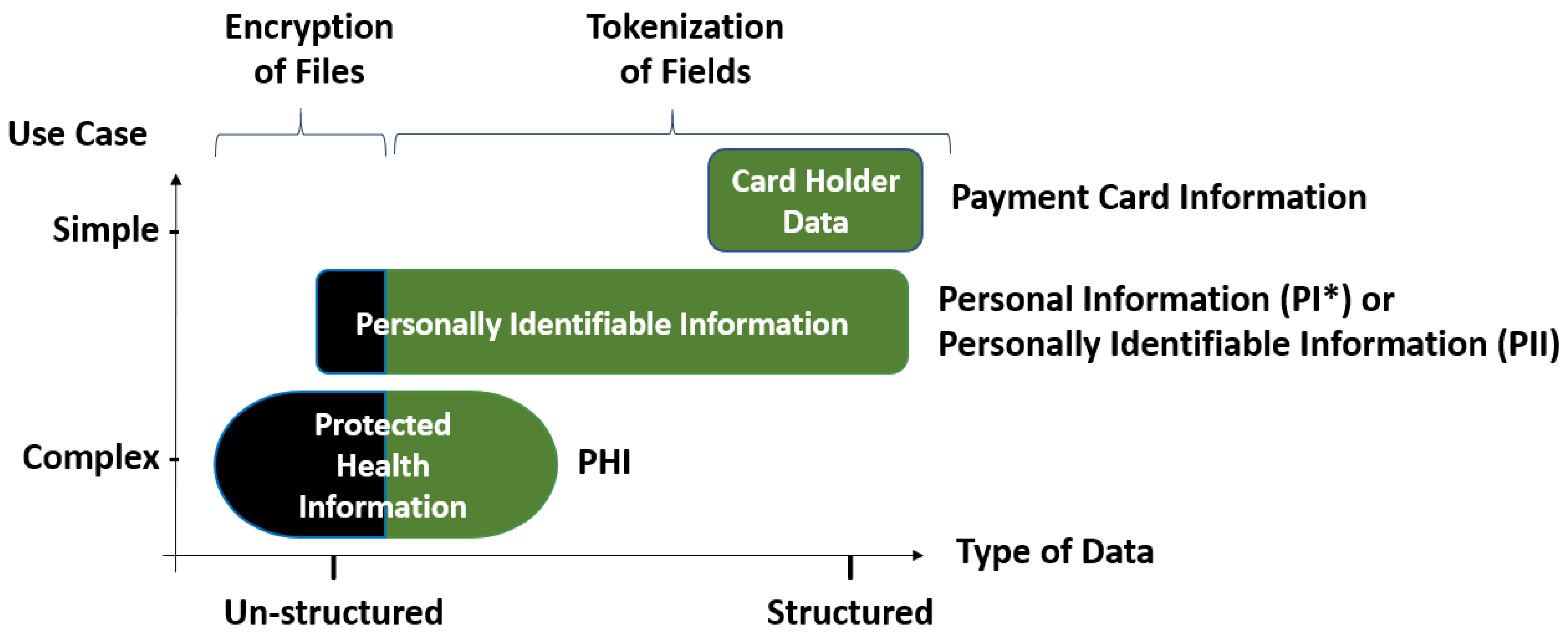
www.Protegrity.com

PROTEGRITY



Was your organization prepared to enable a fully remote workforce before the pandemic?
Yes
56.02%
No
31.95%
Unsure
12.04%
Has your organization loosened its security policies and settings now that most people are presumably working from home?
Yes, we have loosened things.
11.16%
No, more people are working from home, but we haven't changed anything in our security.
49.45%
No, we have tightened our security policy and settings.
34.57%
46.36%
A
In my organization, there has not been an increase in the amount of people working from home now.
4.81%

How to protect different types of data with encryption and tokenization



*: California CCPA

*: California CCPA

Data Security Management for Hybrid Cloud

Consistency

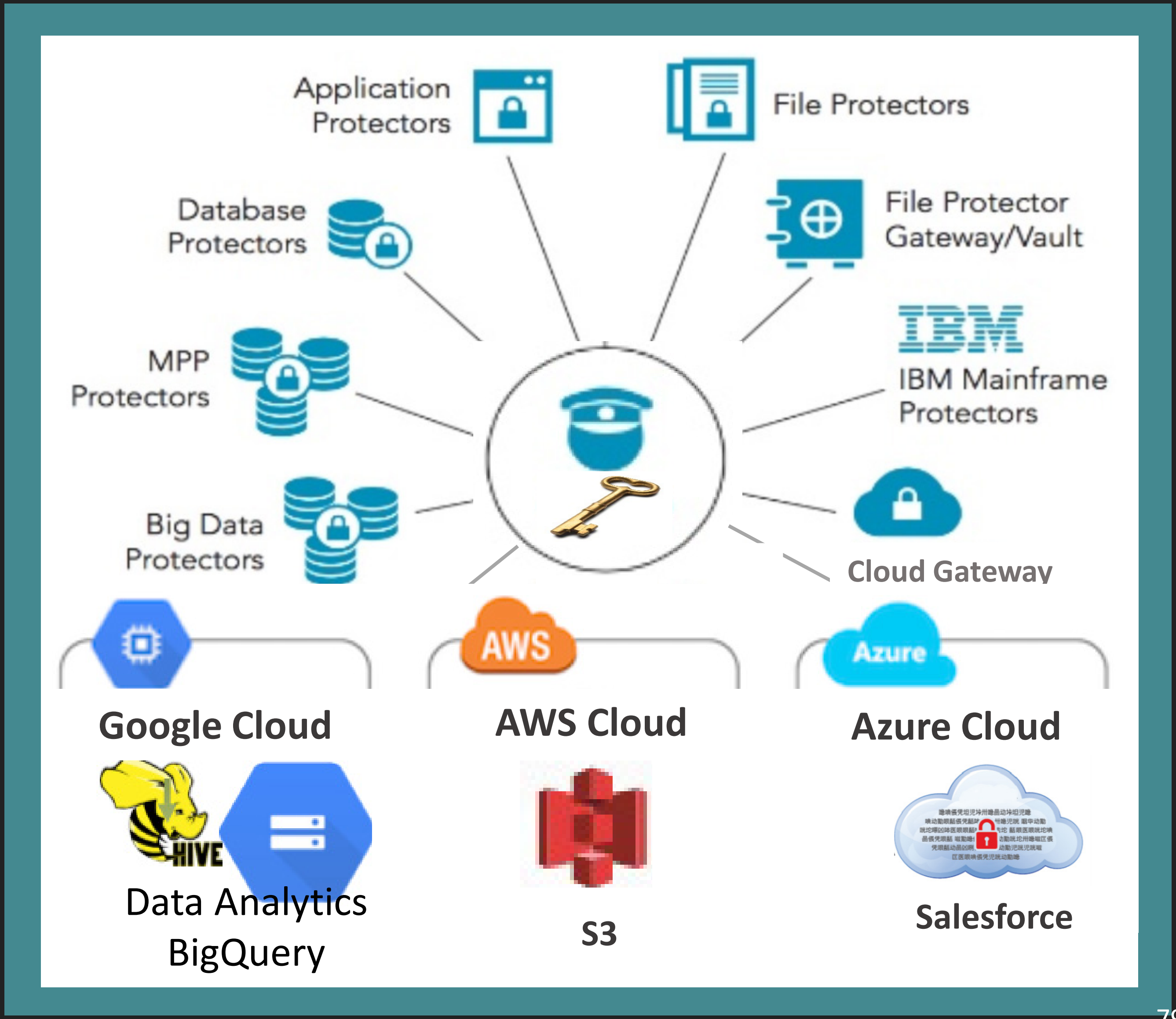
- Most firms are quite **familiar with their on-premises encryption and key management** systems, so they often prefer to leverage the **same tool and skills across multiple clouds**.
- Firms often adopt a **“best of breed” cloud** approach.

Trust

- Some customers simply do not trust their vendors.

Vendor Lock-in and Migration

- A common concern is **vendor lock-in**, and an inability to migrate to another cloud service provider.



Securosis, 2019

PROTEGRITY

Weakness of Searchable Encryption

Unencrypted Data (Plaintext)

1: 123-12-1234

2: 123-12-1235

3: 123-12-1234

Encrypted Data (Ciphertext)

1: MjM0MjM0MjM0LTEyMz Q1NjIzNDM0DQo=

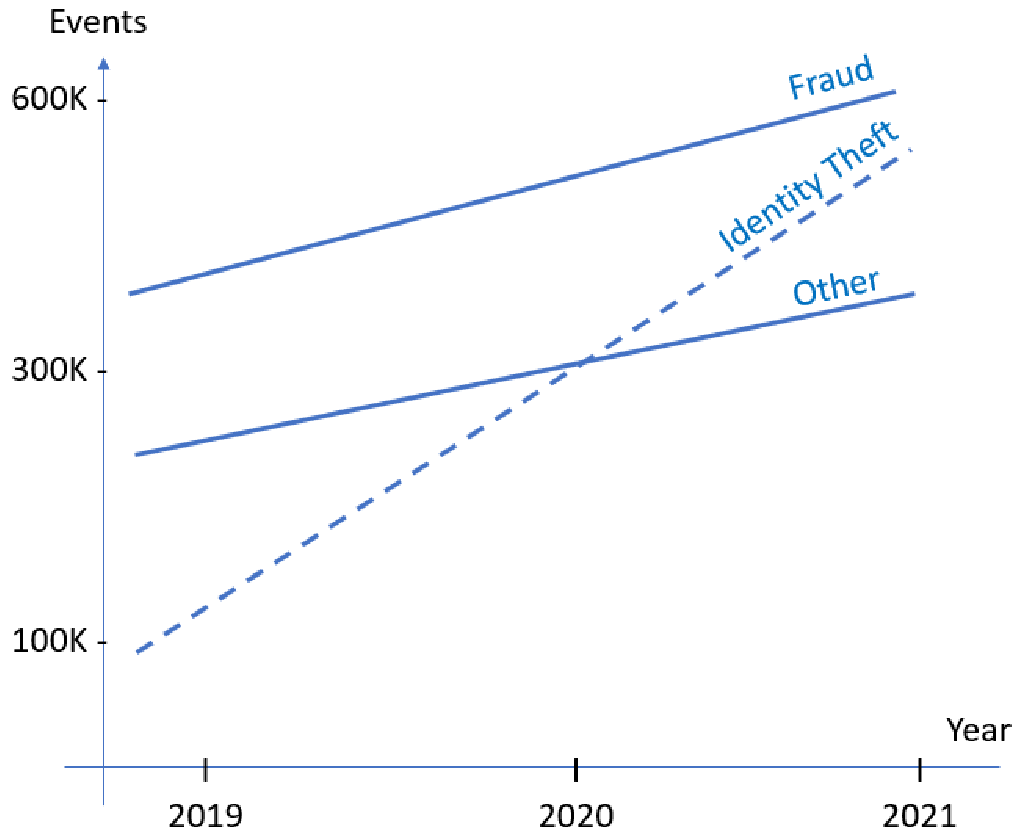
2: VGhpcyBpcyBhIHNhbXB lwaGVydGV4dC43=

3: MjM0MjM0MjM0LTEyMz Q1NjIzNDM0DQo=

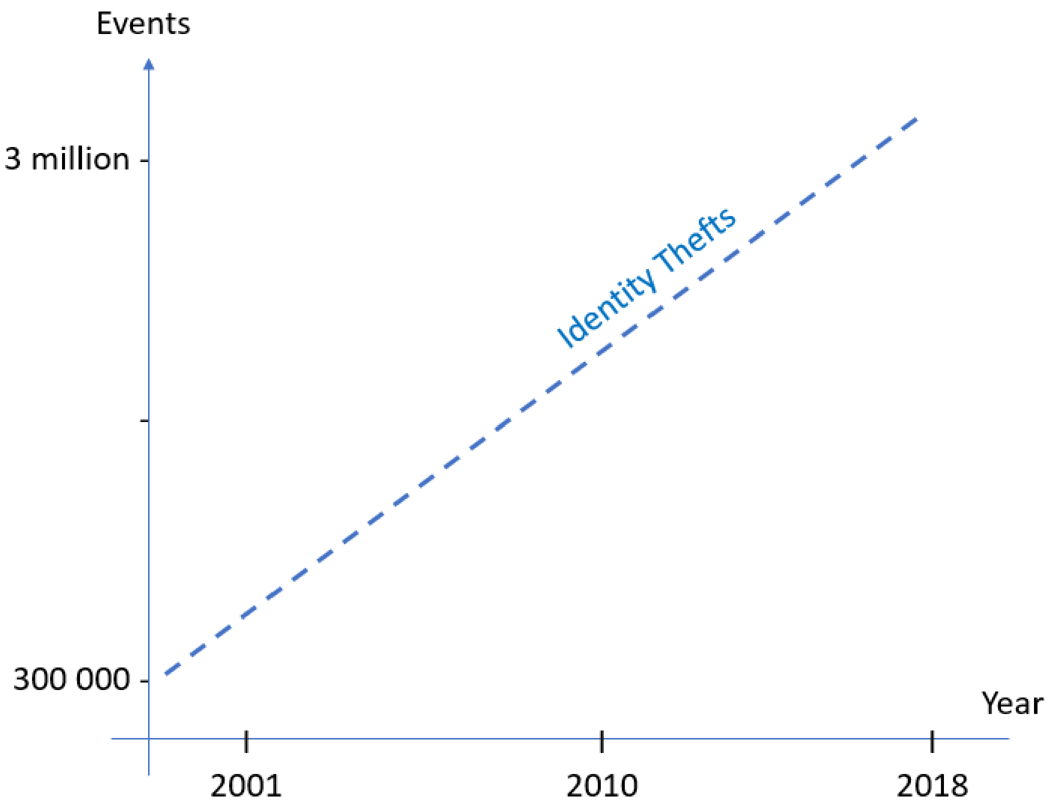
Search is utilized in virtually every application and is critical in a collaborative cloud environment. As mentioned, regular encryption hides data so well that search is not feasible. However, it is possible to efficiently search on encrypted data if one is willing to sacrifice some security. In general, any efficiently searchable encryption algorithm shares a common security weakness: equality of keywords is leaked, making certain statistical attacks possible.

The Landscape

Fraud & Identity Thefts



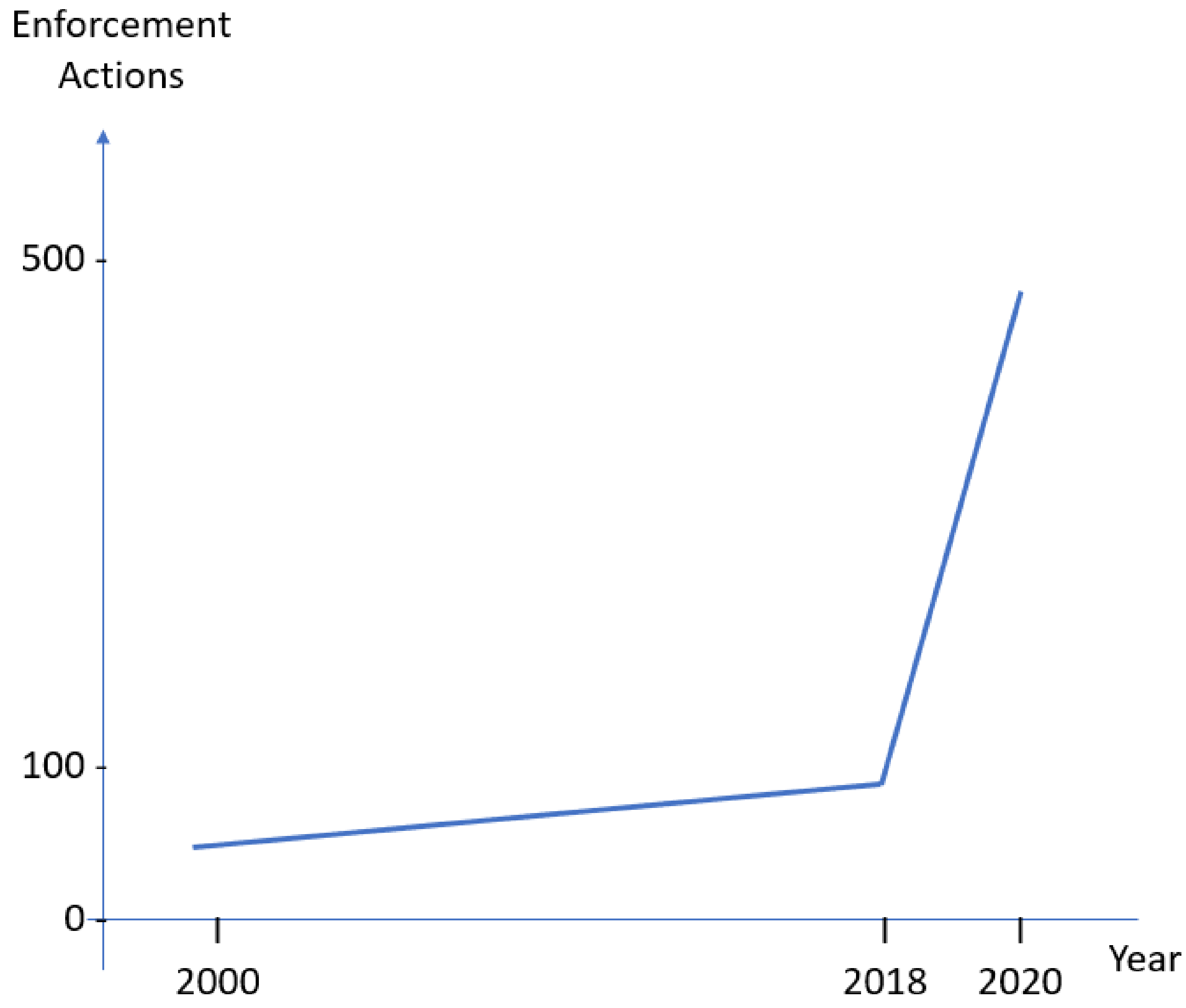
Source: Adapted from FTC Identity Thefts Report, US Federal Trade Commission (FTC)



Source: Adapted from FTC Identity Thefts Report, US Federal Trade Commission (FTC)

PROTEGRITY

Data Privacy Enforcement Actions Worldwide

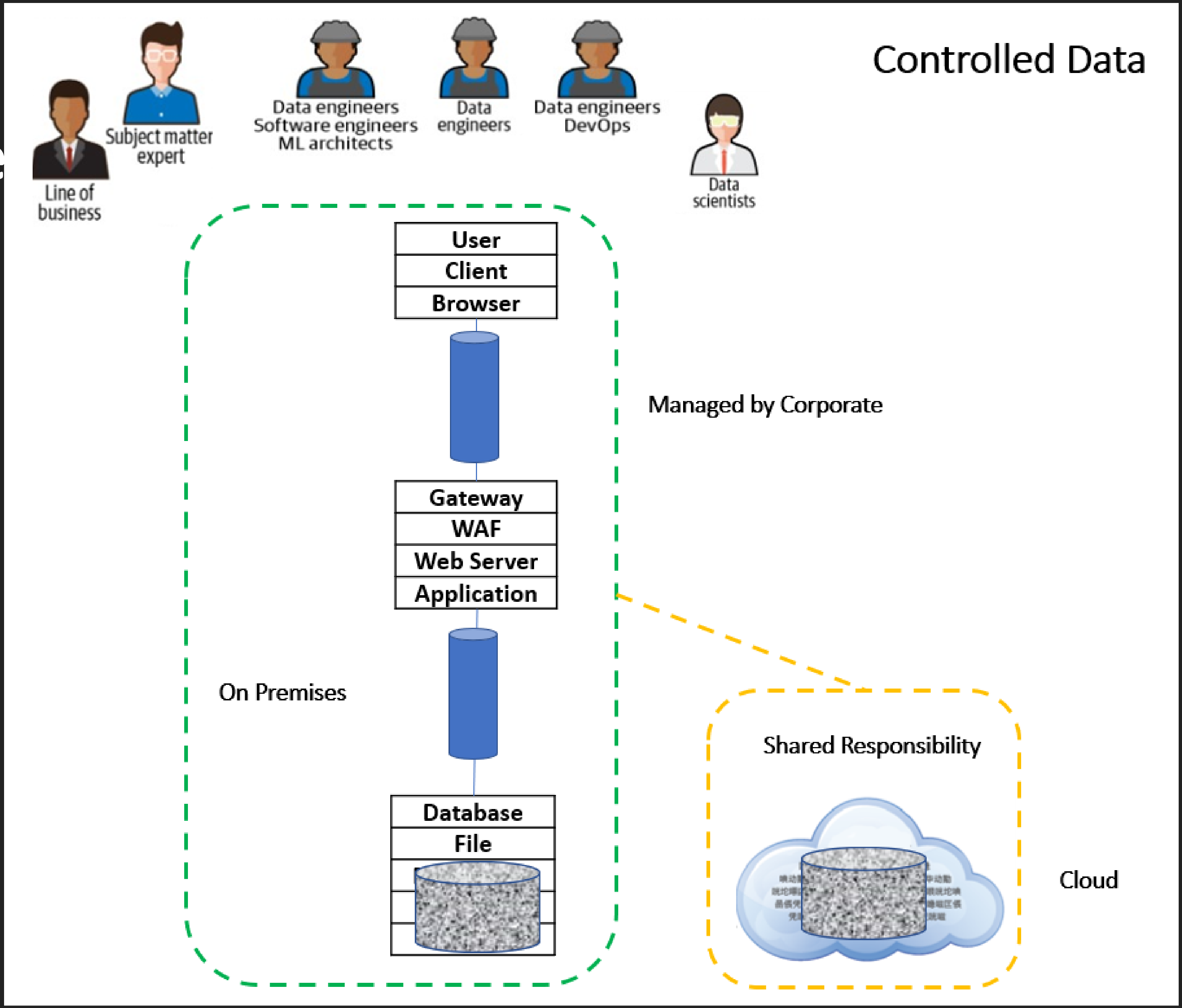


Source: Adapted from The PwC Privacy Policy Database, Data Privacy Enforcement Actions Worldwide reported PwC

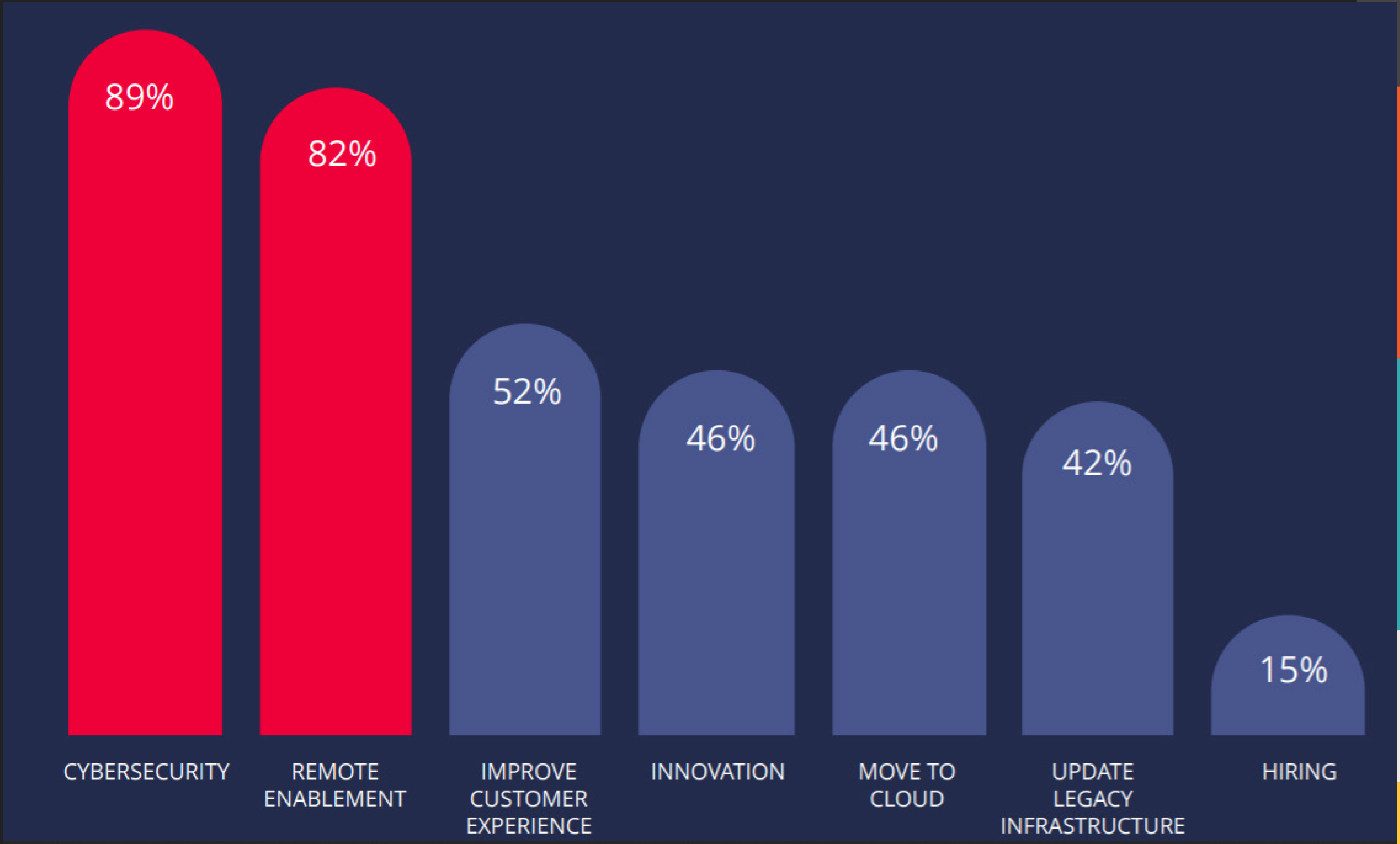
[redacted]

[redacted]

The Old Corporate IT Environment



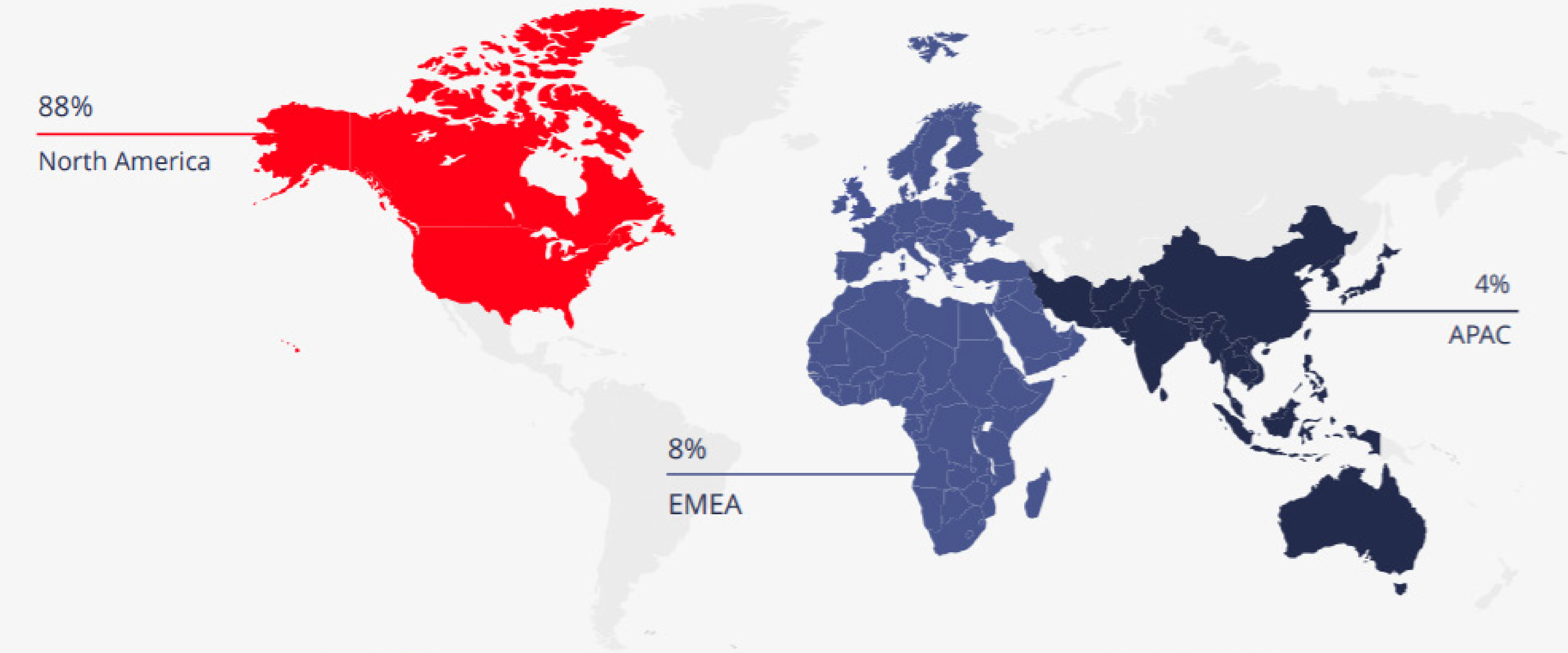
Study: Top Priorities in 2020



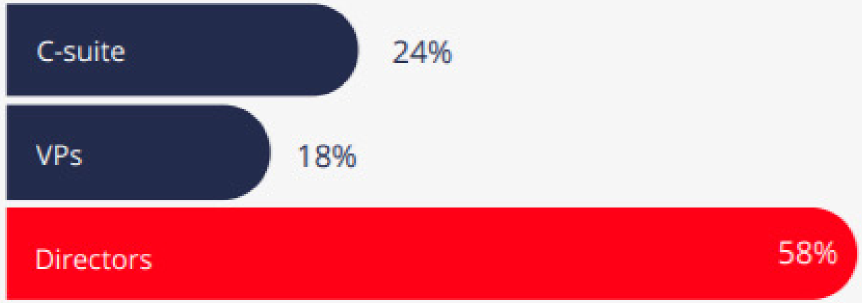
Source: Hitachi

PROTEGRITY

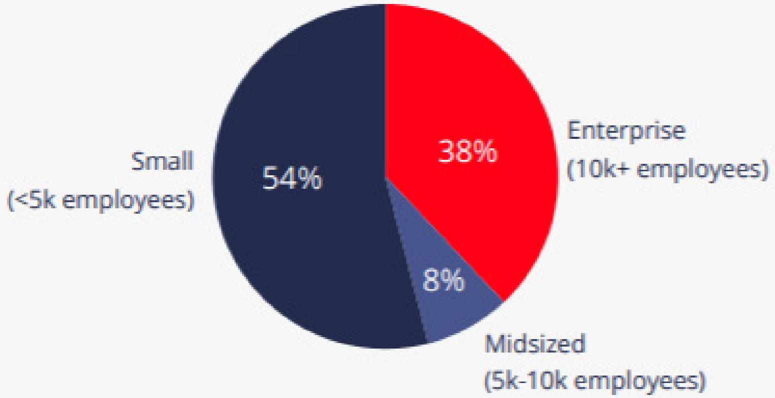
Study: Top Priorities in 2020



TITLES

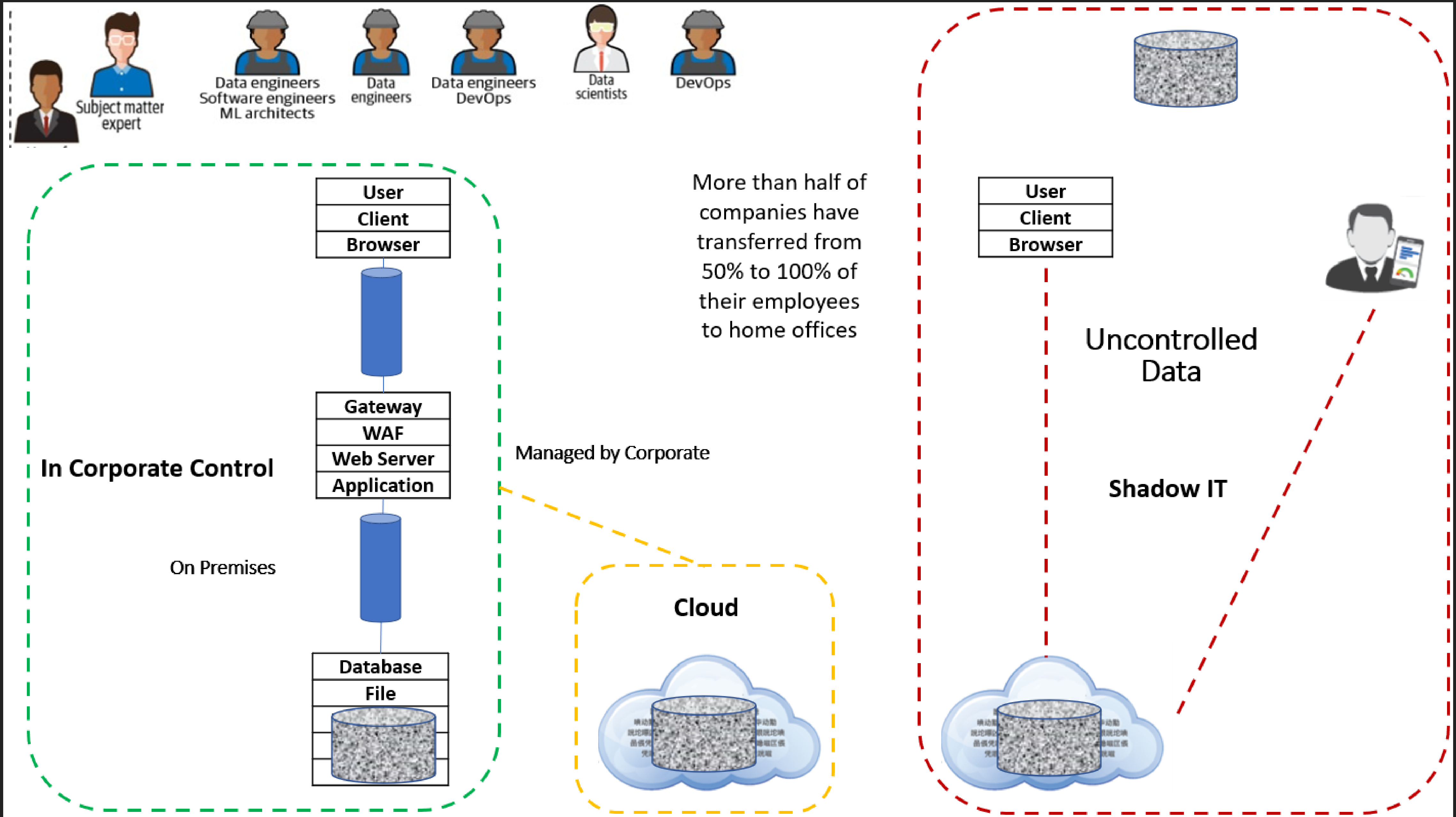


COMPANY SIZE

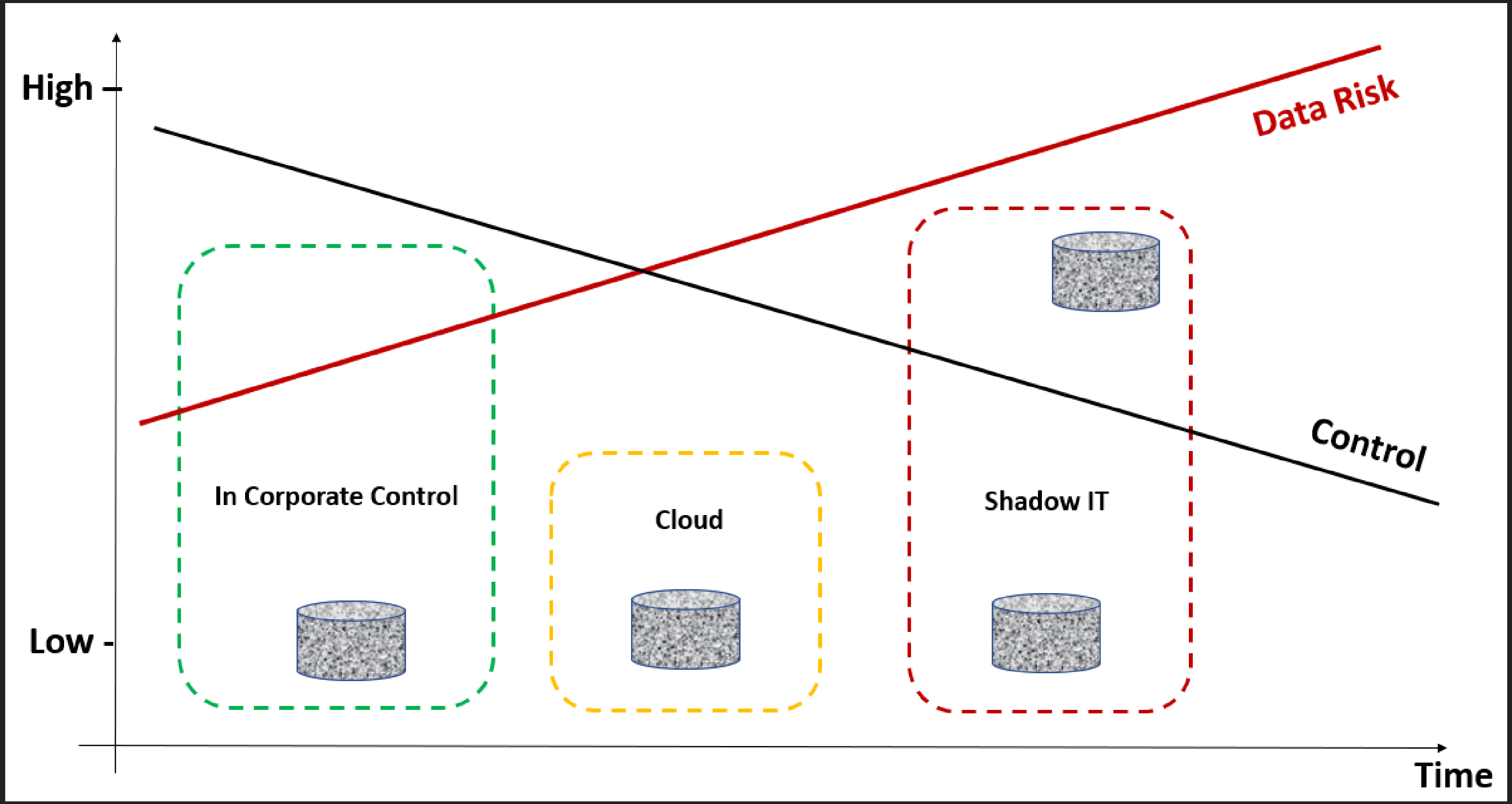


Hitachi

An Increasingly Distributed Environment



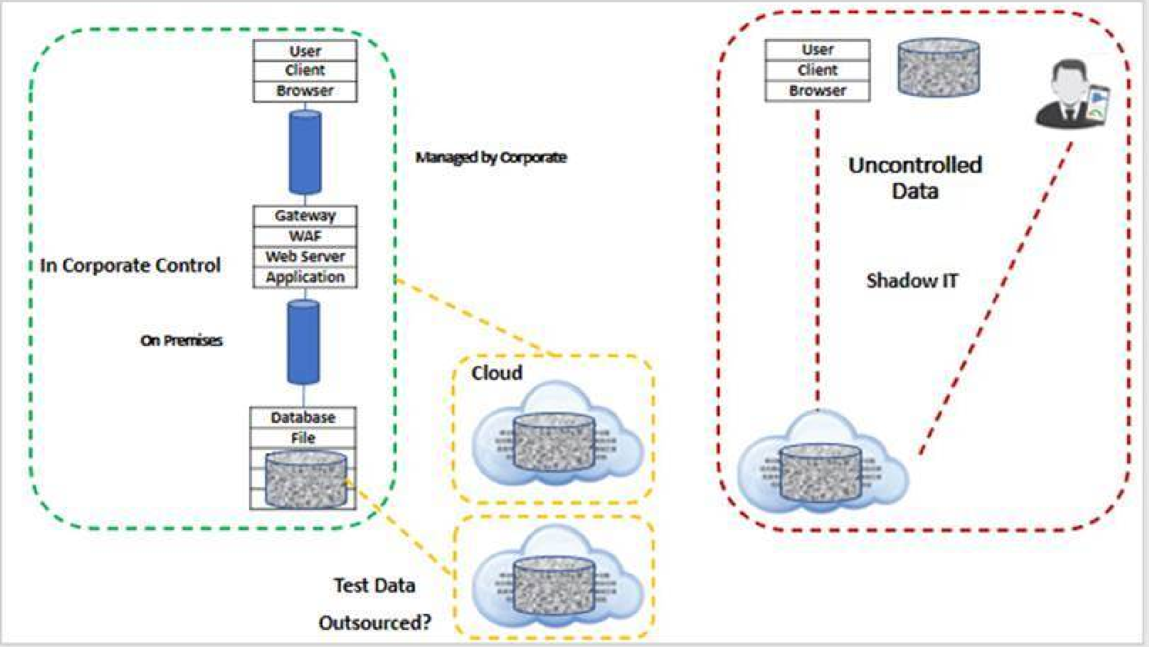
Risks & Control in our New Distributed Environment

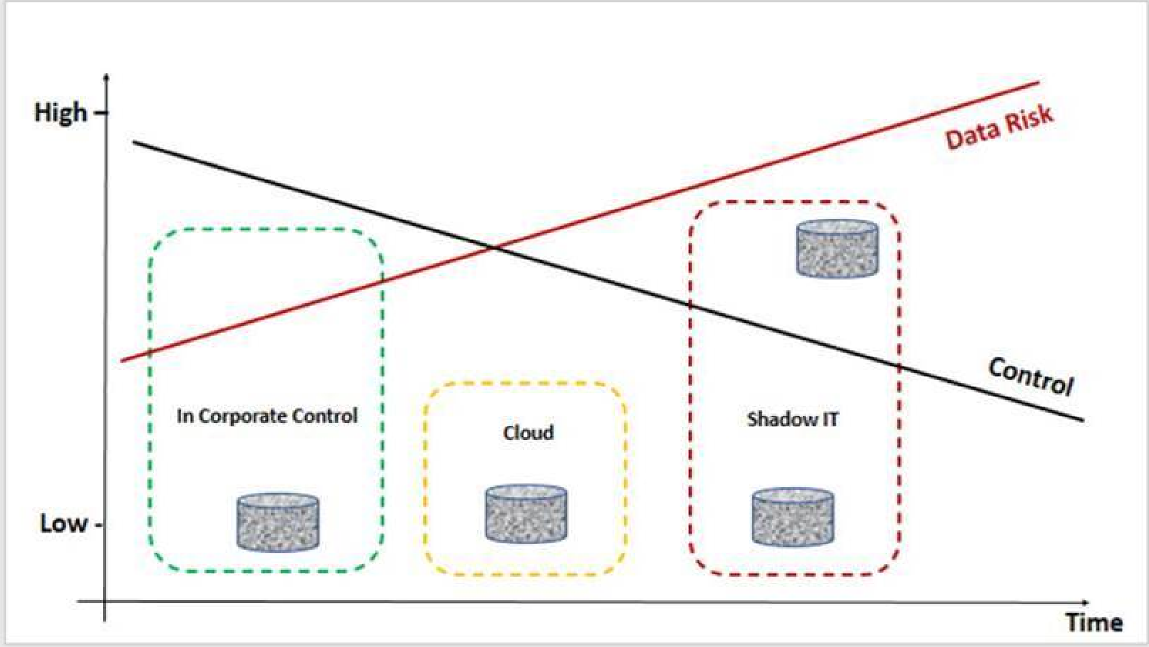


PROTEGRITY

Under control? Is the situation getting worse?

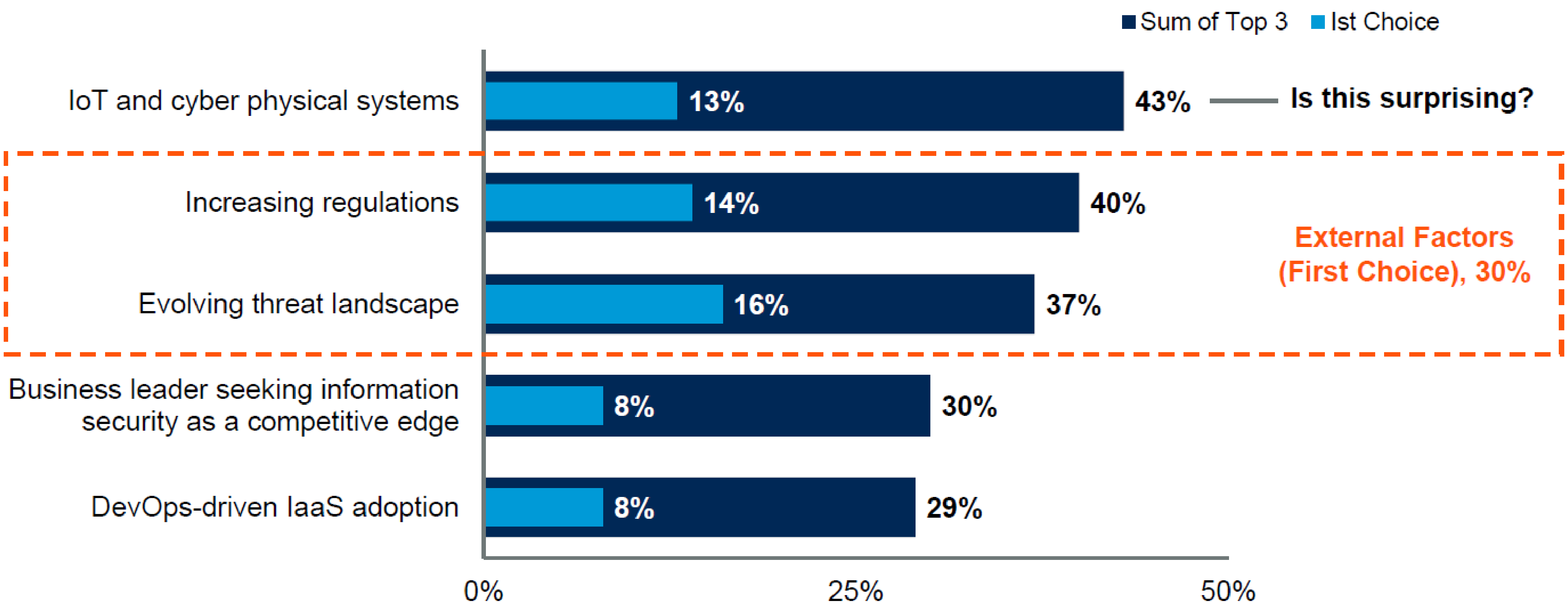
1. How do we control privacy of Test Data? Using Prod Data to meet Their Goals? Outsourced testing?
2. Do we have increasingly less control over distributed data when working from home? IAttack Surface increasing?
3. Is compliance under control? Is the situation getting worse?
4. How much is End-point security helping? How can we protect against Supply Chain Attacks? Solarwinds?





PROTEGRITY

Factors Impacting Information Security Functions in Three to Five Years



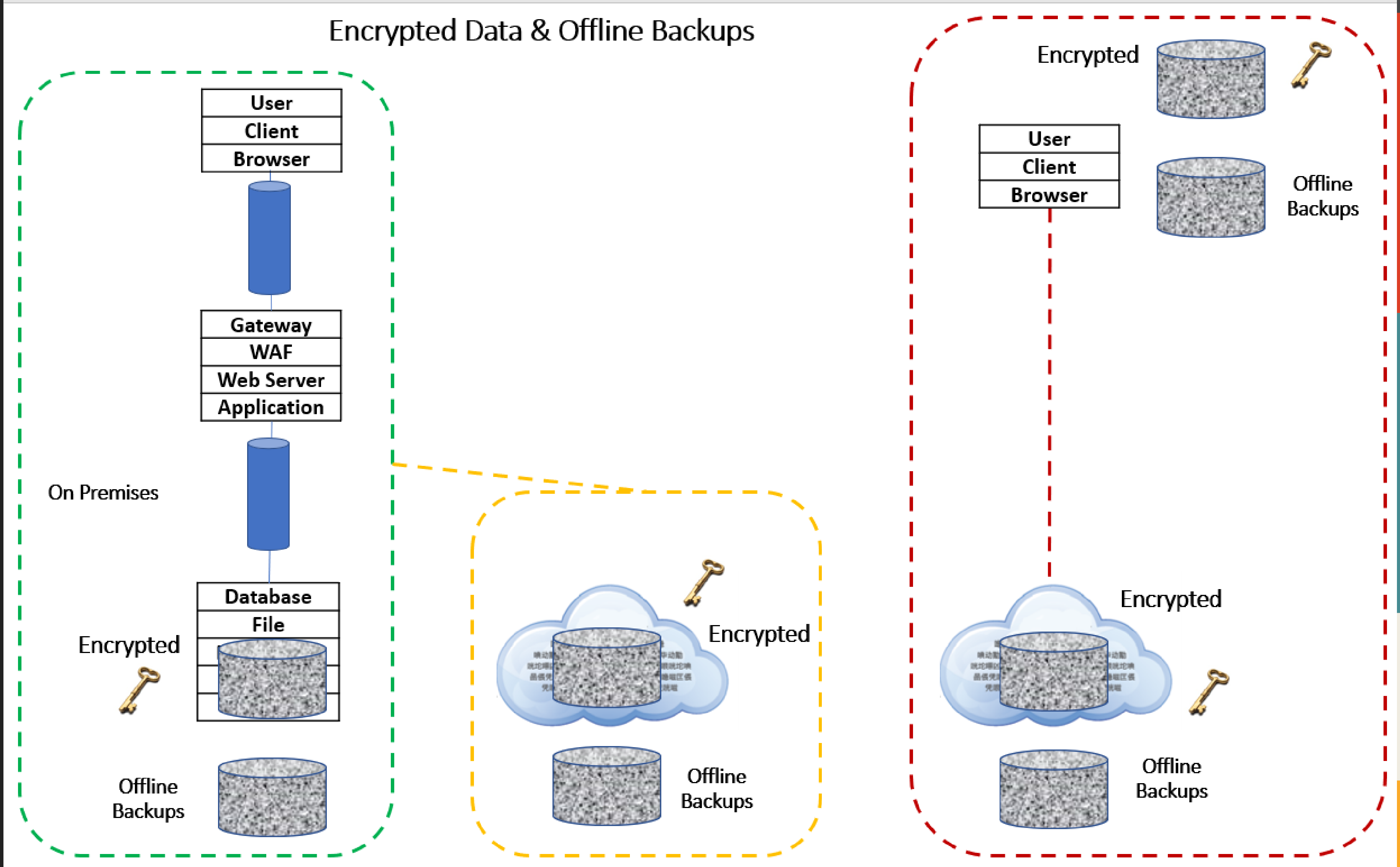
n = 403, All Respondents, Excluding Don't Know or Refused
Q01. What are your organization's top three drivers that are likely to impact its information security function and controls in the next three to five years?
Source: Gartner 2020 Security & IAM Solution Adoption Trends Survey

Gartner

Do I need to address Ransomware and other Attacks on Data?

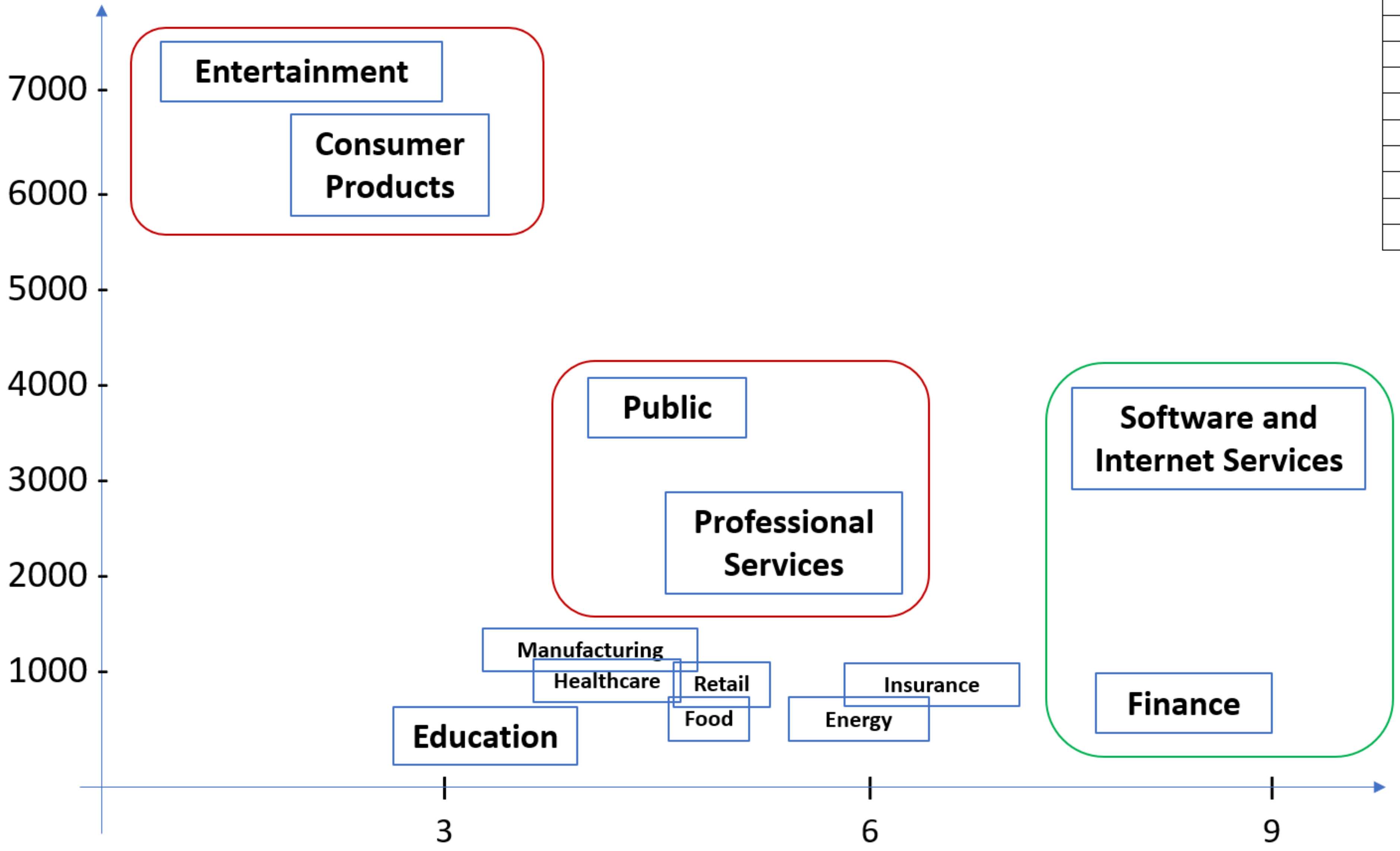
How can we protect our Data?

1. Monitor Access
2. Protect Data
3. Offline Backups
4. Test restore procedures



Number of Security Incidents and Breaches by Victim Industry (Verizon DBIR 2021)

Breaches by Security Spending Budget



% IT Security Spending as Percent of Total IT Spending by Industry (Gartner)

Breaches in Recent Years

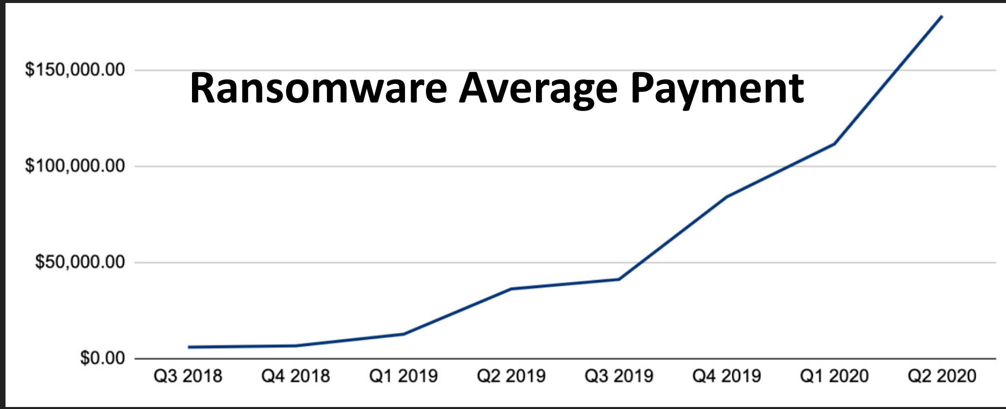
Victim	# Records
Yahoo	3 billion
First American Financial	885 milion
Facebook	540 million
IdentityForce	280 million
TikTok	235 million
Marriott	500 million
Adult Friend	412 million
Experian	200 million
Deep Root Analytics	200 million
Anthem	80 Anthem
Target	70 million
MGM Resorts	11 million
Mashable	2 million
Ancestry	60 k

Organizations using Fine-grained encryption or tokenizing

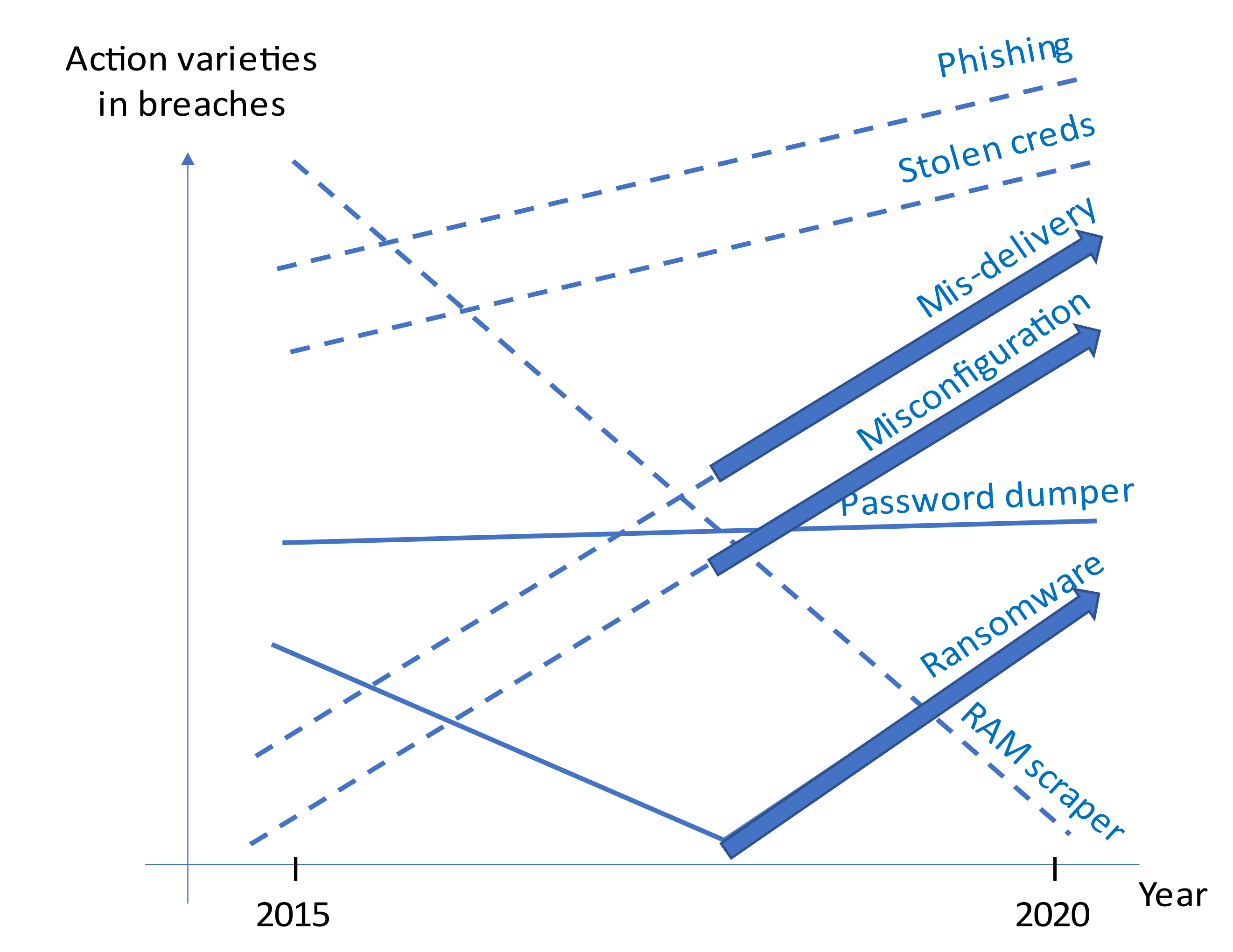
- 30+ Retail
- 20+ Financial Services
- 15+ Government
- 10+ Travel
- 5+ Energy
- 5+ Manufacturing
- 5+ Payment Providers
- 5+ Energy

PROTEGRITY

Ransomware and other Breaches on The Rise



BankinfoSecurity.com



Verizon DBIR 2021

Electronic Health Records:

- **Healthcare's attack surface has grown** considerably over the last two decades.

https://threatpost.com

Ransomware

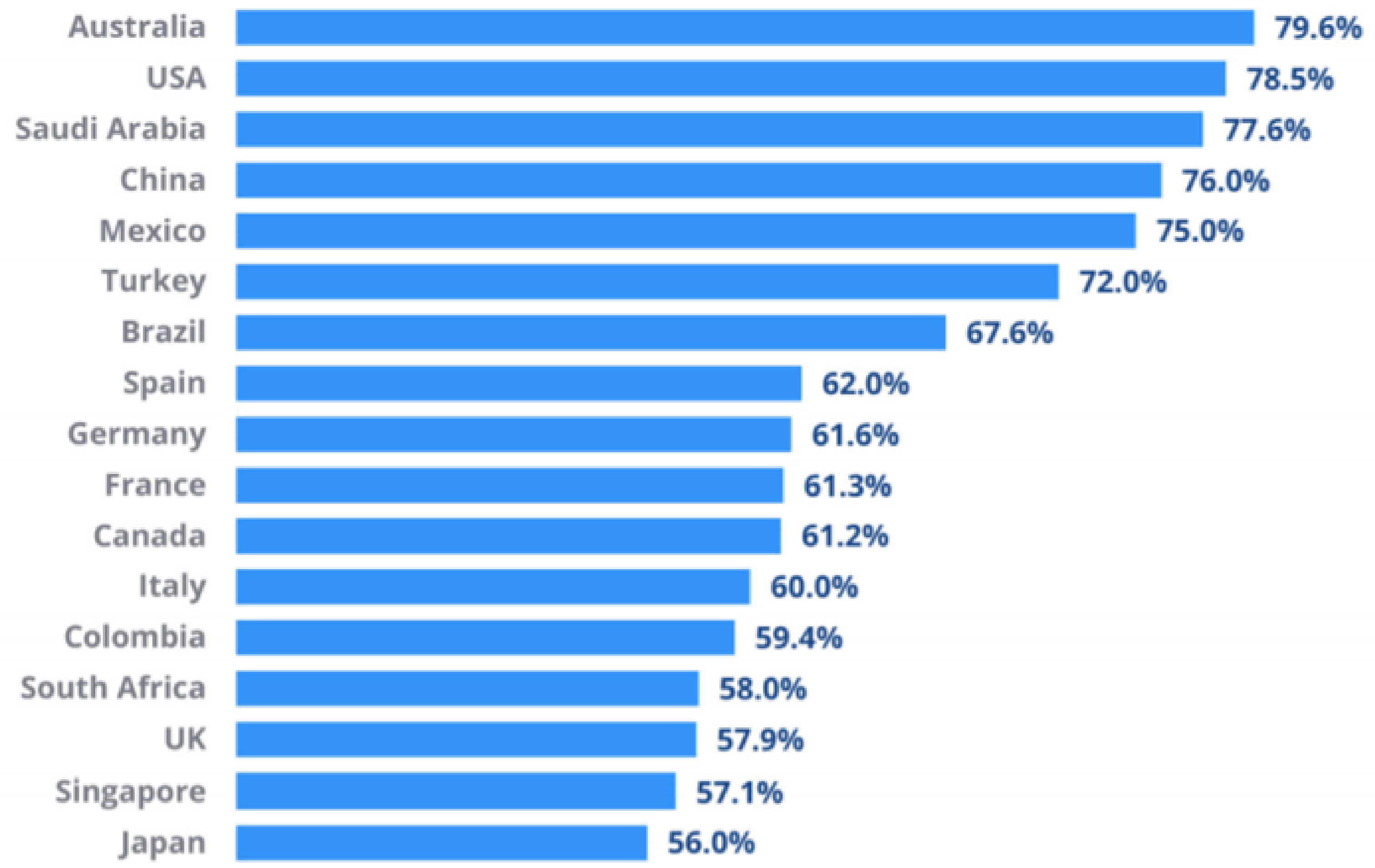


Figure 17: Percentage of organizations affected by ransomware in the last 12 months, by country.

Source: https://www.isc2.org/-/media/ISC2/Research/Cyberthreat-Defense-Report/2021/CyberEdge-2021-CDR-Report-v10--ISC2-Edition.ashx?la=en&hash=60BC7C7969857E2FF07B714896F079EF5C9C1C39

Ransomware

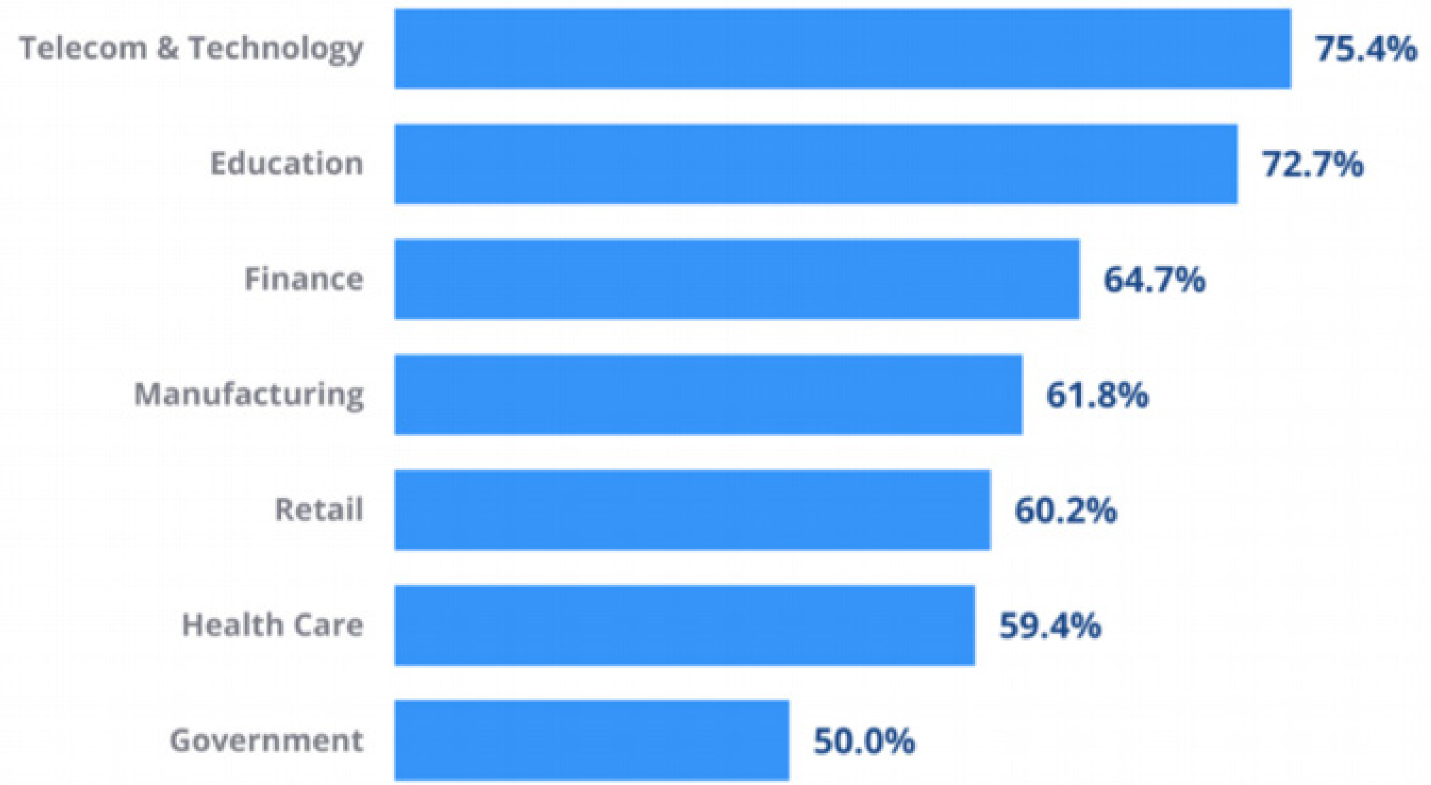
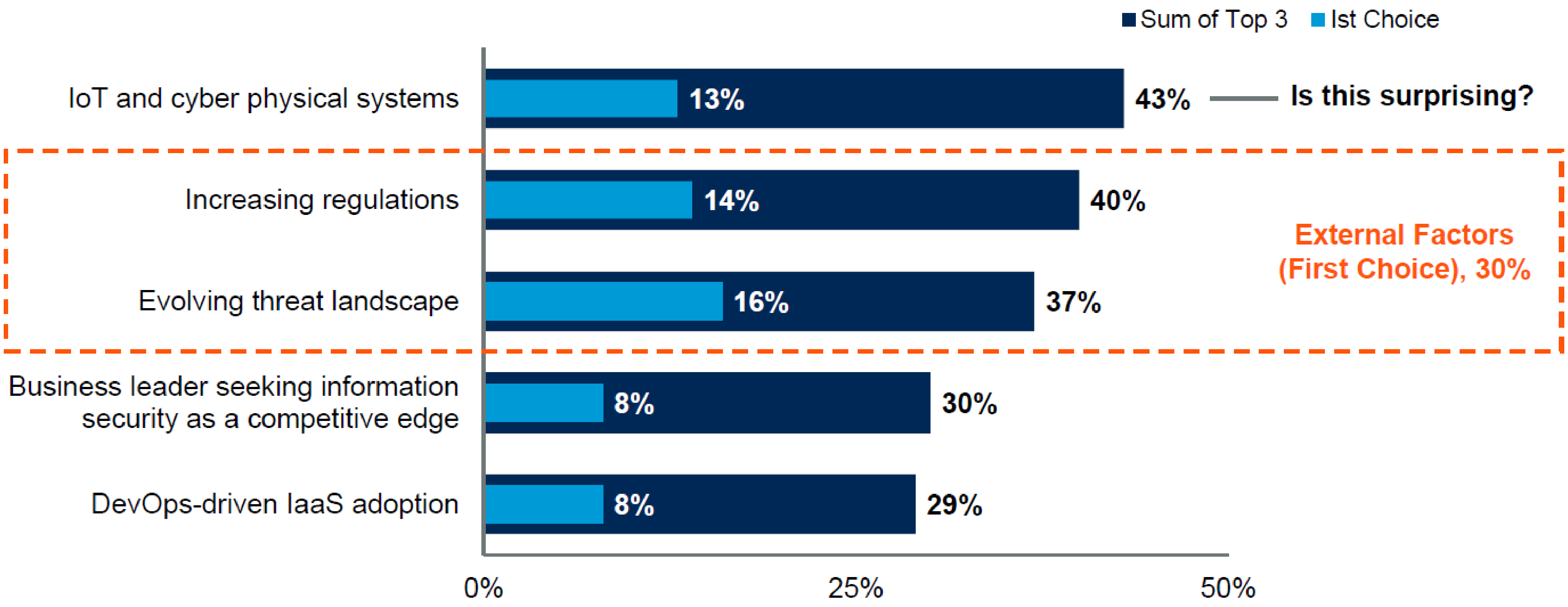


Figure 18: Percentage of organizations affected by ransomware in the last 12 months, by industry.

Source: https://www.isc2.org/-/media/ISC2/Research/Cyberthreat-Defense-Report/2021/CyberEdge-2021-CDR-Report-v10--ISC2-Edition.ashx?la=en&hash=60BC7C7969857E2FF07B714896F079EF5C9C1C39

Factors Impacting Information Security Functions in Three to Five Years



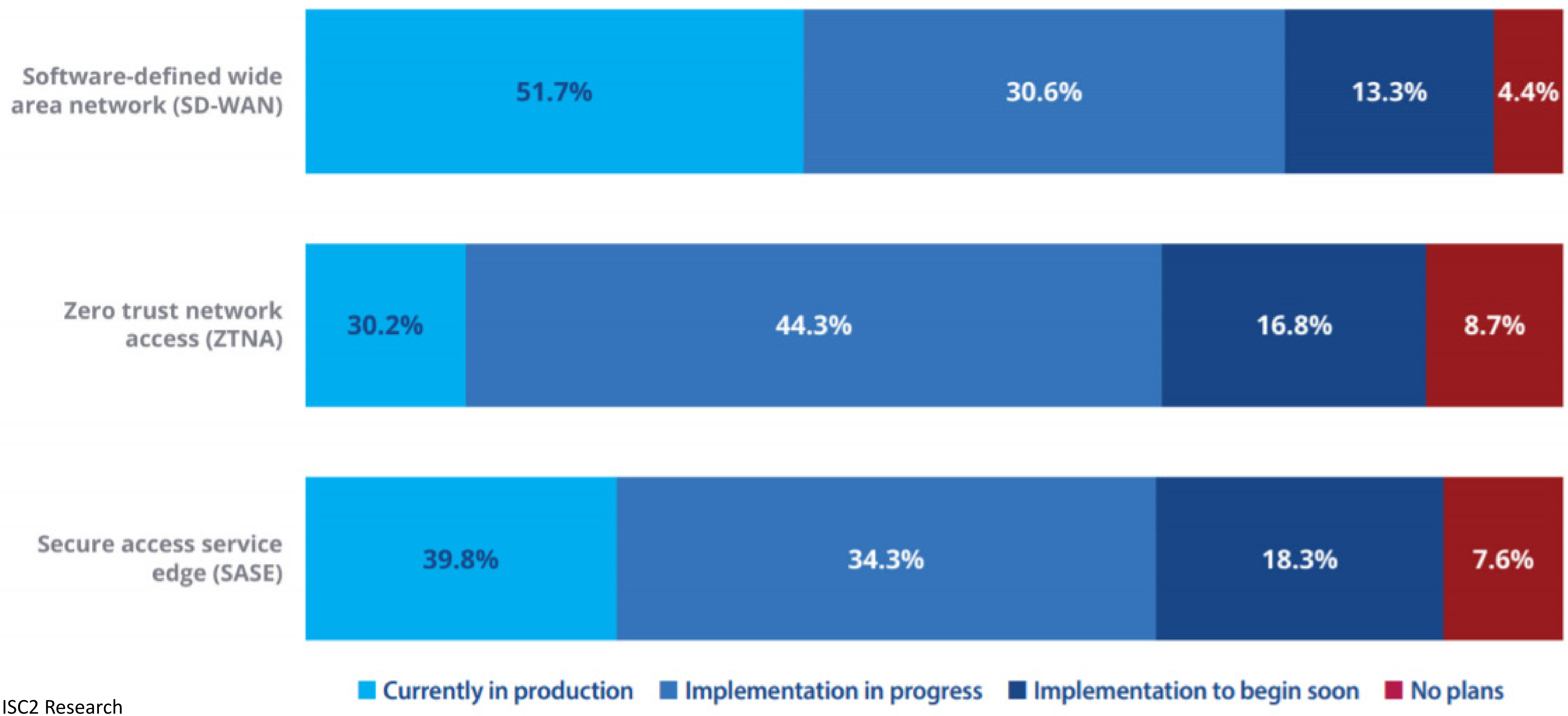
n = 403, All Respondents, Excluding Don't Know or Refused
Q01. What are your organization's top three drivers that are likely to impact its information security function and controls in the next three to five years?
Source: Gartner 2020 Security & IAM Solution Adoption Trends Survey

10

Gartner

Emerging IT Security Technologies

Describe your organization's deployment plans for each of the following emerging IT security technologies / architectures.



ISC2 Research

Cloud

Multi-cloud Policies & Controls

Microsoft Azure

aws

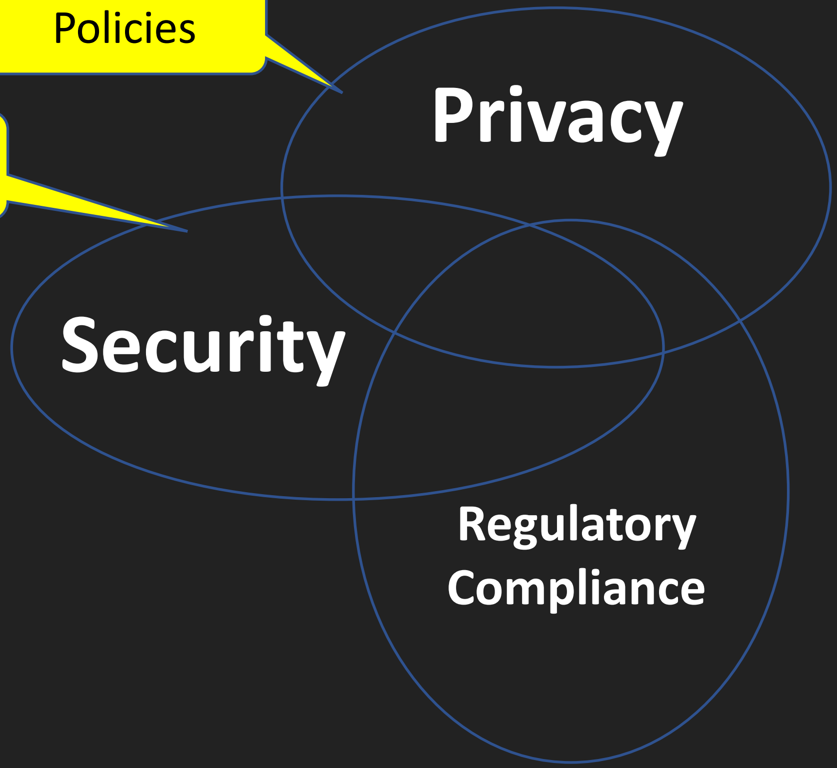
Google Cloud

Applications, Data and Users
Configuration as code (software)
Cloud provider infrastructure

Configuration

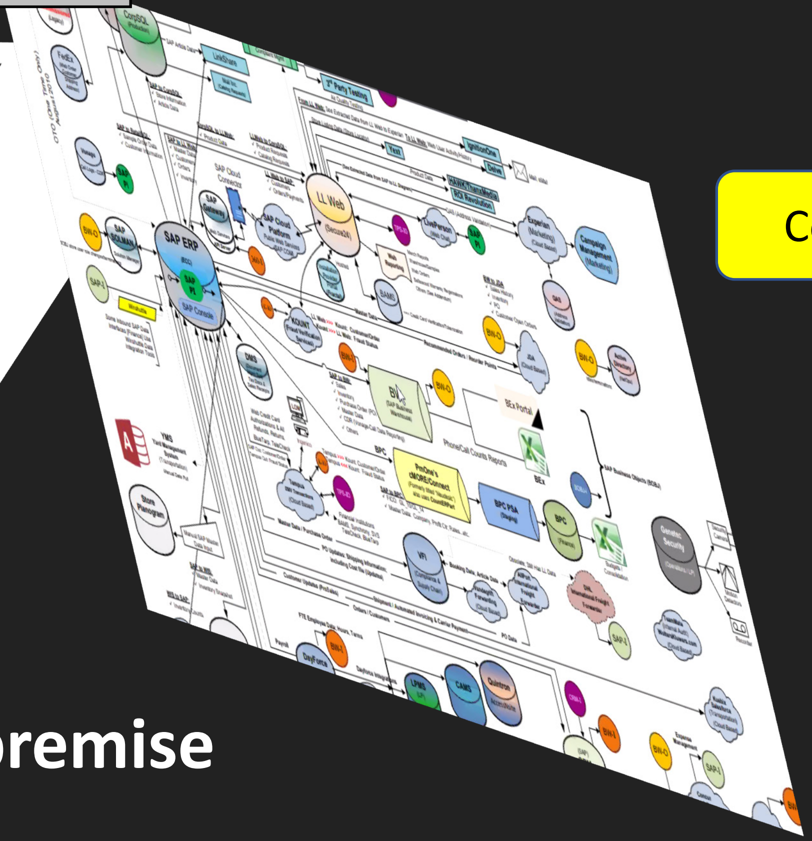
Policies

Controls



Active Directory , Load Balancers

- *WAF*
- *SIEM*
- *Firewall*
- *Encryption*
- *Tokenization*
- *Key Management*
- *AV – Anti Virus*
- *Network Sec*
- *And more*



On-premise

Organizations migrating workloads to the Cloud

	Biz Apps	Data Warehouses
Currently Migrating	41%	30%
Migrating in 12 mo	17%	15%
Migrating in 1-3 years	14%	15%
% of the market moved or moving with 3 years	72%	60%

The 2020 IDG Cloud Computing Survey

(Represents the 551 IT decision-makers)

Area	Timing	Focus	Comments Use case: Bank
Requirements	Short	Internal requirements	International regulations
Cloud	Short	Machine Learning	Start with basic ML training and inference on senstivie data in cloud
Competition	Short	Competitive advantage	ML and NLP-powered services can give banks a competitive edge
Data	Short	Encrypted data	Important
	Long	Synthetic data	Computing cost?
Analytics	Medium	AML / KYC	What are other Large banks doing?
	Short	Analytics	Initial focus
	Short	Operation on encrypted data	Computation on sensitive data to the cloud. Trade-offs with performance, protection and utility?
Industry	Short	Industry dialog	Working groups in standard bodies (ANSI X9, Cloud Security Alliance, Homomorphic Encryption Org)
Model	Short	Encrypted model	Important
Pilot	Short	Experimentation	What are other Large banks doing?
	Short	Scotia Bank case study	Query solution for AML / KYC
Proven	Medium	Fast follower	What are some proven solutions?
Quantum	Short	Homomorphic Encryption post-	Lattice-based cryptography is a promising post-quantum cryptography family, both in terms of foundational properties as well as its application to both traditional and homomorphic encryption
	Medium	Quantum	Plan for quantum safe algorithms
	Long	Quantum	Plan for quantum ML algorithms
Sharing	Short	Secure Multi-party Computing (SMPC)	Without revealing their own private inputs and outputs. Encrypted data and encryption keys never comingled while computation on the encrypted data is occurring or an encryption key is split into shares
Solutions	Short	Vendor positioning	Nonlinear ML regression needed? Linear Regression is one of the fundamental supervised-ML. Linear and non-linear credit scoring by combining logistic regression and support vector machines
	Short	Framework integration	Important
3rd party	Long	3rd party integration	Mining first
Training ML	Long	Federated learning	Complicated
	Long	TEE	Emerging