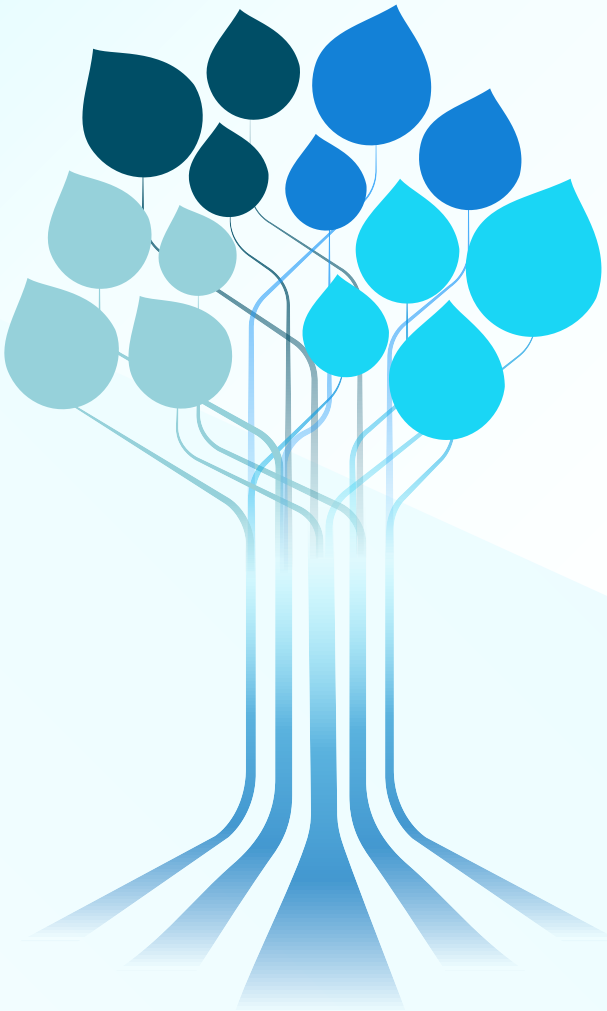
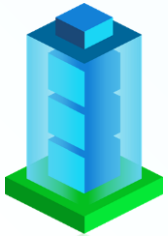
From DevOps to AI:

Building a data strategy to unlock the potential of rapid, secure data delivery



“The world is changing very fast. Big will not beat small anymore. It will be the fast beating the slow.”
– Rupert Murdoch

Delphix DevOps Data Platform



Data
Governance

Data Control Tower

Data
Compliance

Masking

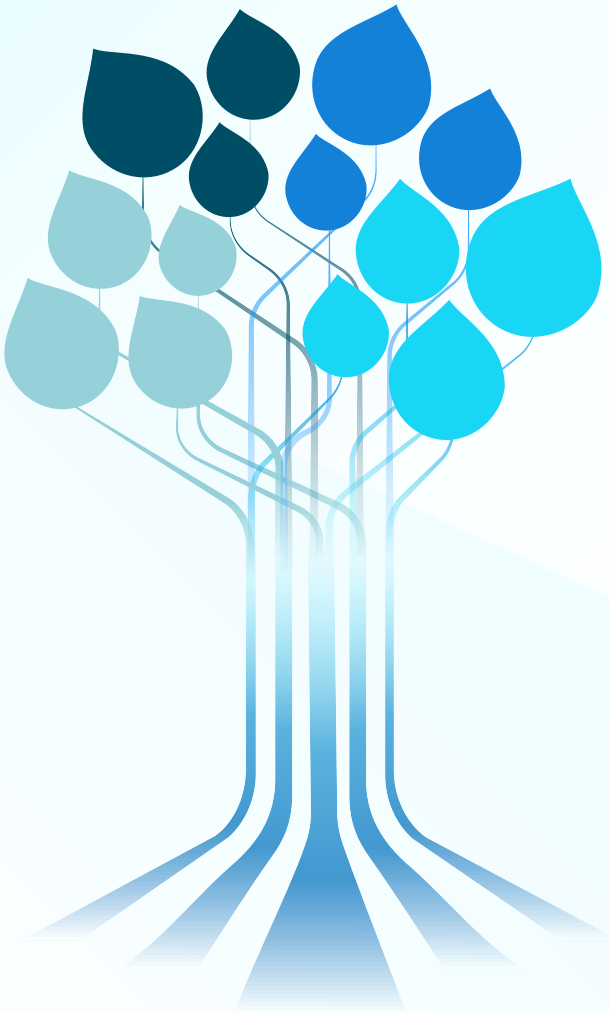
Data
Automation

Virtualization

ORACLE E-BUSINESS SUITE · ORACLE DATABASE · ORACLE SECURE BACKUP · SAP SAP IQ · SAP HANA · SAP ASE · ORACLE CLOUD · Microsoft Azure

SQL Server · salesforce · IBM i · z/OS · IBM DB2 · MariaDB · MySQL · PostgreSQL · mongoDB. · IBM Cloud · Google Cloud · aws

The DevOps Data Challenge



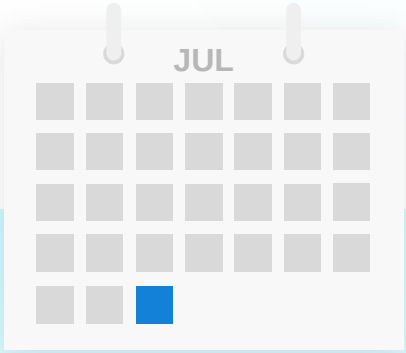
Slow to Fast Application Releases

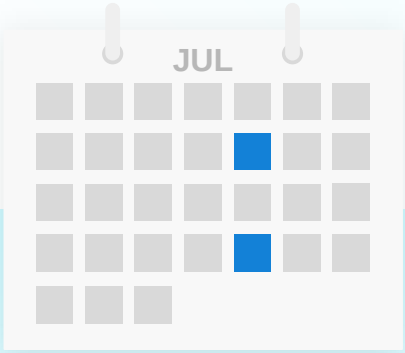
Waterfall
Monthly releases

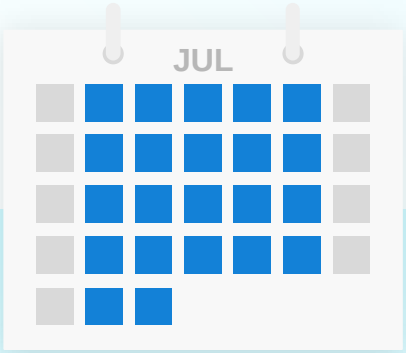
Agile
Biweekly releases

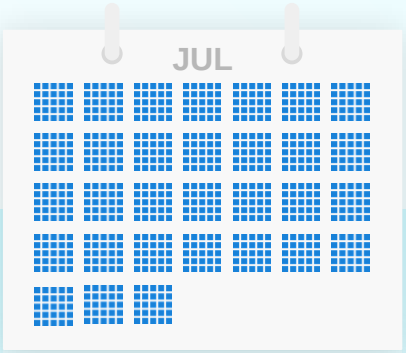
DevOps
Daily releases

CI/CD
Millions of CI/CD runs per month

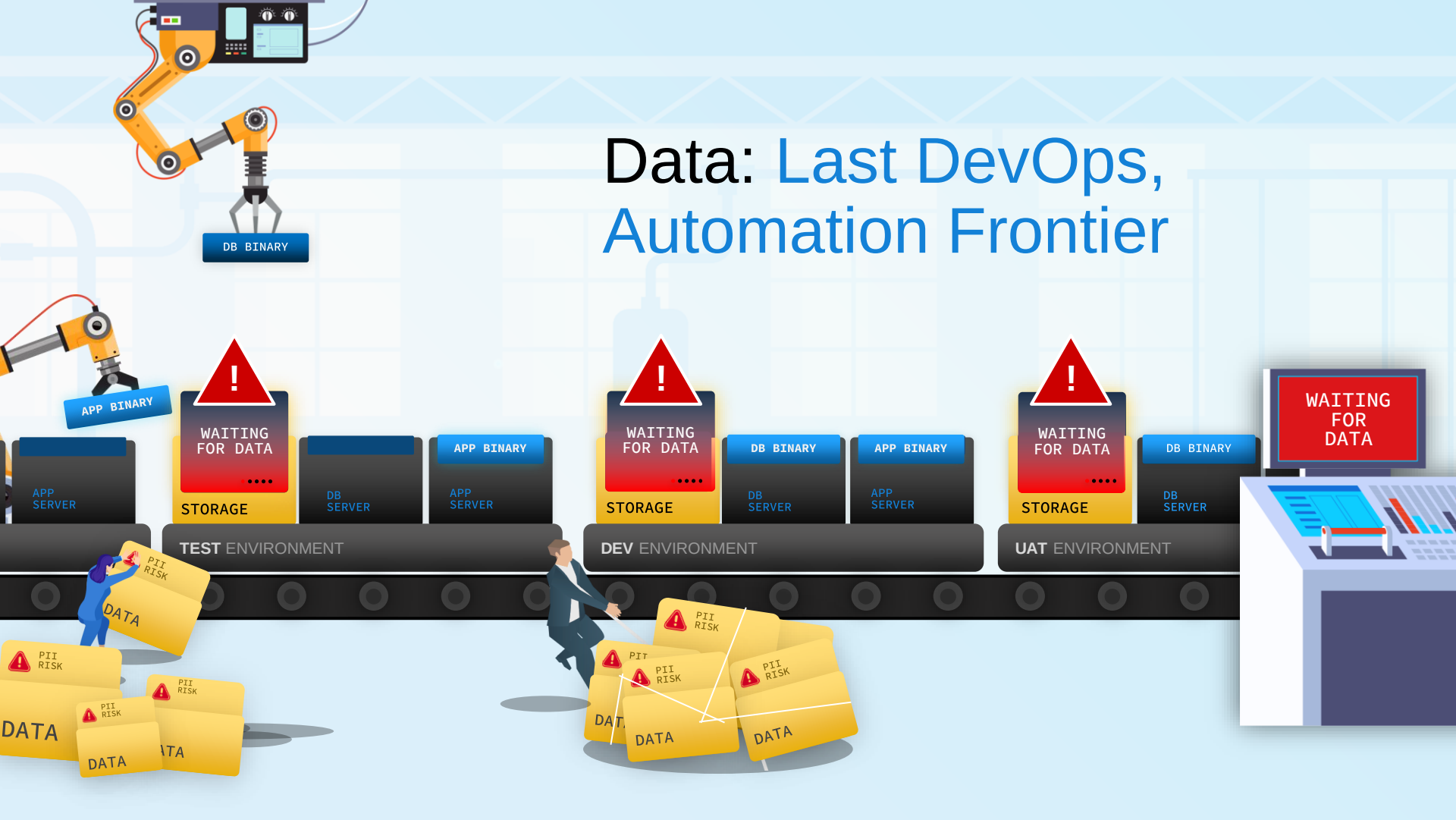




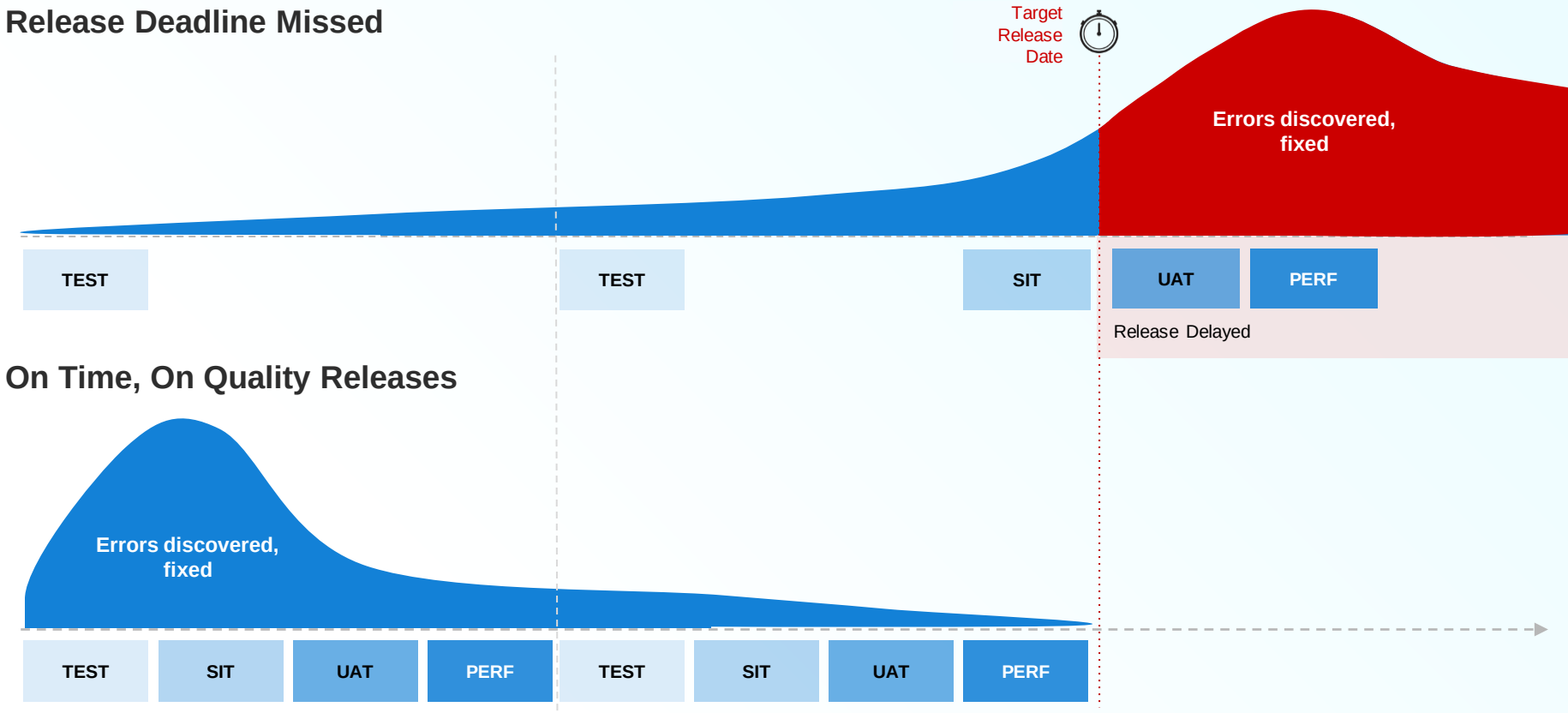




Data: Last DevOps, Automation Frontier



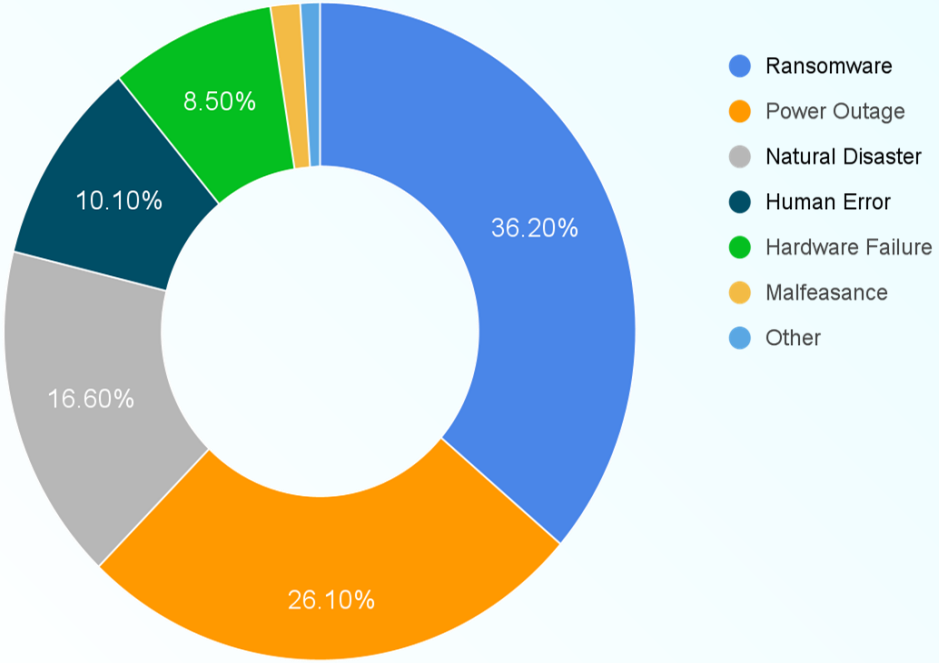
Lack of Data-Ready Environments Delays Testing, Releases



Most Common DR Events

$1.4M*

Average total cost of a Ransomware breach in Australia.



*Sophos report survey data of 5,400 IT managers for "The State Of Ransomware 2022". Average cost was for Australian companies, reported as US$1.01M.

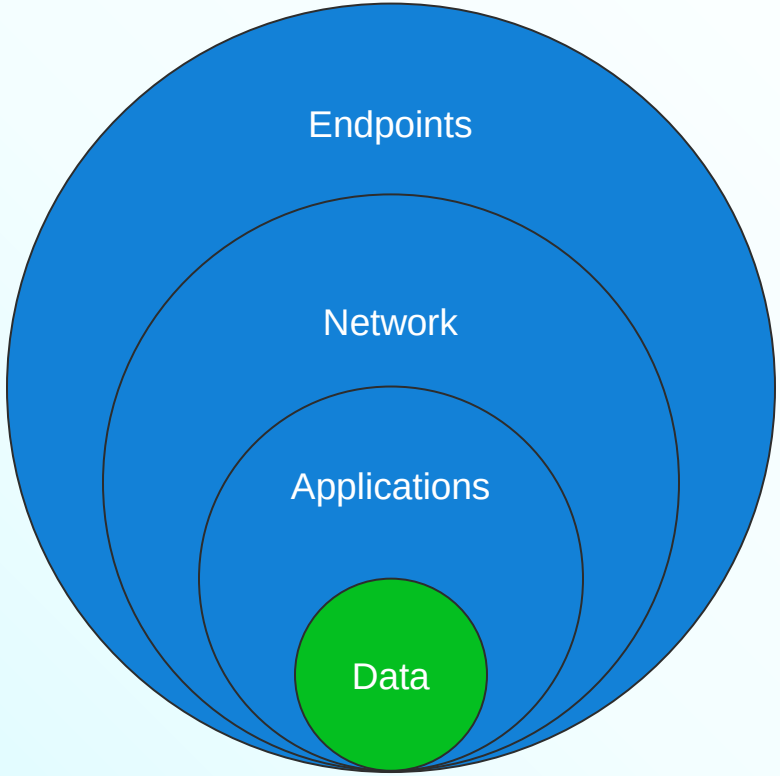
Source: Datrium Survey - The State of Enterprise Data Resiliency and Disaster Recovery

Criminals want your ***data***

We're now truly in the era of ransomware as pure extortion without the encryption

Why Screw around with cryptography and keys when just stealing the info is good enough

Jessica Lyons Hardcastle
Sat 25 June 2022



Real data in non-production is a liability

Cost of Bad **Test Data Management**

77M Customer Records Exposed

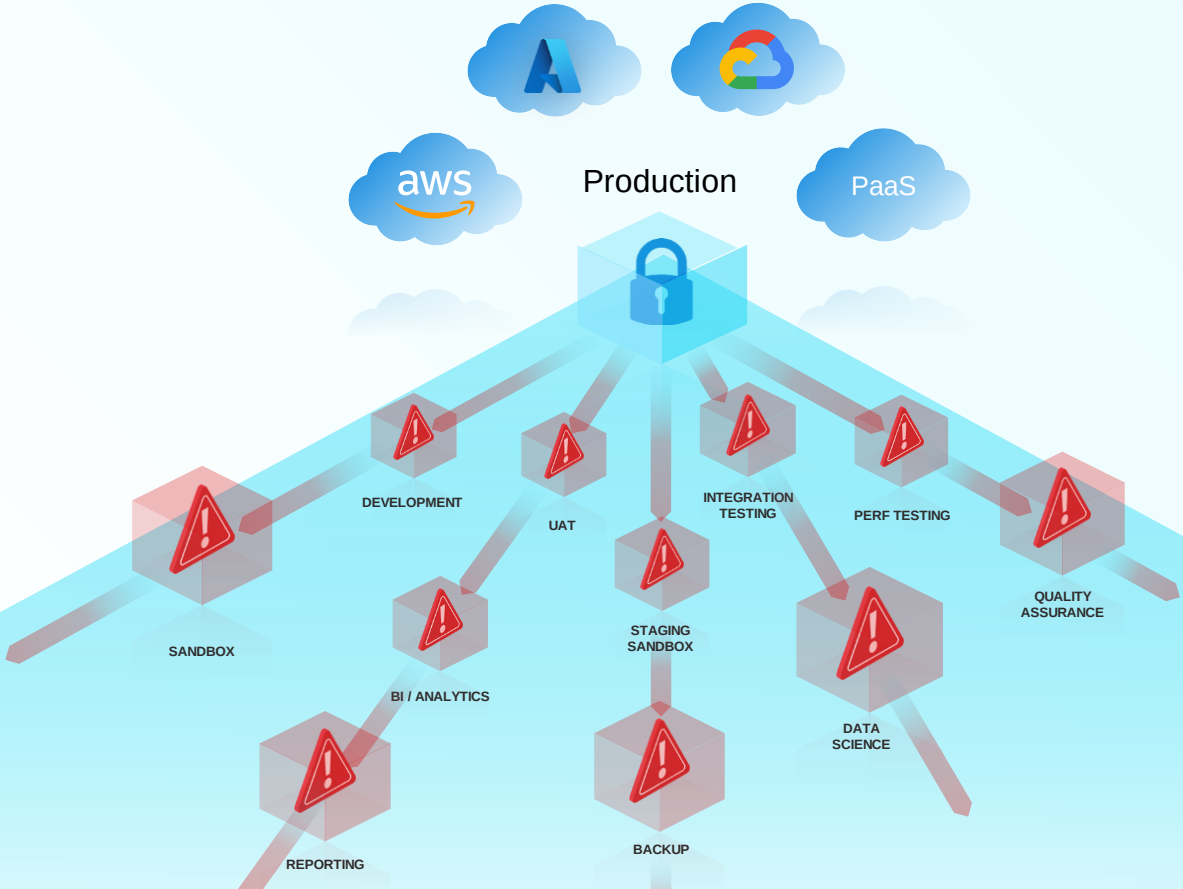
$150M Remediation

$350M CCPA Settlement

"The bad actor leveraged their knowledge of technical systems, along with specialized tools and capabilities, to **gain access to our testing environments.**"

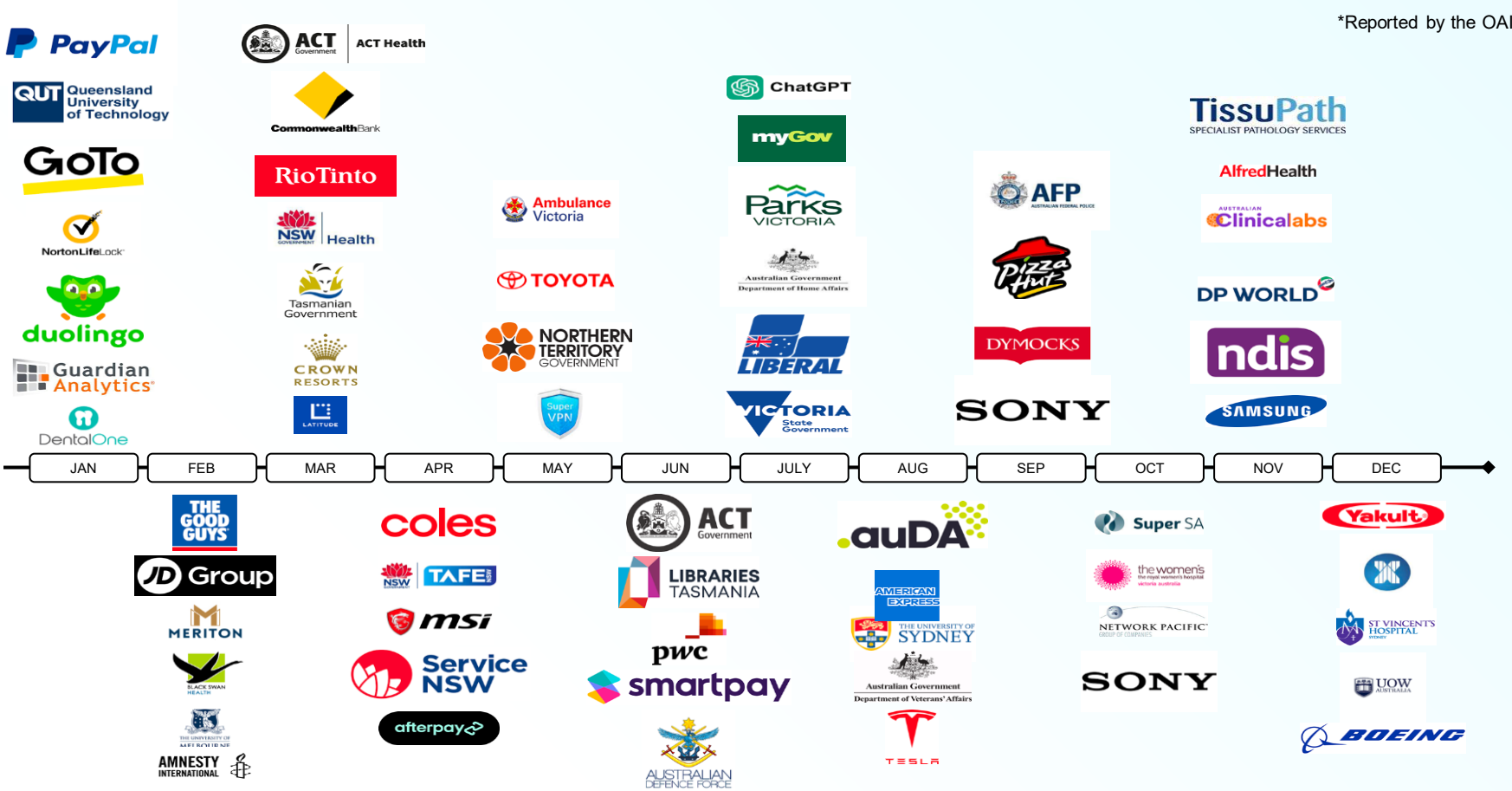
T-Mobile

— Mike Sievert, T-Mobile CEO



Notable Australian Breaches in 2023*

*Reported by the OAIC



The Real Cost of Synthetics and Subsets

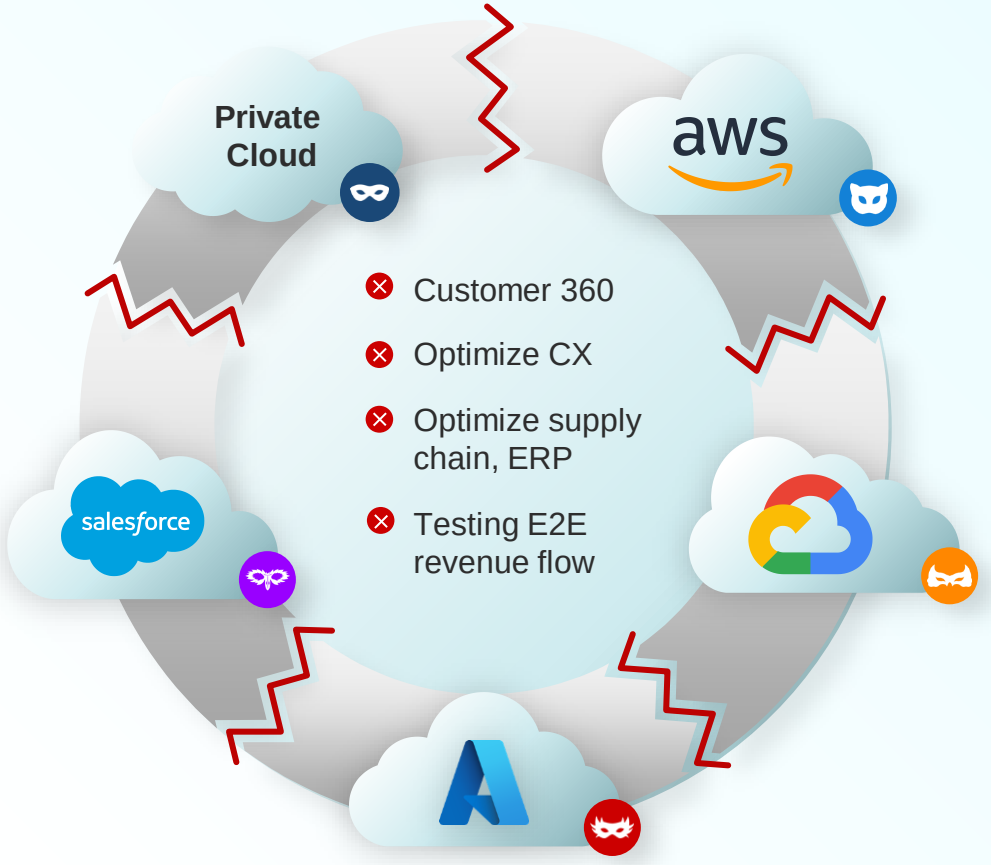
Synthetic Data

- **Ongoing, and Rising Investment**: Requires either a large/ongoing AI investment, or a growing investment in rule writing & maintenance.
- **Integration**: Typically, must still integrate to non-Synthetic datasets, which can be expensive to do and poses a large re-identification risk.
- **Relies on Fallible Human Semiotics**: Humans are still the arbiter of what/how things are synthesized, missing some relationships or sensitive data.
- **Not a Data Protection Silver Bullet**: Subject to Leakage (training data bias, membership inference attack (MIA), metadata bias, inferences from preserving raw distributions, adversarial GaN attacks, et al).

Subset Data

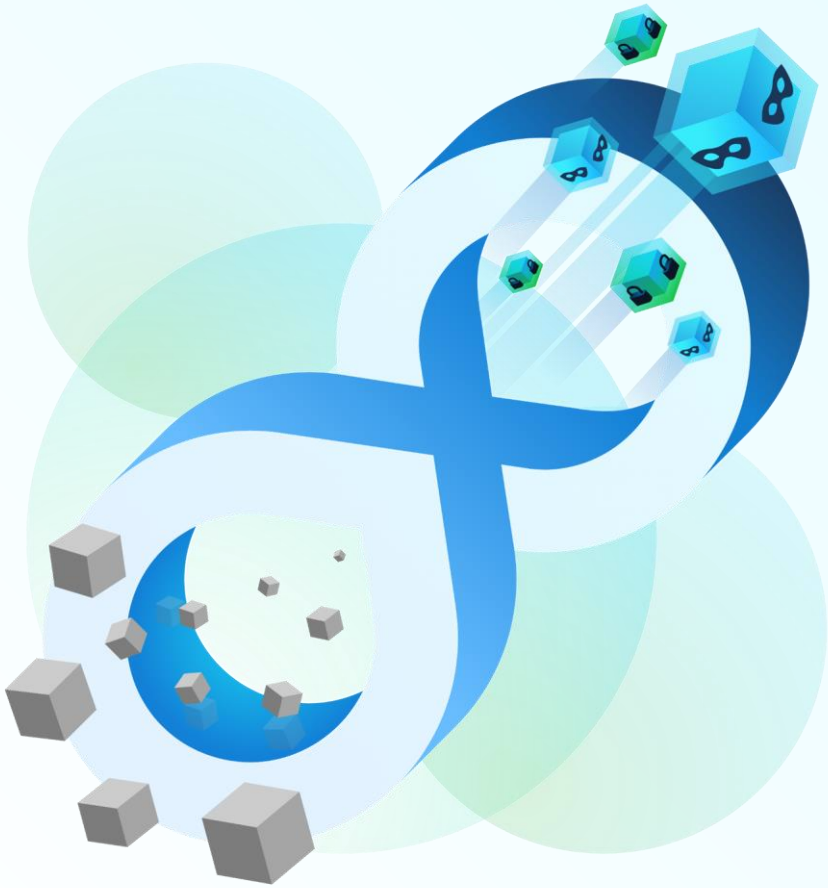
- **Ongoing, and Rising Investment**: Significant and ongoing investment in rule writing & maintenance. Humans spend as much time writing and maintaining as they did building the model in the first place.
- **Incompleteness**: Subsets often miss corner cases, force Type I/II errors, & fail to test basic performance cases.

Compliance in silos blocks innovation



Rapid, Secure Data Delivery

Delivering useful data where and when it's needed



Enterprise Approach to Data Masking

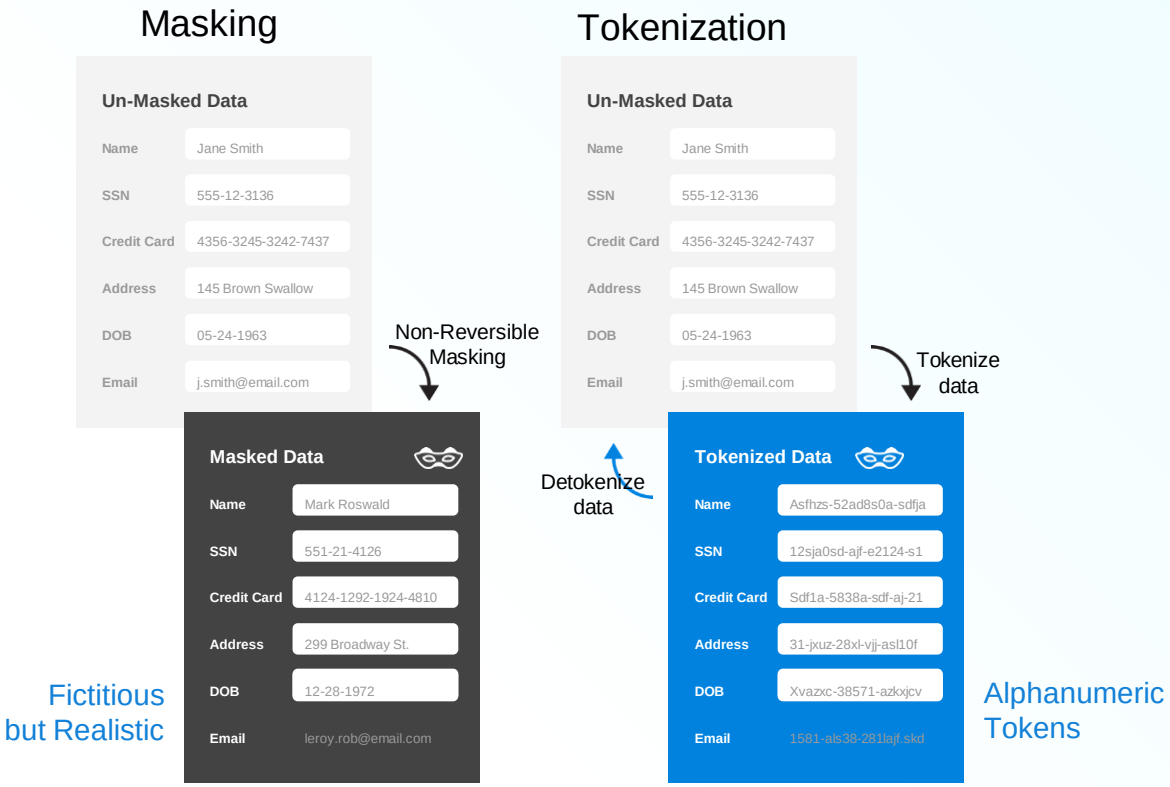
Building secure data assets

INTELLIGENT & USABLE
Keep data in a usable state with no impact to developers, testers, trainers, even vendors.

CONSISTENT
Masked data with consistency to preserve referential integrity

COMPREHENSIVE
Provide irreversibly masked data across a wide range of cloud or on-prem data sources

Intelligent and Useful



Transform sensitive data to comply with privacy regulations in two ways:

- Irreversibly mask data for non-production environments
- Tokenize data to enable teams to reverse transformation

Key benefits:

- Single solution for both masking and tokenization
- Masking completely and irreversibly neutralizes compliance risks in non-production environments
- Tokenization enables use cases requiring secure collaboration with third parties

Consistent

PLM

ERP

PROD

Contacts	
cust_id	last_name
150	Lee
151	Rogers
152	Johnson

Products	
prod_id	activation_code
A12	Alpha
B23	Beta
D63	Delta

Orders			
prod_id	quant	last_name	activation_code
X23	5	Lee	Gamma
A42	15	Jones	Charlie
D63	11	Dach	Delta

- **Transform sensitive data values**
- **Preserve referential integrity**
- **Use real, masked data for tests**

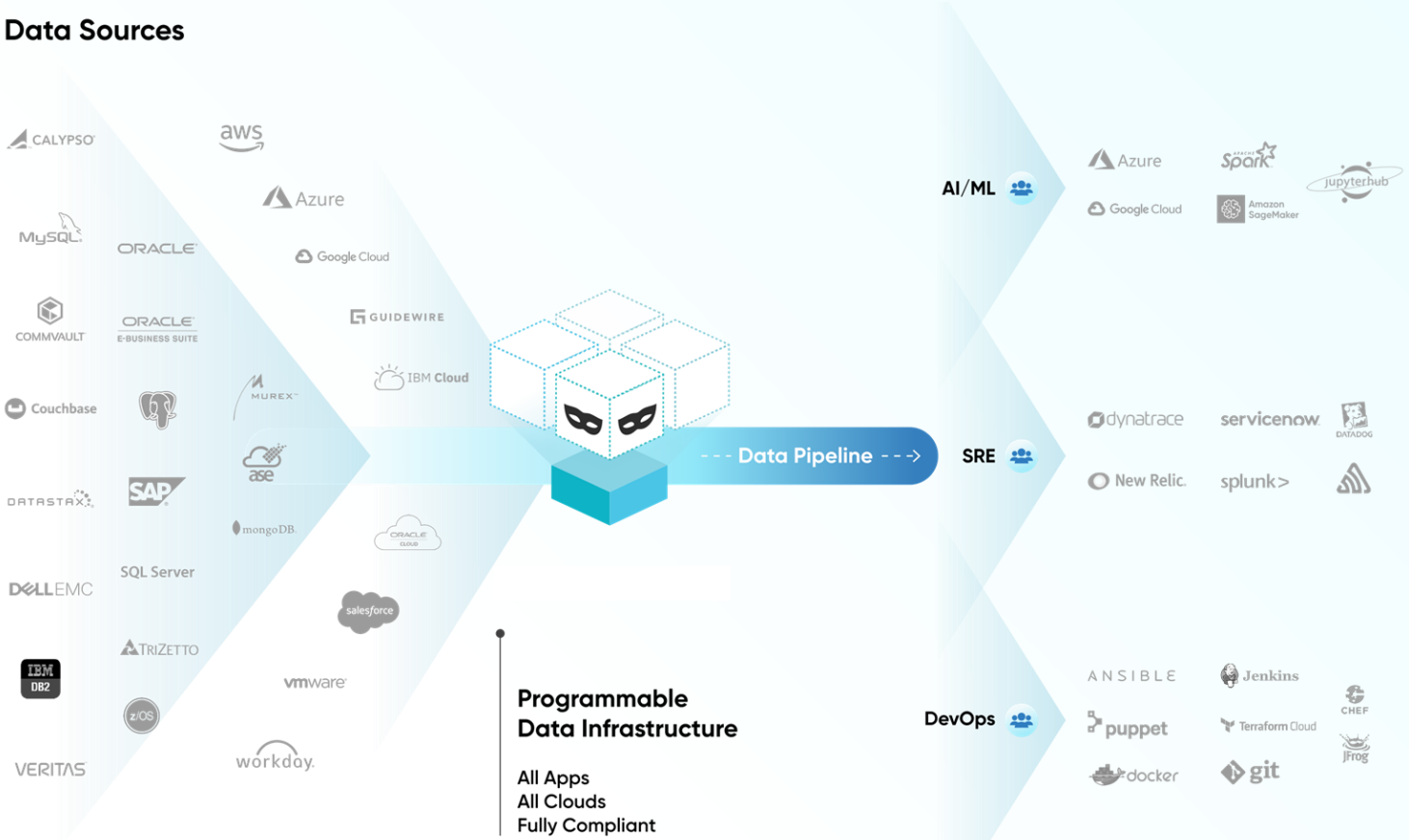
TEST

Contacts	
cust_id	last_name
150	Yang
151	Williams
152	Haas

Products	
prod_id	activation_code
A12	Xray
B23	Yankee
D63	Zulu

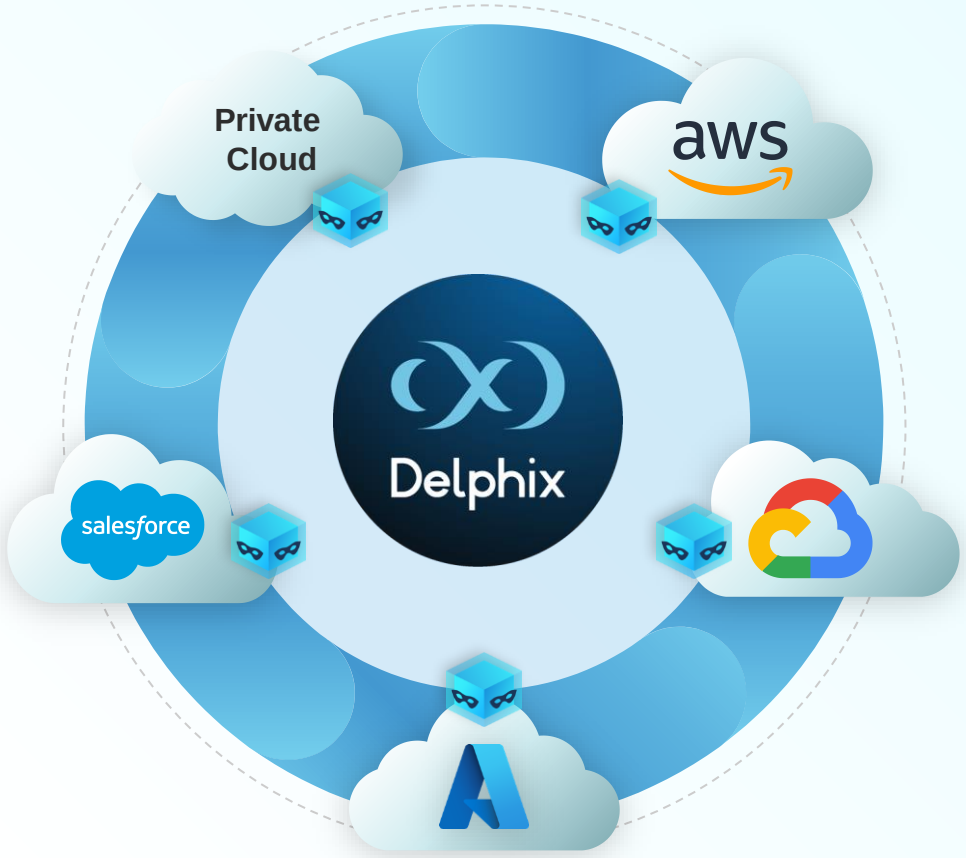
Orders			
prod_id	quant	last_name	activation_code
X23	5	Yang	Helo
A42	15	Williams	Echo
D63	11	Lin	Zulu

Comprehensive



Unified Compliance

Multicloud Referential Integrity Enables Unified Development, Analytics



Enterprise Approach to Data Masking Lifecycle

MANAGE POLICY

DEFINE DATA RISK POLICY

1. What is sensitive?
2. How to identify what is sensitive?
3. How to protect what is sensitive?

DISCOVER

IDENTIFY DATA RISK

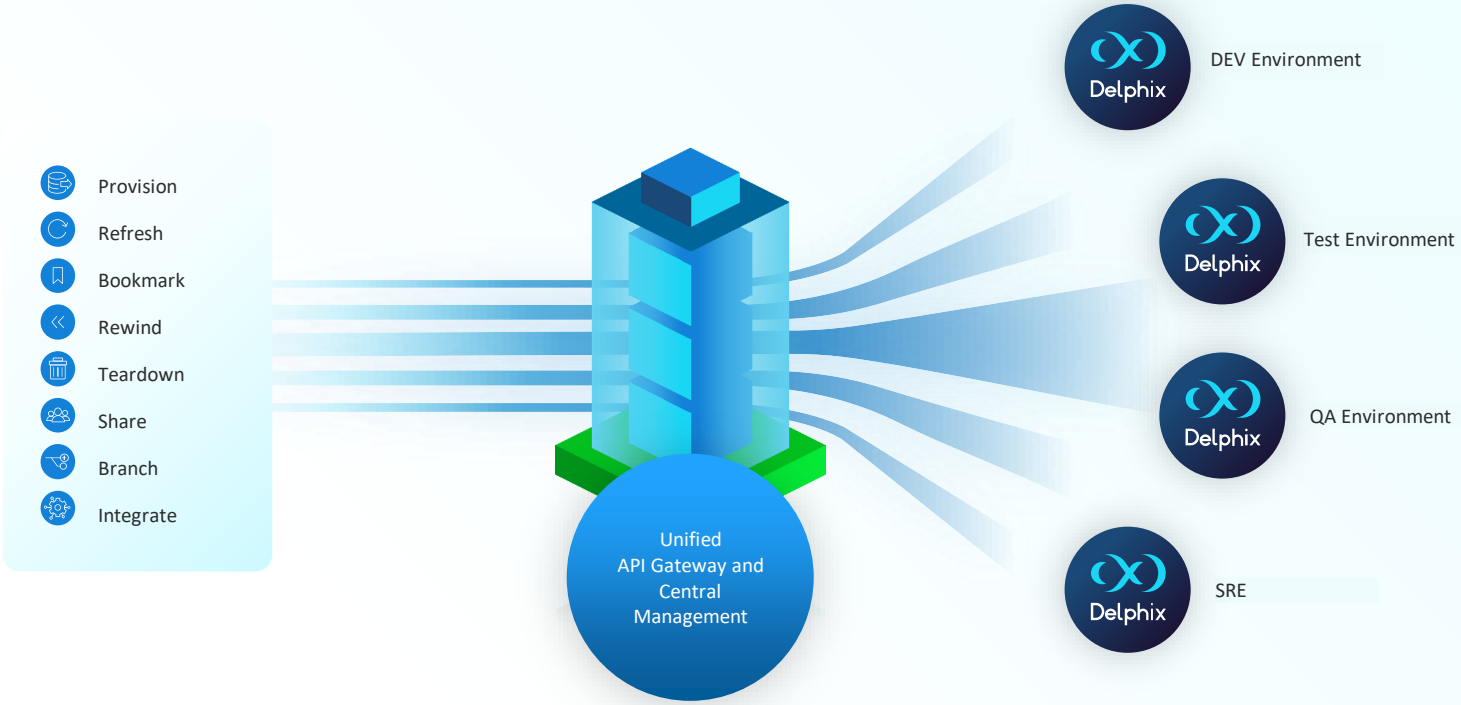
1. Identify what is sensitive based on policy.
2. Establish automatic consistency/integrity.

SECURE

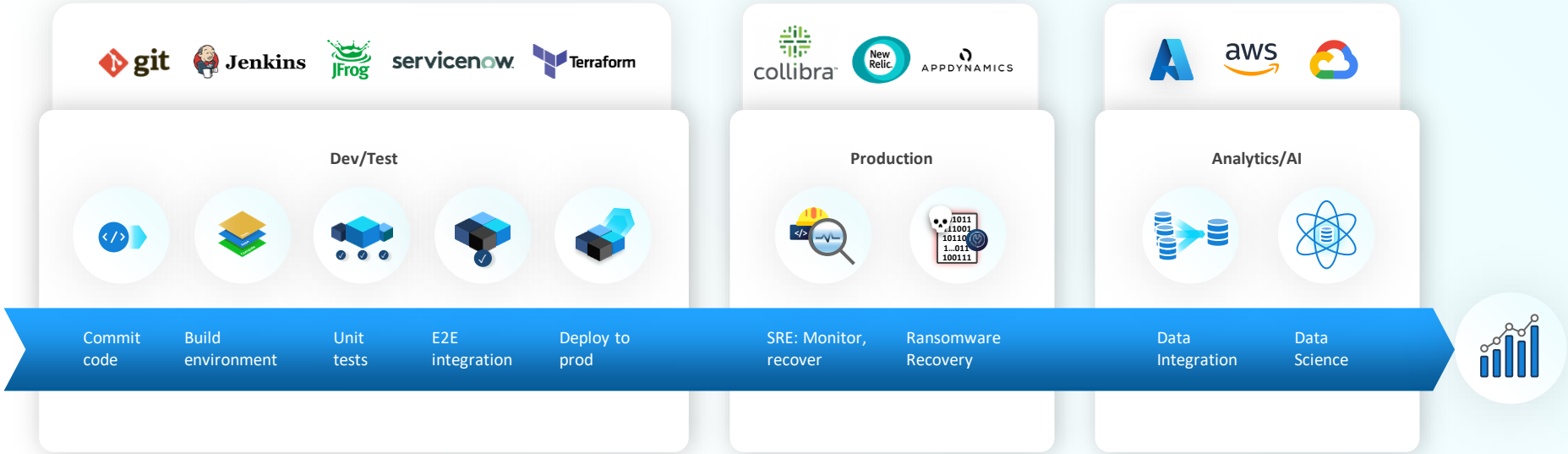
MANAGE DATA RISK

1. Consistently mask/de-identify data based on policy
2. Structured/Unstructured
3. Databases, Mainframe, Files
4. Across data centers
5. On-Premise, AWS, Azure

Automate: Ephemeral Data for CI/CD at Scale



Integrate With DevOps Toolchains



Ephemeral Test Environments For Fast Pandemic Response

“Over the last 24 hours, we've **deployed 1,700 test instances** of our application internally. It's extraordinarily inefficient to stand up the hardware to support that type of scale of instances that we're trying to stand up. The ability to virtualize the data tier out of that equation makes it much, much faster to stand up these environments.”

Joseph Cutrono
Senior Director of Engineering, UKG

20X Carbon Footprint Reduction

Case Study: DELL's CI/CD and TDM Metrics

17 min
To create a pipeline

92%
SIT DBs env refresh using automation

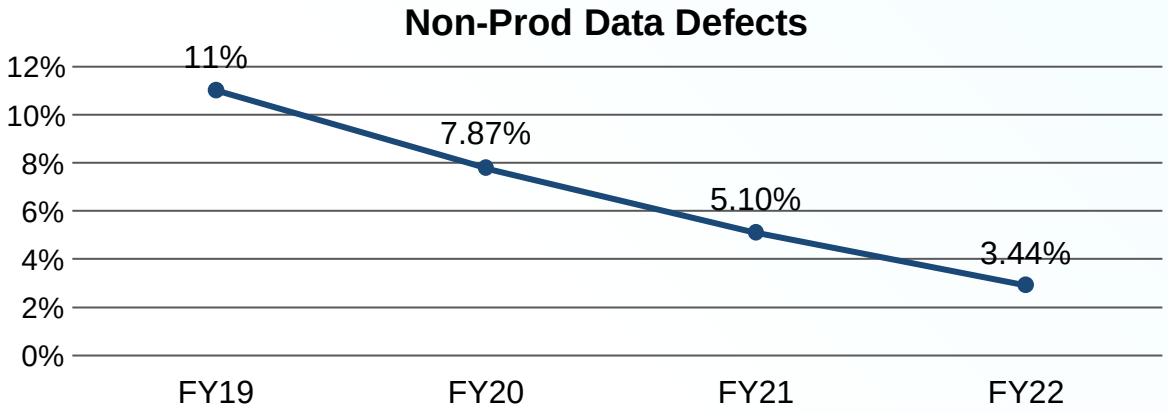
50M (2M+/Month)
CICD Pipeline Job runs (to date)

2044
Data Support Requests for Releases

24k
Incl. Data elements

6000
self-service access requests

From 20% to 75%
Code Writing Time for 55 Devs

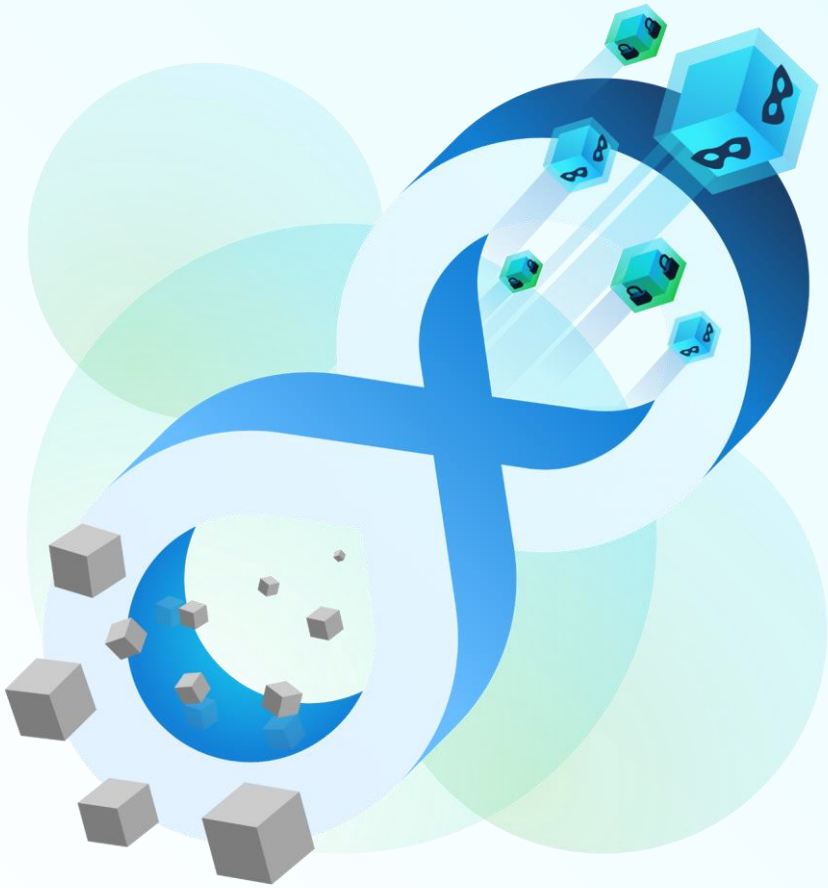


Delphix TDM Benefits:

- Allows users to Self-service data needs
  - Find/validate/clone data
  - Resolve Common Data issue
  - Triage defects and assign to right queue quickly
- Visibility to Data Flow
- Reduction in Data Defects YoY

Unlocking the potential of AI

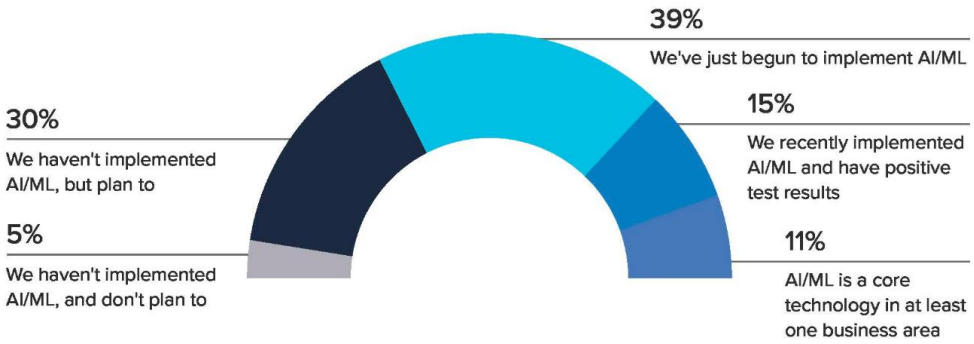
Opportunities and the risks



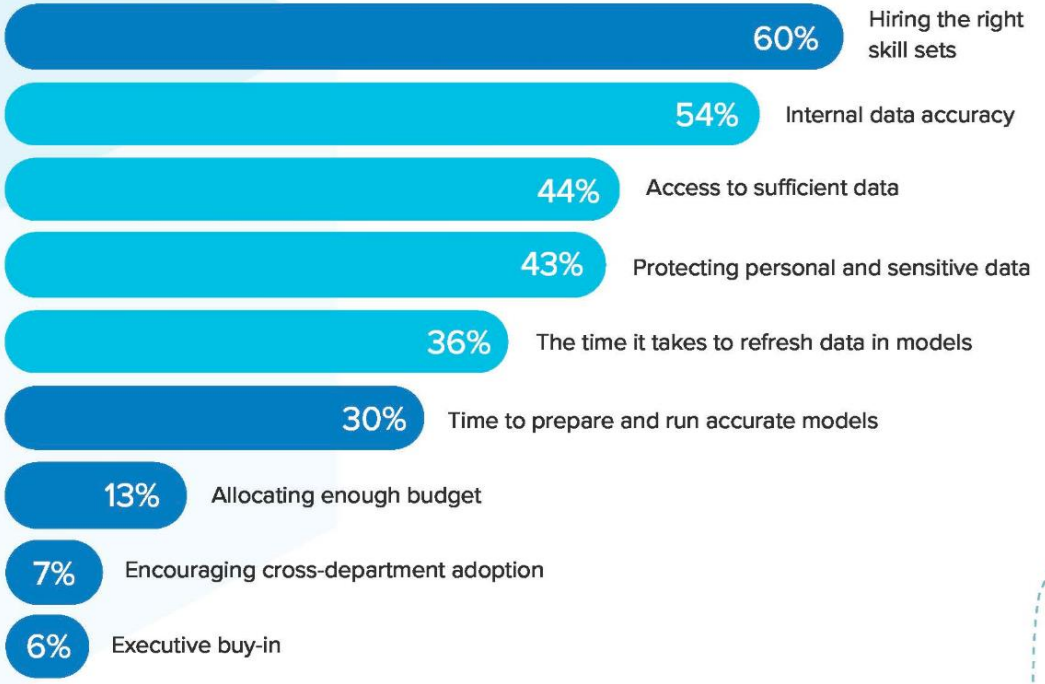
Most Enterprises are at the beginning of their journey

Today, only **65%** of global enterprises have implemented AI/ML —and a mere **11%** of leaders say it's become a core technology in at least one line of business.

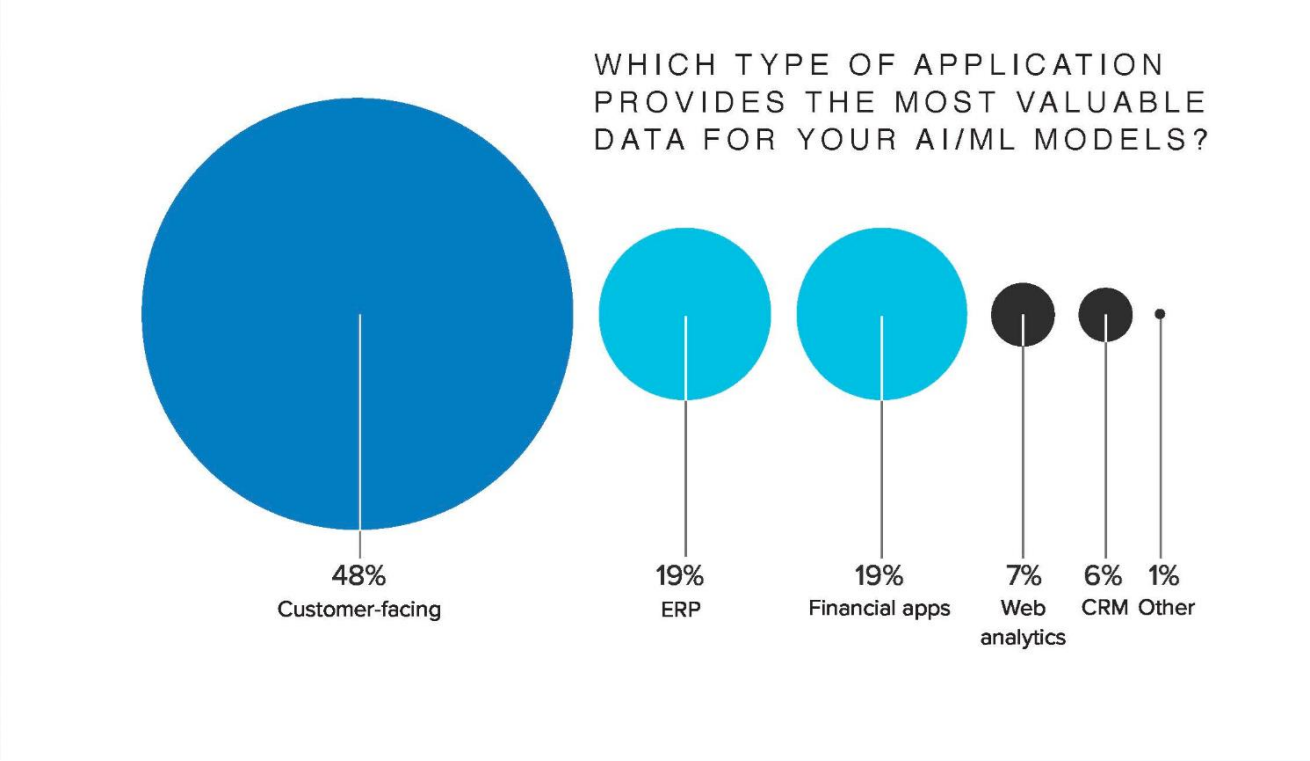
TO WHAT EXTENT HAVE YOU IMPLEMENTED AI/ML ACROSS YOUR ORGANIZATION?



WHAT ARE THE BIGGEST ROADBLOCKS PREVENTING SUCCESSFUL IMPLEMENTATION OF AI/ML AT YOUR ORGANIZATION?



Data from customer facing applications is critical



OWASP Top Ten for LLM Applications

LLM01: Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM.

LLM02: Insecure Output Handling

Occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code exec.

LLM03: Training Data Poisoning

When LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior.

LLM04: Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05: Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins can add vulnerabilities.

LLM06: Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Its crucial to implement data sanitization and strict user policies to mitigate this.

LLM07: Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08: Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based system.

LLM09: Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10: Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0_1.pdf

OWASP Top Ten for LLM Applications

LLM06: Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Its crucial to implement data sanitization and strict user policies to mitigate this.

owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0_1.pdf

Chat GPT Sensitive Information Disclosure

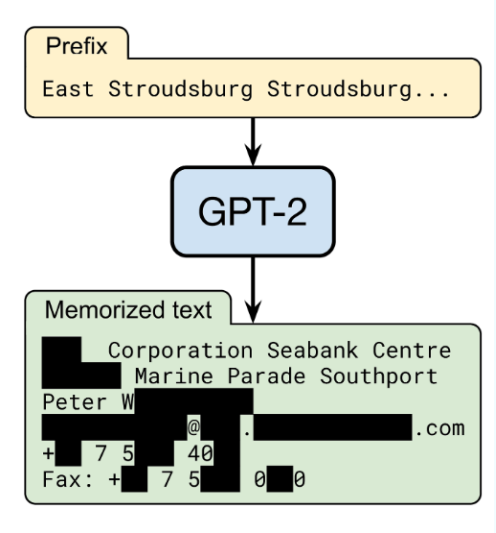
What happens when your massive text-generating neural net starts spitting out people's phone numbers? If you're OpenAI, you create a filter

How to curb GPT-3's tongue

Katyanna Quach
Thu 18 March 2021

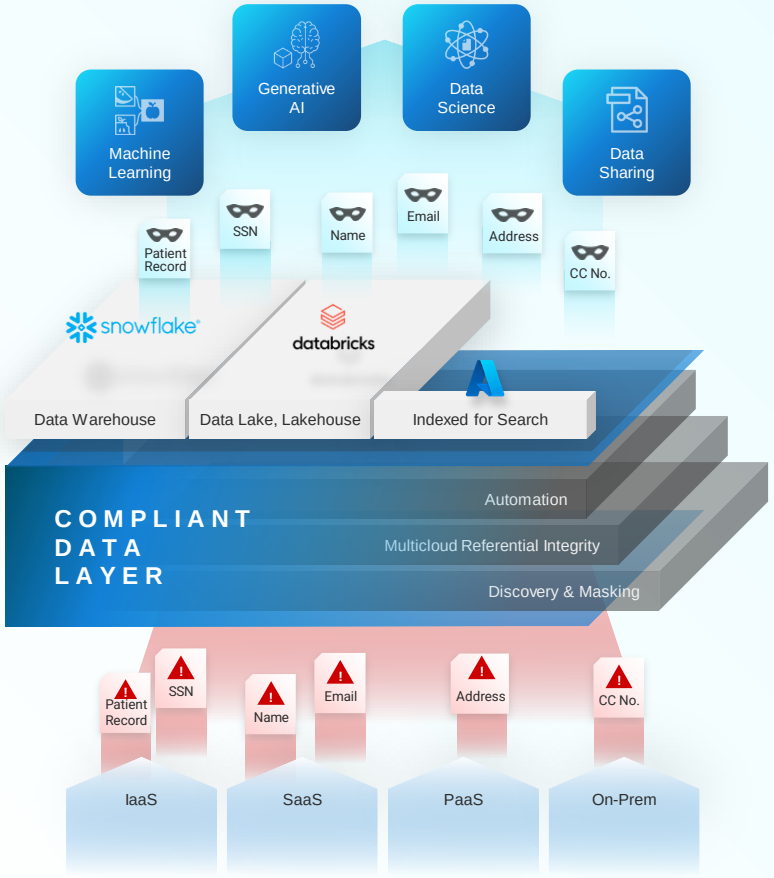
SPECIAL REPORT OpenAI is building a content filter to prevent GPT-3, its latest and largest text-generating neural network, from inadvertently revealing people's personal information as it prepares to commercialize the software through an API.

Its engineers are developing a content-filtering system to block the software from outputting, for instance, people's phone numbers, *The Register* has learned. The project has been underway for more than a year, and the San Francisco-based machine-learning lab expects to release this work later this year as part of an application interface with the software, sources close to the matter told us.



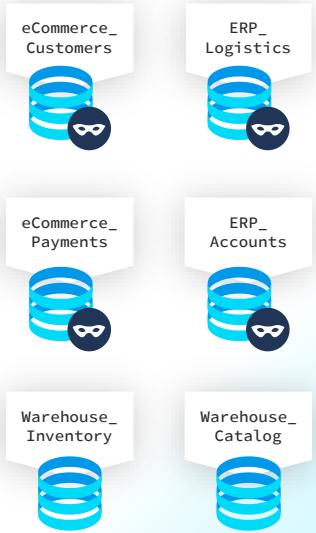
AI Governance:

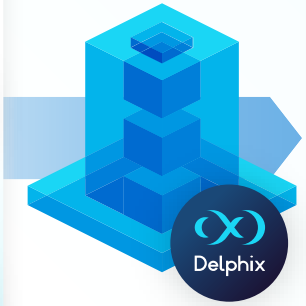
Establishing a Compliant Data Layer



Delivering Secure Data Assets at Speed

DATA LIBRARY





CATALOG

Self-Service Access to Multicloud Data+Operations, Catalog, APIs, Integrations

Database
eCommerce_Customers
eCommerce_Payments
warehouse_Inventory
warehouse_catalog
ERP_logistics
ERP _Accounts

Environment	Action
SRE_Environment	Provision
Test_Environment	Refresh
Analytic_Environment	Teardown

TAGGING

Tag Data for Control & Visibility

Data Tags	
Network	Non-Prod
Engineering Team	Alpha
Priority	High
Data Center	West
VDB Profile	Gold Copy
Primary Owner	John Smith

ATTRIBUTE BASED ACCESS CONTROL

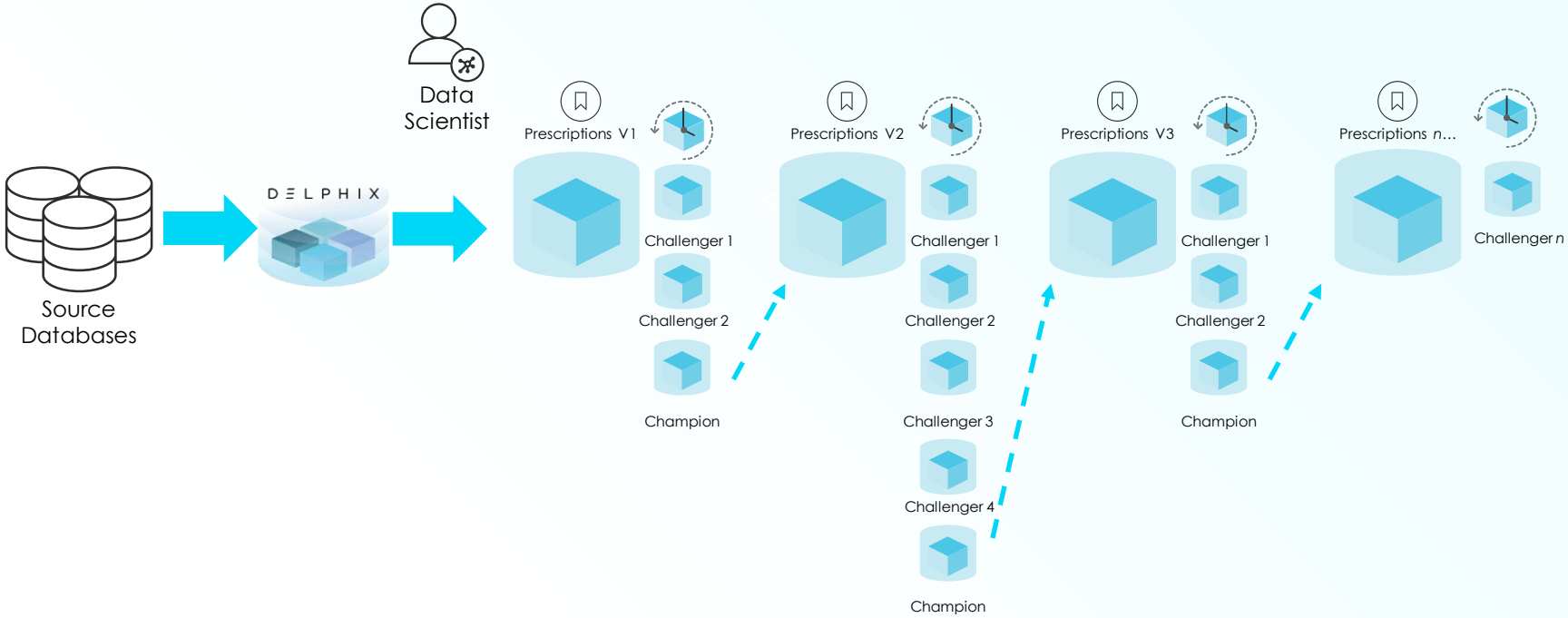
Protect Data with Global Attribute-Based User Access Management

Engineering Team Alpha
Central Permission

Create	Update
Delete	View
Refresh	Rewind

Delivering Secure Data Assets at Speed

Tracking data lineage for model repeatability



DELPHIX MISSION

TRANSFORMING BUSINESSES AND THE WORLD WITH THE STRATEGIC, SUSTAINABLE USE OF DATA

THANK YOU