Removing Danger From Data

Kevin Scott, Senior Consultant @ CloverDX

Businesses face increasing risk around safety of data.

- **Data is increasingly taking central role in many businesses.**

… while at the same time

- **Government regulations around safeguarding practices are rapidly emerging and are being actively enforced.**

Rising Regulations

- **European Regulation – GDPR**

  Enforcement began in May 2018

  Fines already levied against Google, British Airways, Marriot

- **US Regulation status**

  No national regulation

  California CCPA regulation takes effect January 2020

  Microsoft will comply with CCPA nationwide

  Other states following California model

It’s not <u>only</u> PIIs that needs safeguarding

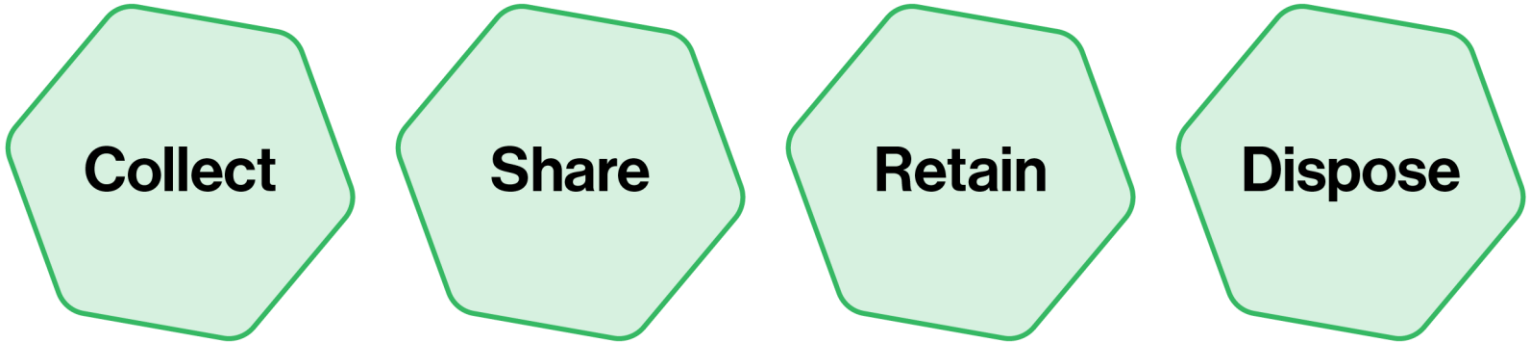
- **PII danger more obvious and includes potential legal costs**

  name, birthdate, SSN, account number

- **Non-PII data is also potentially dangerous.**

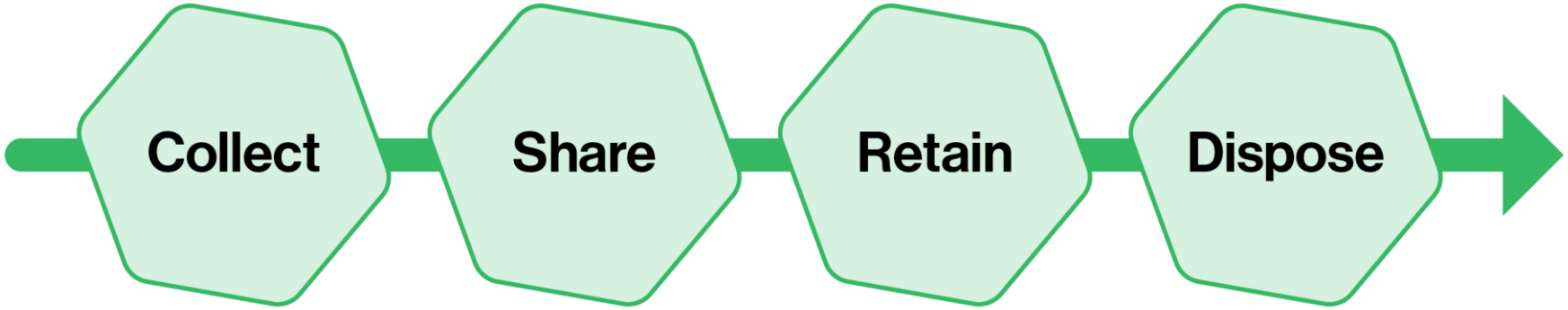
  sales forecasts, product plans, KPIs

Danger lurks in all phases of the Data Life Cycle



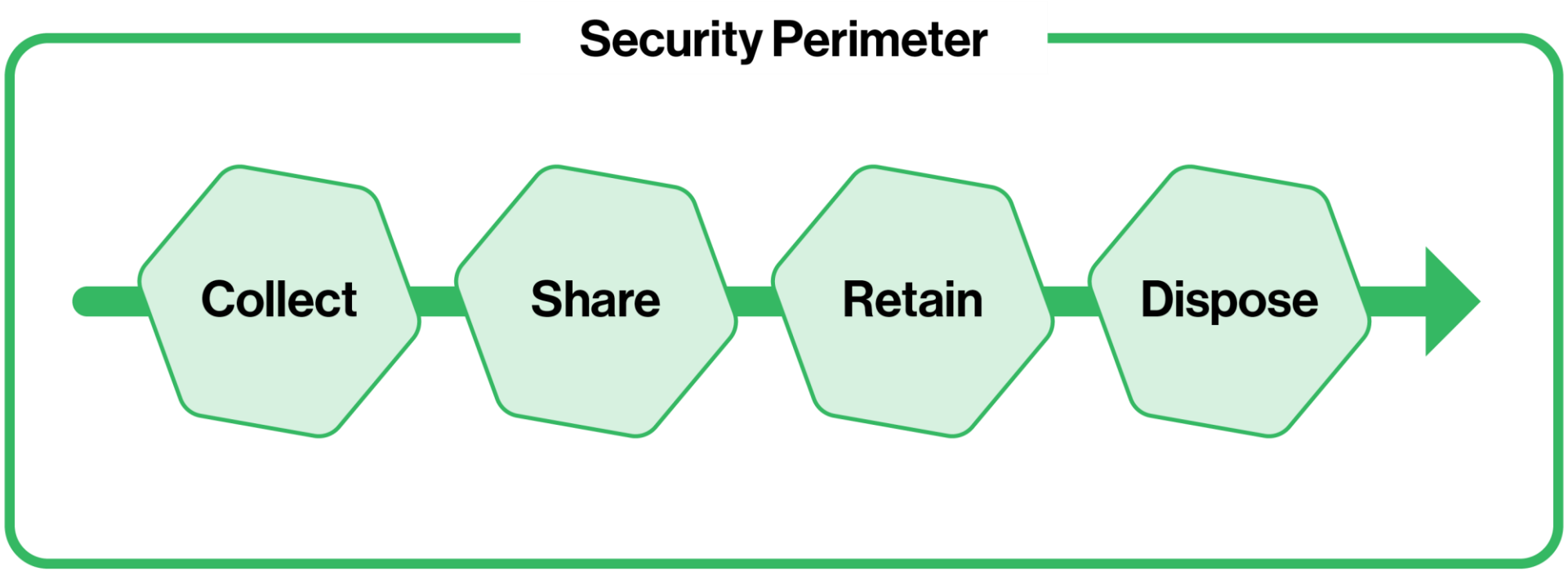
Danger lurks in all phases of the Data Life Cycle

Typically there's **processing** required to move data between stages

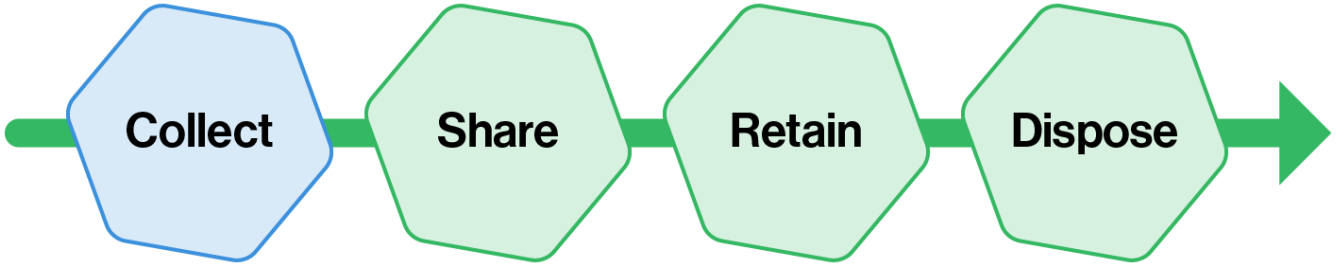


This **processing** can include **actions to remove danger**

Controlling danger from within



Collection Phase

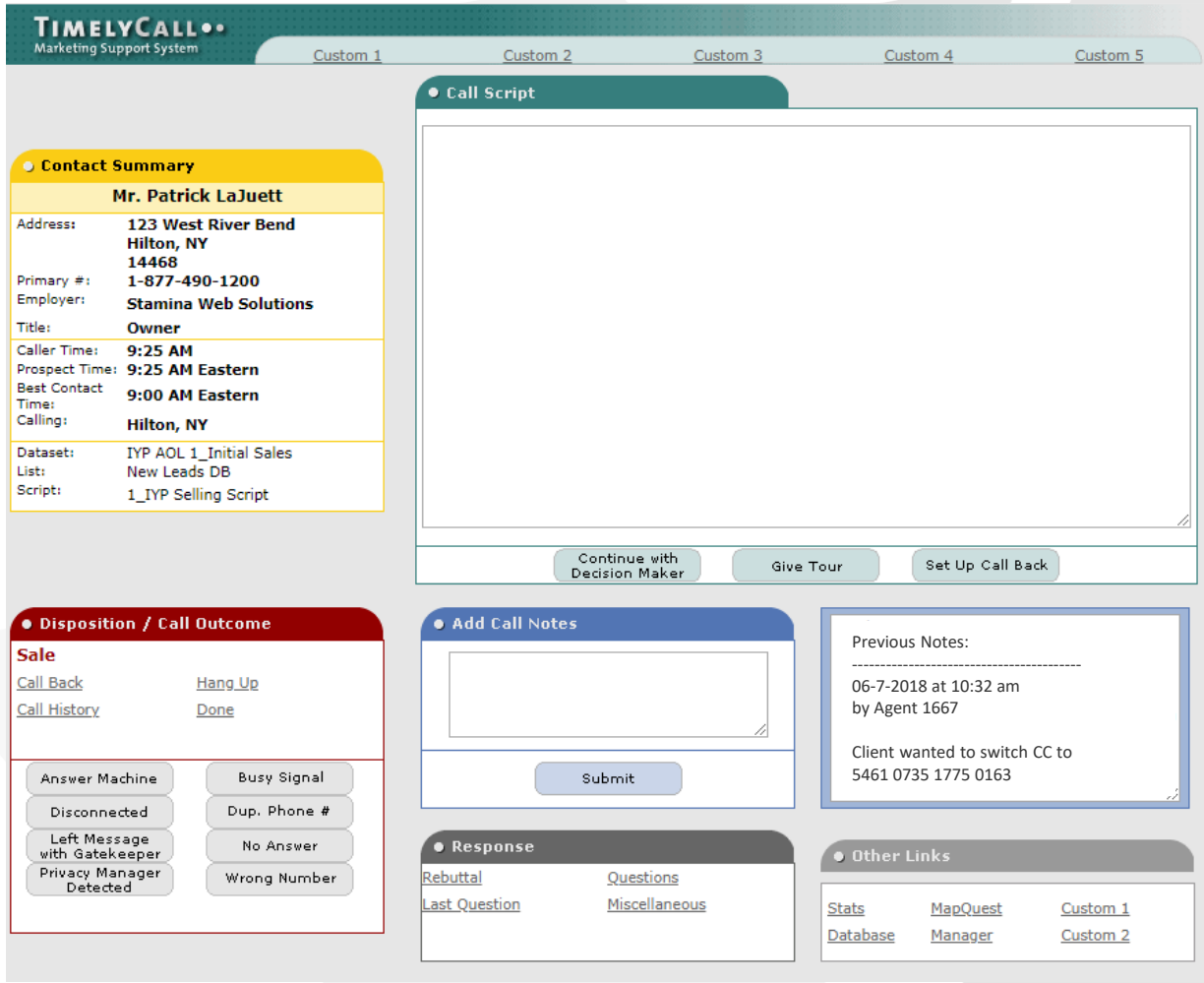


- Danger originates in collection phase
- Not knowing if PII data has been introduced into system

Collection Phase:

Look for dangerous data everywhere

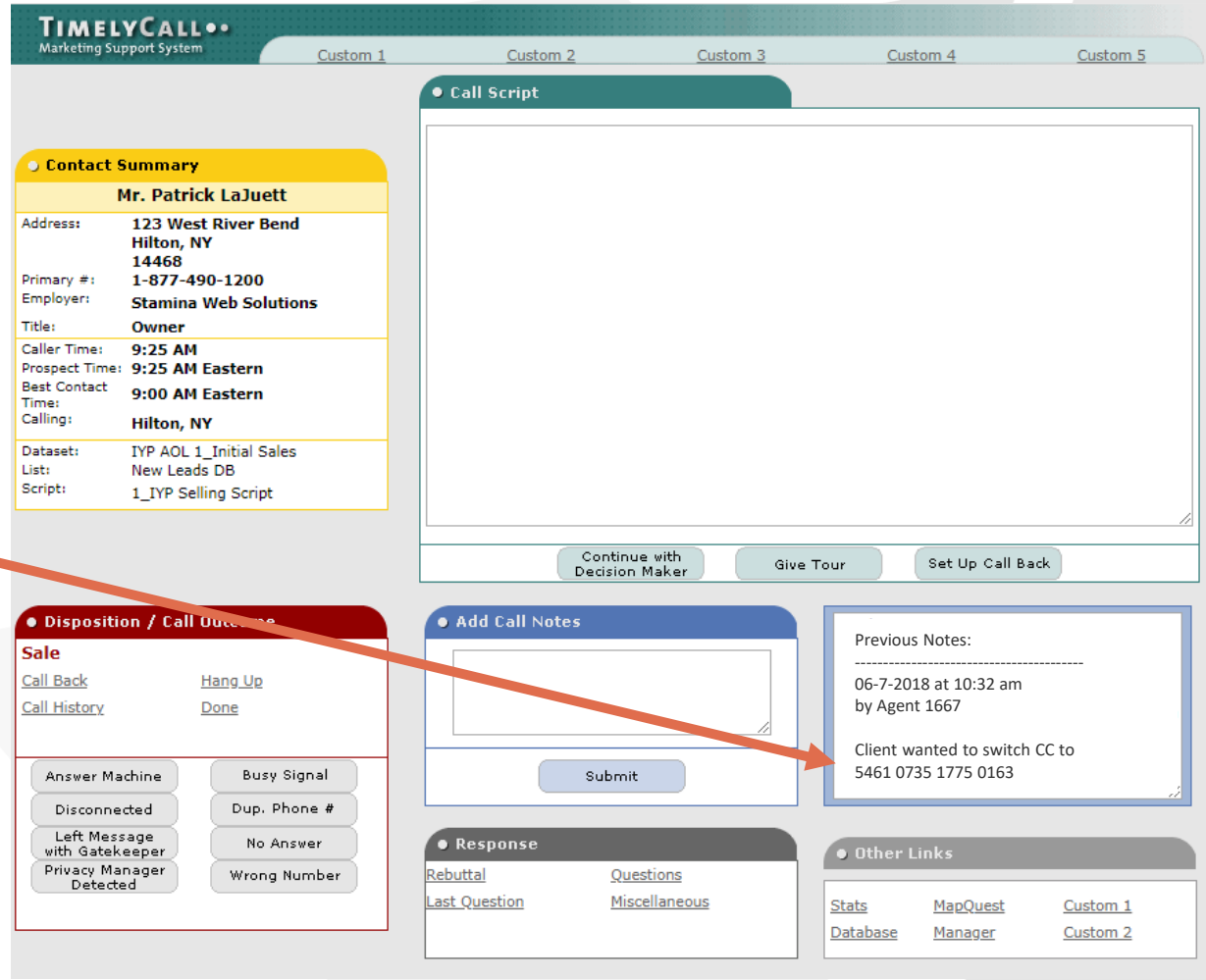
expected



Collection Phase:

Look for dangerous data everywhere

unexpected



Collection Phase

Remedies:

- Data minimization
- Choose ETL tools that provide input classification
- Consider Software to catalog and tag collected data

Collection Phase:

- **Data Classification tools can help track incoming sensitive data**

GDPR Check Tool

Pending deletions | Value search | John Doe (settings) | Logout

Home / Reports / Data Map

Data map - list of tables and related domains

Records per page 10 | Export

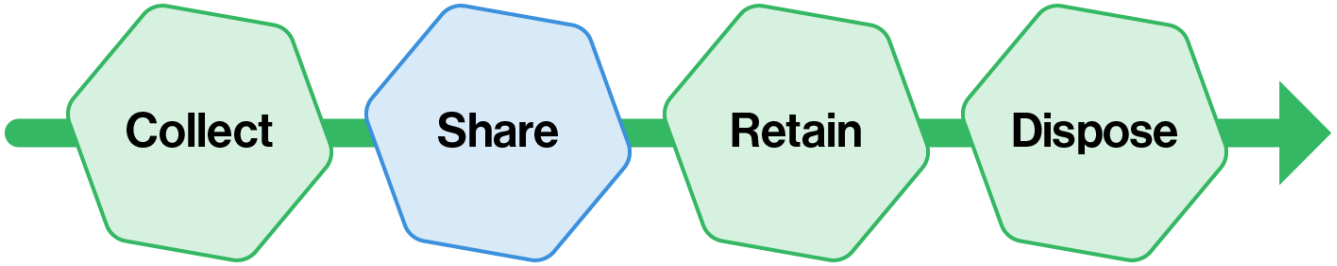
Database	Schema	Table	Column	Domains	Processes	Sample data	
							Remove filter
firmdb	erpp	ta_person	name	First name, Last name, Address - Street	P016, P101	alex, andrej, daniel ...	Actions
firmdb	erpp	ta_person	name_alt	Last name, First name	P092	alex, andrej, daniel ...	Actions
firmdb	erpp	ta_person	last_name	First name, Last name	P023, P101	adam, adamek, barton ...	Actions
firmdb	erpp	ta_purchord	last_name	Last name, Address - Street	P092, P101	bláha, borovička, cín ...	Actions
firmdb	erpp	ta_purchord	addrtxt	Address - Street, Address - City, Address - ZIP code	P016, P023	brno, bayerova, dlouhá ...	Actions
firmdb	erpp	ta_purchord	addrcontxt	Address - Street, Address - City, Address - ZIP code	P023	60200, brno, bezručova ...	Actions
firmdb	erpp	ta_sup_contact	cvalue	E-mail, Last name, First name, Address - City	P016, P023, P092, P101	aberka@firma.cz, adamec@jinafirma.cz, bures.jiri@oknaprovas.cz	Actions
firmdb	erpp	ta_supplier	city	Address - City	P016	beroun, brno, chrudim ...	Actions
firmdb	erpp	ta_supplier	txtDisplayName	First name, Last name, Address - City	P023, P092, P101	adam, adamec, beron ...	Actions
firmdb	erpp	ta_supplier	conperson	Last name, First name	P016, P023, P092, P101	andrej, ber, cilek ...	Actions

Showing **1 - 10** of **2160** results.

First Previous 1 2 3 4 5 ... 216 Next Last

Data Classification tags can be stored and accessed in a Data Catalog

Share phase:



- **Sharing usually requires transforming data assets so they can provide desired business value.**
  searching, mapping, sorting, merging categorizing, analyzing, reshaping.

- **Primary danger is oversharing**
  Sharing too much data
  Sharing to an unnecessarily large audience
  Accidental sharing

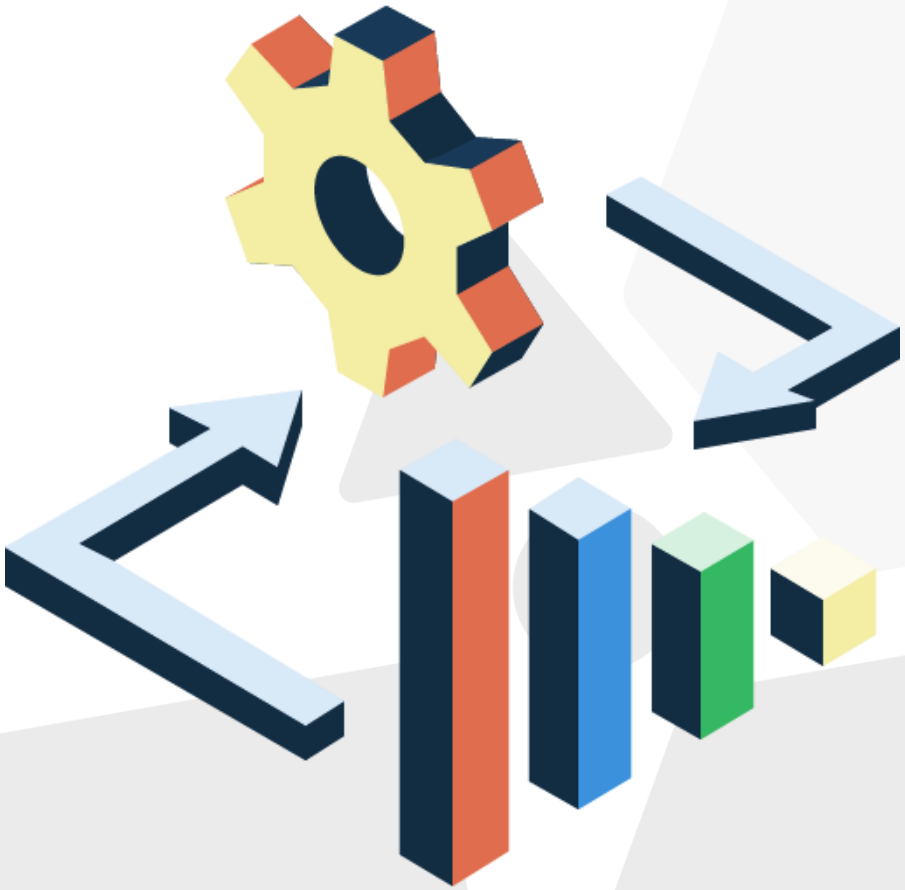
Share phase:

Processing techniques available to deliver business value without unnecessarily sharing sensitive data.

- **Anonymization**
- **Privacy preserving analytics**

Anonymization

Process of obfuscating data, so they keep some of their original attributes but not to extent they could be used to infer relation to real people or entities.



Anonymization – Shuffle

First Name	Surname	Age
Alice	Smith	42
Bob	Johnson	21
Dave	Doe	74
Eve	Jackson	44
Grace	Chang	32

First Name	Surname	Age
Alice	Doe	42
Bob	Jackson	21
Dave	Chang	74
Eve	Smith	44
Grace	Johnson	32

Anonymization – Jitter

Birth Date	Gender	Education Level		Birth Date	Gender	Education Level
7-Jul-82	Male	High School	+3 days	10-Jul-82	Male	High School
18-May-85	Male	High School	-1 day	17-May-85	Male	High School
5-Mar-87	Female	Bachelor	+3 days	8-Mar-87	Female	Bachelor
17-Jun-97	Male	Associate	-1 day	16-Jun-97	Male	Associate
27-Feb-82	Male	Graduate	-3 days	24-Feb-82	Male	Graduate
9-Nov-58	Male	Graduate	+1 day	10-Nov-58	Male	Graduate

Anonymization – Masking

Naively masked `4024 XXXX XXXX XXXX`

Intelligently masked `4024 0071 4314 0399`

Keeping Issuer code
VISA Credit card
Issued by Bank of America

Randomized
Account Number

Valid Luhn checksum

Preserves card types, issuers, preserves validity

Anonymization – Aggregation/Generalization

Birth Date	Gender	Education Level
7-Jul-82	Male	High School
18-May-85	Male	High School
5-Mar-87	Female	Bachelor
17-Jun-97	Male	Associate
27-Feb-82	Male	Graduate
9-Nov-58	Male	Graduate

Age	Gender	Education Level
30's	Male	High School
30's	Male	High School
30's	Female	Bachelor
20's	Male	Associate
30's	Male	Graduate
60's	Male	Graduate

Privacy Preserving Analytics

- **Methods developed at Linked-In for data mining on very large (web-scale) result sets**
- **Provide accurate analytics without inadvertently identifying individual users.**

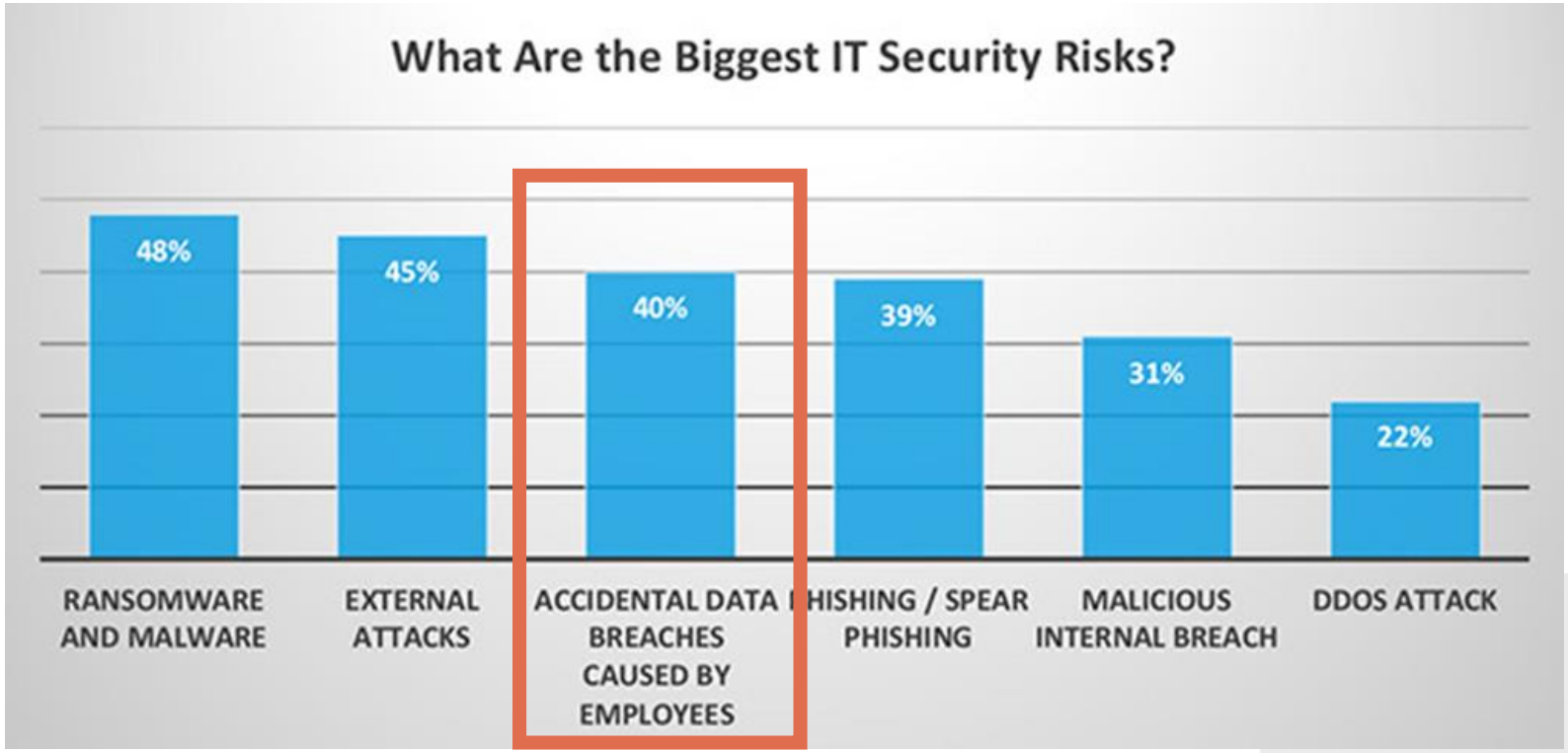
Inject small amounts of random noise to query/search

Offset noise with more processing to maintain data consistency

See [Privacy-Preserving Analytics and Reporting at LinkedIn]

Share Phase

- Employees simply doing their job can be dangerous.

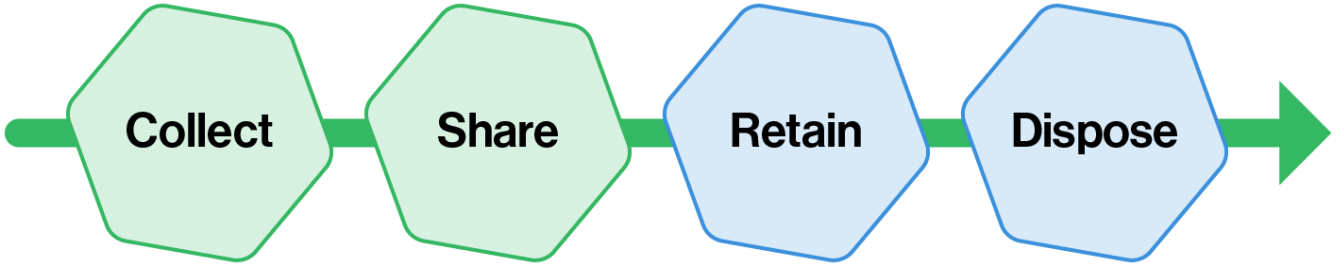


Share Phase:

- 5 most common technologies that lead to accidental data breaches by employees:

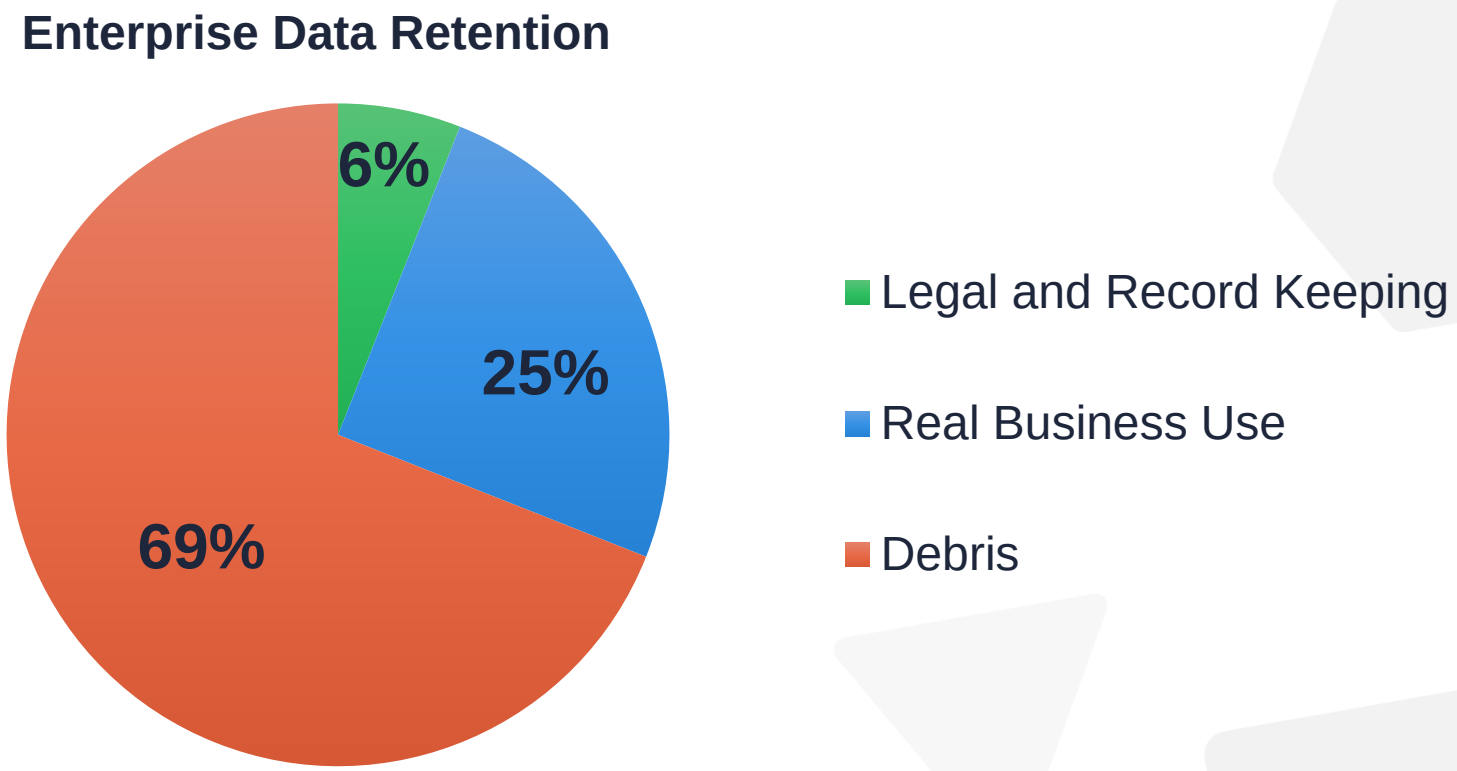
1. External email services (Gmail, Yahoo!, etc.) (51 percent)
2. Corporate email (46 percent)
3. File sharing services (FTP sites, etc.) (40 percent)
4. Collaboration tools (Slack, Dropbox, etc.) (38 percent)
5. SMS / messaging apps (G-Chat, WhatsApp, etc.) (35 percent)

Retain and Dispose Phases



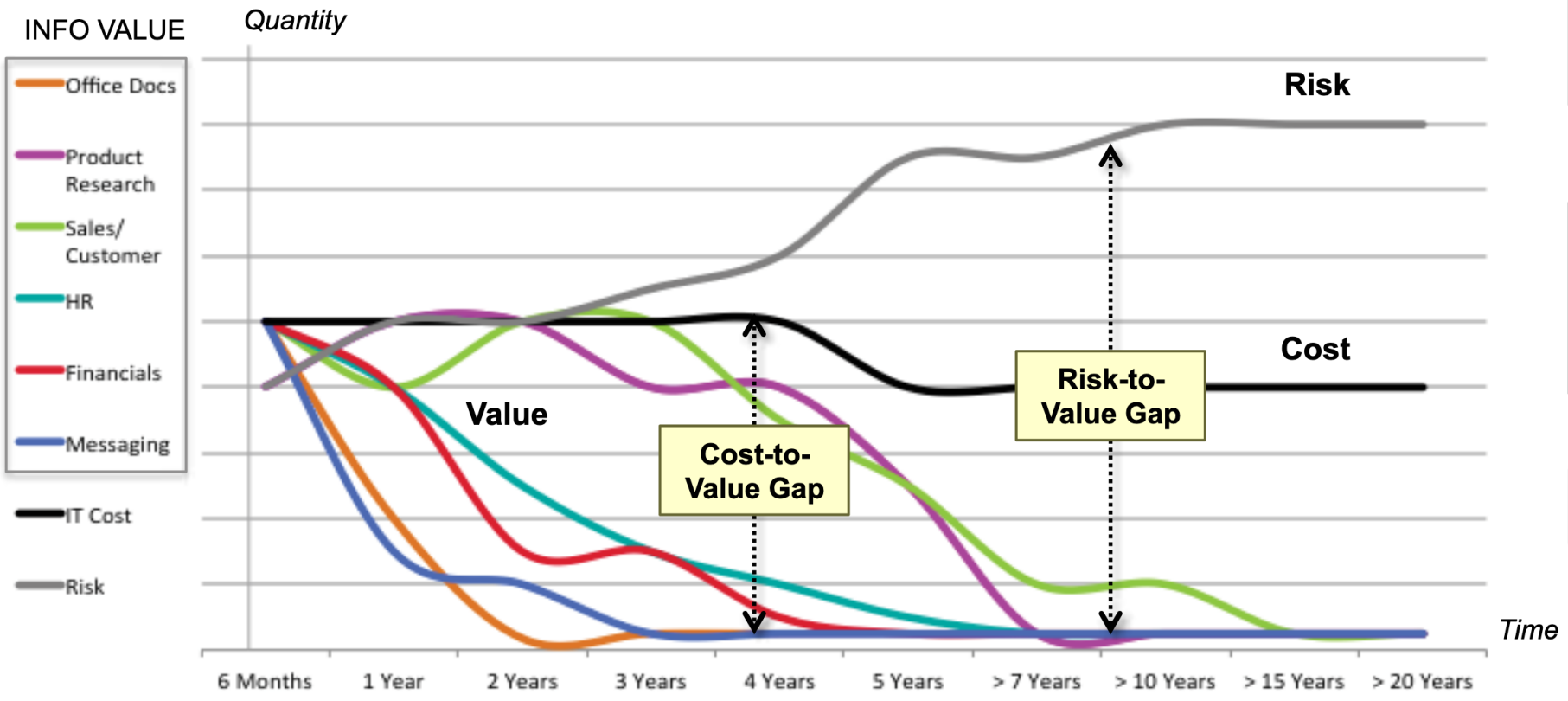
- Minimize time data is kept
- A Data Catalog can also house retention and disposal status

Most enterprise data is retained indiscriminately



https://cedar.princeton.edu/sites/cedar/files/media/information_lifecycle_governance.pdf

Information Value Declines Over Time, Costs and Risks of retention do not.



https://cedar.princeton.edu/sites/cedar/files/media/information_lifecycle_governance.pdf

Retain and Dispose Phases

In 2019 CapitalOne Breach exposed sensitive data that included *rejected* credit card applications from as far back as *2005*.

Remediation expected to cost $200-300 million *without any explicit regulatory fines*.

Retain and Dispose Phases

GDPR Check Tool

Pending deletions | Value search | John Doe (settings) | Logout

Home / Reports / Pending deletions

Retentions

Records per page 5 | Export

Search for a person: Michael Jordan

Showing retention records for name: **Michael Jordan**

System	State	Applicable date	Comment	
				Remove filter
SalesForce CRM	Deleted	2018-01-07	Deletion requested on 20th June 2017	Actions
AR Data	Pending deletion	2020-04-12	Legal requirement - invoices. Deletion requested on 2017-12-29 June 2017	Actions
ERPSource	Pending deletion	2018-01-07	Legal requirement - invoices. Deletion requested on 2017-12-12 June 2017	Actions
SugarCRM	Deleted	2018-01-07	Deletion requested on 12th December 2017	Actions
Sales Tools	Deleted	2018-01-07	Deletion requested on 12th December 2017	Actions

Showing **1 - 5** of **12** results.

First | Previous | 1 | 2 | 3 | Next | Last

Summary

- **Emerging Regulations are increasing danger in data**
- **Removing Danger = Identification + Remediation**

  Realizing the dangers exist
  Software tools exist to help

- **Develop Risk Intelligence**

  Size of risk vs cost of remedy

  Implement reasonable precautions

CloverDX

Thank you

About CloverDX Data Integration Platform

CloverDX is a data integration platform for designing, automating and operating data jobs at scale. We've engineered CloverDX to solve complex data movement and transformation scenarios with a combination of visual IDE for data jobs, flexibility of coding and extensible automation and orchestration features.

www.cloverdx.com

References

- **Data minimization**
  https://www.dataguise.com/gdpr-compliance-data-minimization-use-purpose/
- **Regulation Costs**
  https://www.cnbc.com/2019/07/10/gdpr-fines-vs-marriott-british-air-are-a-warning-for-google-facebook.html
- **Capital One (Data EOL)**
  https://www.theverge.com/2019/7/31/20748886/capital-one-breach-hack-thompson-security-data
  https://www.nytimes.com/2019/07/29/business/capital-one-data-breach-hacked.html
- **Princeton Lifecycle Governance**
  https://cedar.princeton.edu/sites/cedar/files/media/information_lifecycle_governance.pdf