Data Protection Considerations for Artificial Intelligence

Julya Rebstock, CIPT
Information Governance Practice Manager
Symantec by Broadcom

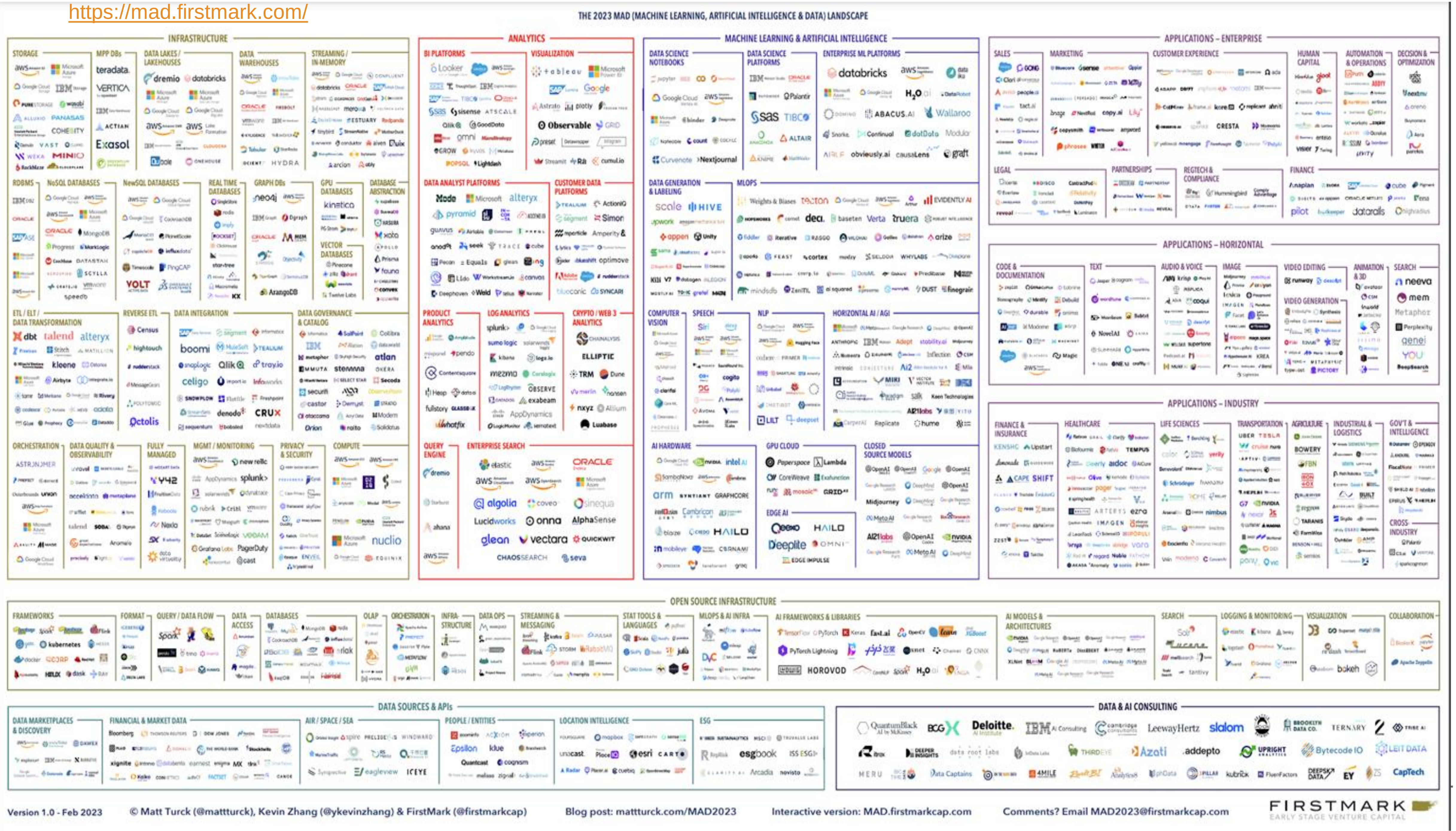
February 19, 2024

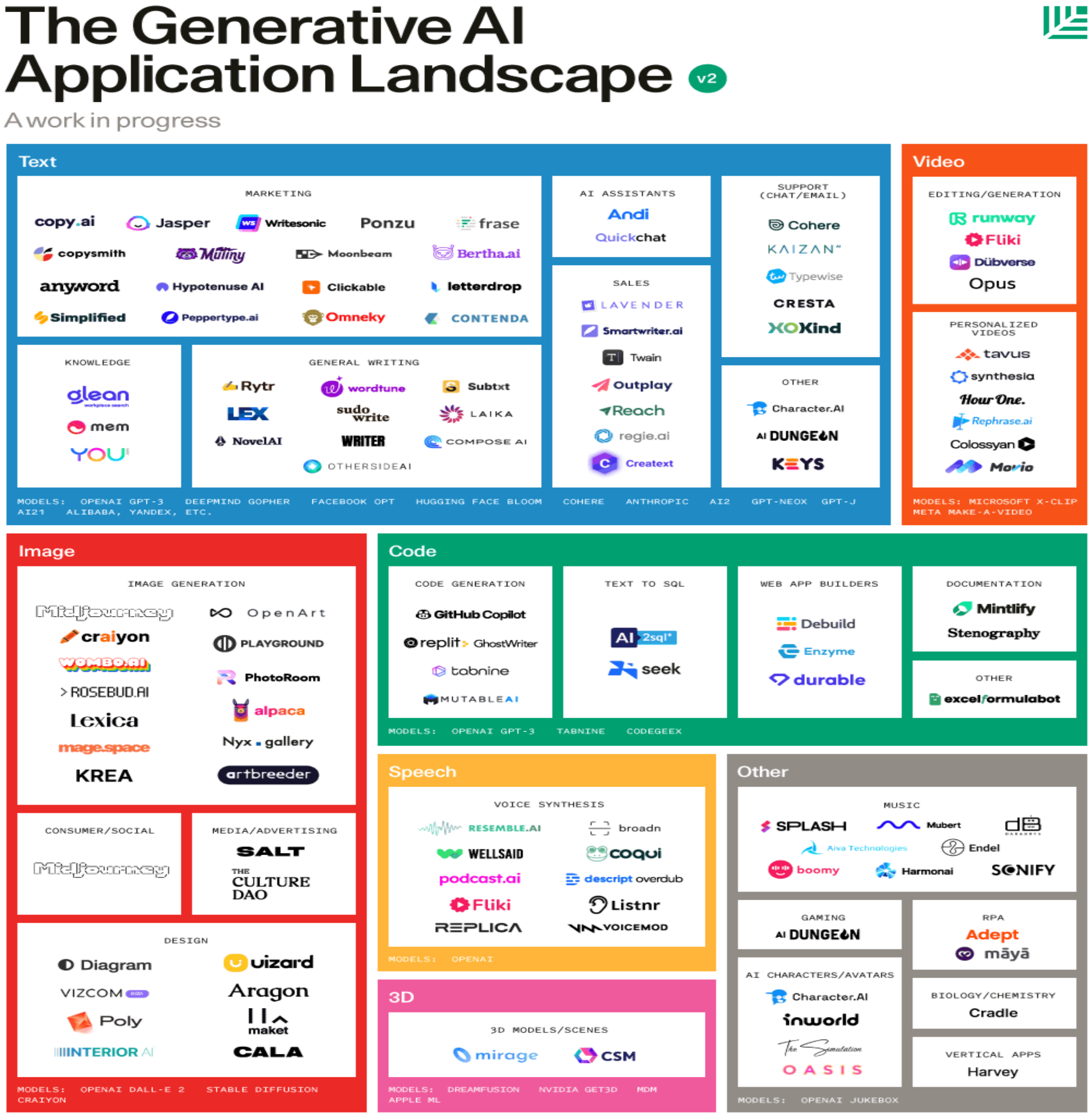
There are fundamental differences in the possible effects on our people depending on how and why they are engaging with AI…

There are typically 3 primary pathways of access to Artificial Intelligence:

1. Users (employees, citizens, etc.) accessing publicly available GenAI such as ChatGPT, BARD, etc.
    a. Presents data exfiltration risks for Organizations
2. Employees/students accessing enterprise contracted AI such as via corporate applications or websites.
    a. Presents data exfiltration risks for Organizations but usually has a security framework
3. Organizations building out their own AI models/apps for use in their business that will be accessed by others.
    a. What data will be used to train? Where could it be shared?

https://mad.firstmark.com/





The Generative AI Application Landscape v2

A work in progress

Text

MARKETING
copy.ai · Jasper · Writesonic · Ponzu · frase · copysmith · Mutiny · Moonbeam · Bertha.ai · anyword · Hypotenuse AI · Clickable · letterdrop · Simplified · Peppertype.ai · Omneky · CONTENDA

AI ASSISTANTS
Andi · Quickchat

SUPPORT (CHAT/EMAIL)
Cohere · KAIZAN · Typewise · CRESTA · XOKind

SALES
LAVENDER · Smartwriter.ai · Twain · Outplay · Reach · regie.ai · Creatext

KNOWLEDGE
glean · mem · YOU

GENERAL WRITING
Rytr · wordtune · Subtxt · LEX · sudowrite · LAIKA · NovelAI · WRITER · COMPOSE AI · OTHERSIDEAI

OTHER
Character.AI · AI DUNGEON · KEYS

MODELS: OPENAI GPT-3 DEEPMIND GOPHER FACEBOOK OPT HUGGING FACE BLOOM COHERE ANTHROPIC AI2 GPT-NEOX GPT-J AI21 ALIBABA, YANDEX, ETC.

Video

EDITING/GENERATION
runway · Fliki · Dübverse · Opus

PERSONALIZED VIDEOS
tavus · synthesia · Hour One. · Rephrase.ai · Colossyan · Movio

MODELS: MICROSOFT X-CLIP META MAKE-A-VIDEO

Image

IMAGE GENERATION
Midjourney · OpenArt · craiyon · PLAYGROUND · WOMBO.AI · PhotoRoom · > ROSEBUD.AI · alpaca · Lexica · Nyx.gallery · image.space · KREA · artbreeder

CONSUMER/SOCIAL
Midjourney

MEDIA/ADVERTISING
SALT · THE CULTURE DAO

DESIGN
Diagram · uizard · VIZCOM · Aragon · Poly · maket · INTERIOR AI · CALA

MODELS: OPENAI DALL-E 2 STABLE DIFFUSION CRAIYON

Code

CODE GENERATION
GitHub Copilot · replit GhostWriter · tabnine · MUTABLEAI

TEXT TO SQL
AI2sql · seek

WEB APP BUILDERS
Debuild · Enzyme · durable

DOCUMENTATION
Mintlify · Stenography

OTHER
excelformulabot

MODELS: OPENAI GPT-3 TABNINE CODEGEEX

Speech

VOICE SYNTHESIS
RESEMBLE.AI · broadn · WELLSAID · coqui · podcast.ai · descript overdub · Fliki · Listnr · REPLICA · VOICEMOD

MODELS: OPENAI

3D

3D MODELS/SCENES
mirage · CSM

MODELS: DREAMFUSION NVIDIA GET3D MDM APPLE ML

Other

MUSIC
SPLASH · Mubert · dB · Aiva Technologies · Endel · boomy · Harmonai · SONIFY

GAMING
AI DUNGEON

RPA
Adept · māyā

AI CHARACTERS/AVATARS
Character.AI · inworld · The Simulation · OASIS

BIOLOGY/CHEMISTRY
Cradle

VERTICAL APPS
Harvey

MODELS: OPENAI JUKEBOX

What are you looking for?

Purpose built for…

- Text
- Video
- Images
- Code
- Speech
- 3D
- Other

AI Apps

Purpose built on…

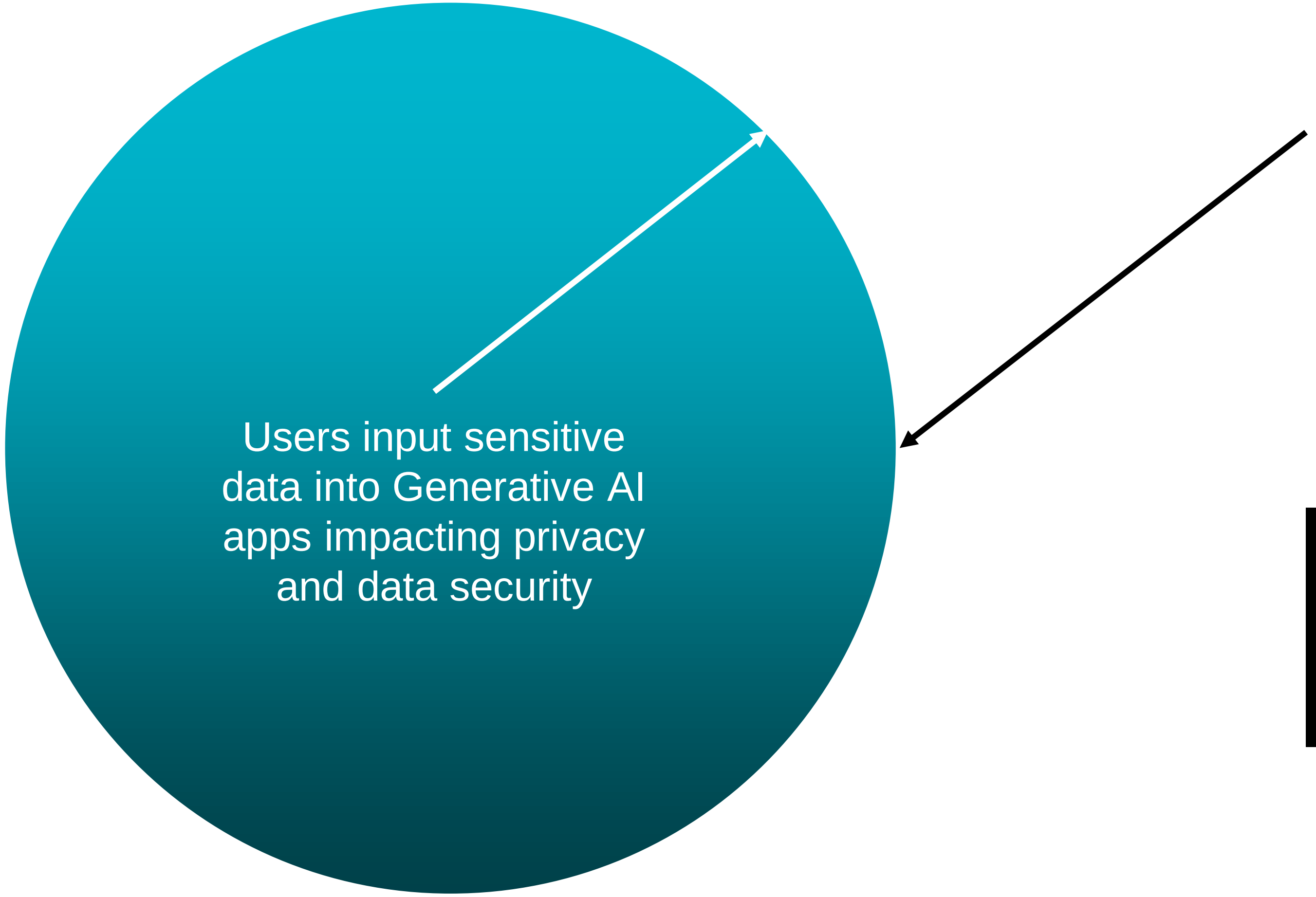
- AWS AI services
- Google AI Platform
- Azure Machine Learning
- IBM Watson

AI as a Service

https://research-assets.cbinsights.com/2022/11/14115447/SC-genAI-map.png

Symantec. by Broadcom

How Generative AI impacts you?



Exfiltration:

Users input sensitive data into Generative AI apps impacting privacy and data security

Infiltration:

Generative AI used by bad actors to design better attacks

- phishing email copy
- fake Gen AI portals
- malware design
- deep fakes

Symantec Security Center

Stay ahead of tomorrow's threats and security incidents with the latest information from the global leader in cyber security.

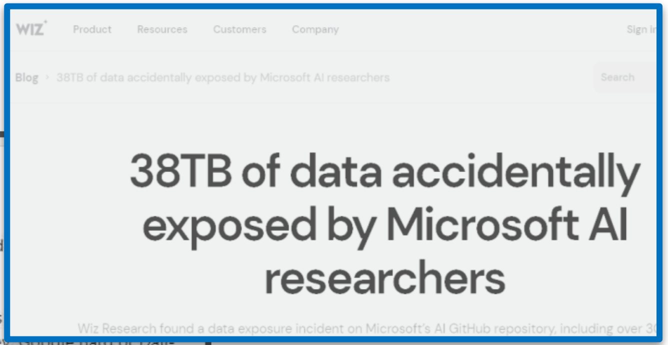
Symantec Security Center

Consider these Questions:

- Is accessing these services allowed?
  - *Should* you allow access to the services?
- How will it be used and by whom?
- What are the potential risks?
  - Security
  - Privacy
  - Legal
  - Regulatory
  - Economic
  - Human Rights
- What can we do?

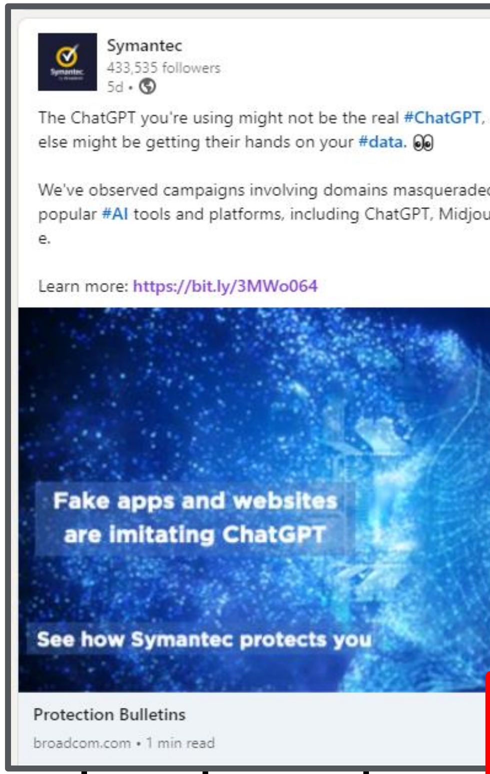
Generative AI: Security Risks

- Already being used to create:
  - Phishing campaigns
  - Keyloggers
  - Malicious code
  - “Deep fakes”
  - Virtual machines “inside” of ChatGPT
  - Undetectable steganography malware
  - AI poisoning & “sleeper agents”
  - Per FTC AI is “turbocharging fraud”
  - Which account profile used?
- ChatGPT has already had more than one data breach (due to an open source vulnerability that revealed other users chat histories)



38TB of data accidentally exposed by Microsoft AI researchers

Wiz Research found a data exposure incident on Microsoft's AI GitHub repository, including over 3



Symantec
433,535 followers
5d •

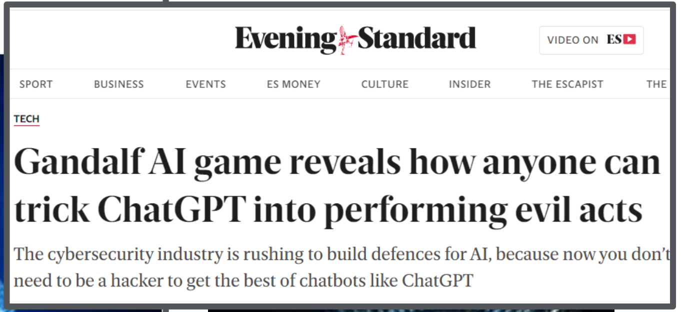
The ChatGPT you're using might not be the real #ChatGPT, and else might be getting their hands on your #data. 👀

We've observed campaigns involving domains masqueraded as popular #AI tools and platforms, including ChatGPT, Midjourney, Google bard or Dall-e.

Learn more: https://bit.ly/3MWo064

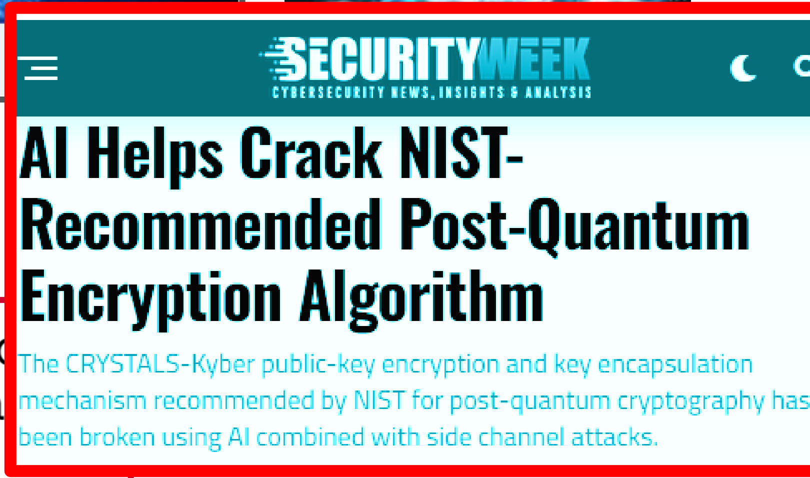
Fake apps and websites are imitating ChatGPT
See how Symantec protects you

Protection Bulletins
broadcom.com • 1 min read



Gandalf AI game reveals how anyone can trick ChatGPT into performing evil acts

The cybersecurity industry is rushing to build defences for AI, because now you don't need to be a hacker to get the best of chatbots like ChatGPT



AI Helps Crack NIST-Recommended Post-Quantum Encryption Algorithm

The CRYSTALS-Kyber public-key encryption and key encapsulation mechanism recommended by NIST for post-quantum cryptography has been broken using AI combined with side channel attacks.

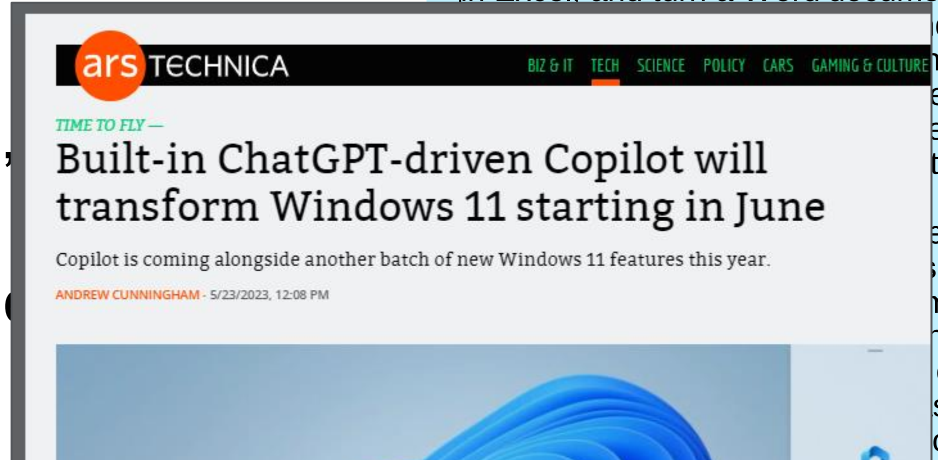
Group-IB Discovers 100K+ Compromised ChatGPT Accounts on Dark Web Marketpla
Asia-Pacific region tops the list

Generative AI: Integrations

- Microsoft announced integration into Windows 11 and O365 for “Copilot which will help edit, summarize, create, and compare documents.”
- Google’s Bard/Gemini for Google Cloud, MakerSuite, and Workspace.
- LinkedIn, Slack, Teams, Box among othe already using ChatGPT
- Anthropic's Claude, Meta's Llama
- Various Android AI from apps and OEM
- iOS AI

Microsoft

“With the new features, users will be able to
transcribe meeting notes during a Skype call,
summarize long email threads to quickly draft
suggested replies, request to create a specific chart
in Excel, and turn a Word document into a
ds.
ncept called
entially rides along
es to understand
t 365 data. The
email and on their
e documents
sentations they’ve
meeting with, and
ms platform,
can then ask
s write a status
ocuments across
d then draft an email
ith an update.” -
tech/openai-gpt-

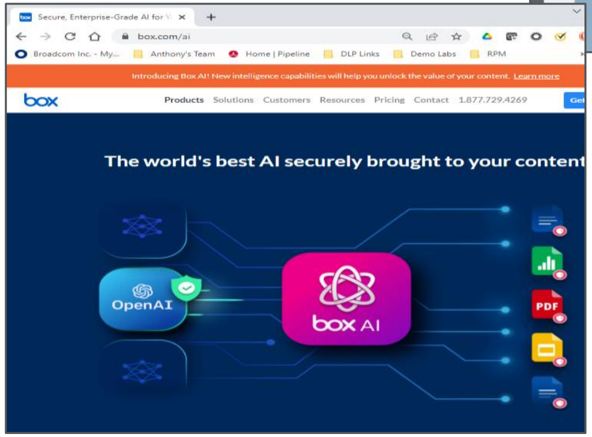


TIME TO FLY —

Built-in ChatGPT-driven Copilot will transform Windows 11 starting in June

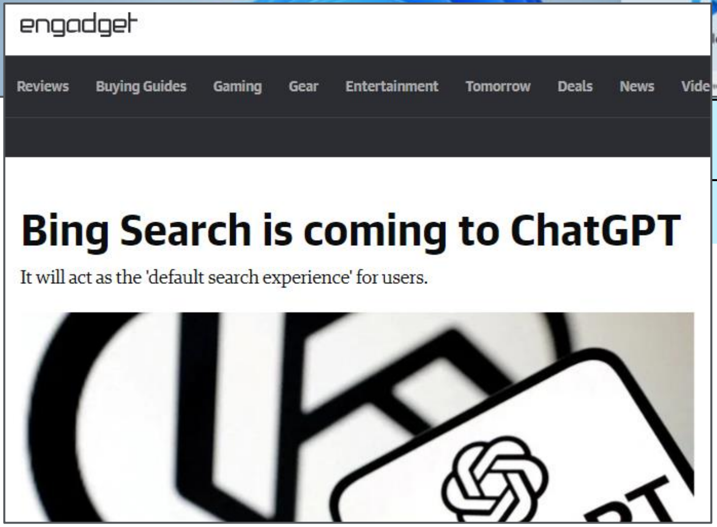
Copilot is coming alongside another batch of new Windows 11 features this year.

ANDREW CUNNINGHAM - 5/23/2023, 12:08 PM



Introducing Box AI! New intelligence capabilities will help you unlock the value of your content. Learn more

The world's best AI securely brought to your content



Bing Search is coming to ChatGPT

It will act as the 'default search experience' for users.

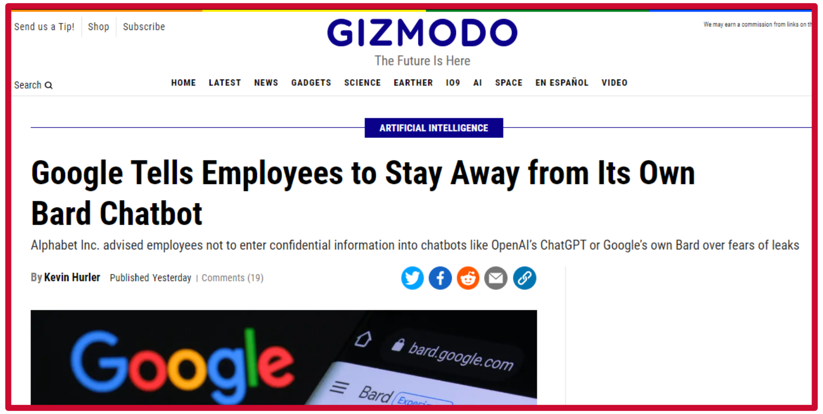
OpenAI - “One stop shop for the world’s data”

- Microsoft Windows 11 OS, O365 “Copilot”, ~~Bing~~ & Edge
- Box
- LinkedIn
- Slack “Einstein”
- Grammarly
- Snapchat
- Koo (X/Twitter rival)
- Quizlet
- Instacart
- Shopify
- Android & iOS apps
- Duolingo
- Expedia
- Insider
- FiscalNote
- Ghost
- KAYAK
- Klarna Shopping
- Milo Family
- OpenTable
- Speak
- Wolfram
- Zapier (interacts w/>5k apps)
- ***OpenAI GPT Store offers >3M Custom AI Bots***

Plugins: web browser, code interpreter, retrieval

Generative AI: Privacy Concerns

- Any data presented to it becomes its property and can be reviewed by staffers.
  - API submitted data not used to train the LLM but still retained 30 days
  - Public site data may be used to train the LLM, users have a specific Opt-Out form to complete if they don’t want their data used
- Lack of Transparency
  - How was data collected?
  - How is it processed?
  - How are decisions made?
  - “Automated discrimination” and bias concerns



Google Tells Employees to Stay Away from Its Own Bard Chatbot

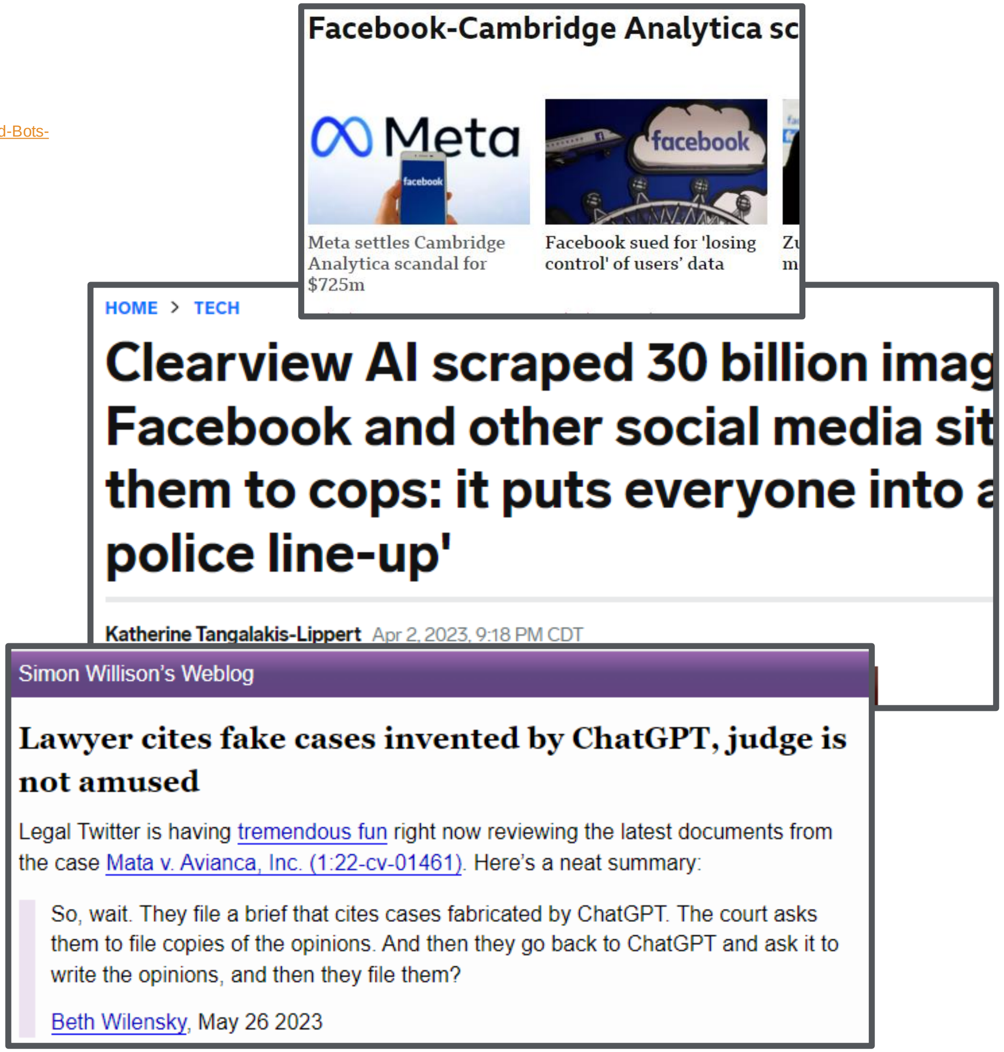
Alphabet Inc. advised employees not to enter confidential information into chatbots like OpenAI’s ChatGPT or Google’s own Bard over fears of leaks

By Kevin Hurler Published Yesterday | Comments (19)

- How do you comply with “*right to be forgotten*” or “*access request*” with a learning model?
- Data scraped from internet is regulated differently across privacy jurisdictions globally and may not meet “legal basis” required for collection, storage, and processing.
- AI used to process data is typically considered “3rd party processing”

Generative AI: Legal Concerns

- Do AI companies have the rights to the training data?
  - Liability risk for web scraping technologies
  - “Posting publicly is not implied consent” https://jolt.law.harvard.edu/assets/articlePDFs/v34/6.-Xiao-Bad-Bots-Regulating-the-Scraping-of-Public-Personal-Information.pdf
- Is the output of generative AI new and original content, or a derivative of the data sets it trained on?
  - Can you guarantee the data returned wasn’t copyrighted?
  - How do you attribute source?
- Can you trust the information you are seeing?
  - AI “**hallucinations**” are an issue.
  - What if the output is misleading or dangerous?
  - How are decisions made?
  - AI now also faces “sleeper agents”
- Could be considered “3rd party software” but YOUR company may not have a license.



Facebook-Cambridge Analytica sc

Meta settles Cambridge Analytica scandal for $725m

Facebook sued for 'losing control' of users' data

Zu
m

HOME > TECH

Clearview AI scraped 30 billion imag
Facebook and other social media sit
them to cops: it puts everyone into a
police line-up'

Katherine Tangalakis-Lippert Apr 2, 2023, 9:18 PM CDT

Simon Willison’s Weblog

Lawyer cites fake cases invented by ChatGPT, judge is not amused

Legal Twitter is having tremendous fun right now reviewing the latest documents from the case Mata v. Avianca, Inc. (1:22-cv-01461). Here’s a neat summary:

> So, wait. They file a brief that cites cases fabricated by ChatGPT. The court asks them to file copies of the opinions. And then they go back to ChatGPT and ask it to write the opinions, and then they file them?
>
> Beth Wilensky, May 26 2023

Generative AI: Regulatory Concerns

EU Artificial Intelligence Act - proposed text approved Feb 2, 2024

- Banned: Social credit scoring, emotion recognition, predictive policing, “real-time” biometric identification, exploitation/manipulation ←
- High Risk: Quite complex to determine, many exceptions to be explored
  - Medical devices
  - Infrastructure
  - Education/training
  - Law enforcement, migration, asylum management
  - Judicial/democratic
- Obligations:
  - Summarize training data used
  - Respect the Copyright Directive
  - Enforce watermarking, metadata identification and cryptographic methods
  - Continuous monitoring of General Purpose AI (GPAI) in case it evolves into GPAI with Systemic Risks
  - Ensure good data and cybersecurity for the AI and its infrastructure

Generative AI: Regulatory Concerns (cont'd)

US AI Governance Law and Policy

- Executive orders:
  - Maintaining American Leadership in AI
  - Promoting the Use of Trustworthy AI in the Federal Government
- Acts and bills:
  - AI Training Act
  - National AI Initiative Act (Division E, Sec. 5001; in force)
  - AI in Government Act (Division U, Sec. 101; in force)
  - Algorithmic Accountability Act (Draft)
  - National AI Commission Act (Draft)
  - Digital Platform Commission Act (Draft)
  - Transparent Automated Governance Act (Draft)
- Relevant:
  - FTC Act, Section 5 (in force)
  - Title VII of the Civil Rights Act (in force)
  - Americans with Disabilities Act (in force)
  - Age Discrimination in Employment Act (in force)
  - Genetic Information and Nondiscrimination Act (in force)
  - American Data Privacy and Protection Act (draft)
  - Health Equity and Accountability Act (draft)

Generative AI: Regulatory Concerns (cont’d)

US Healthcare Specific

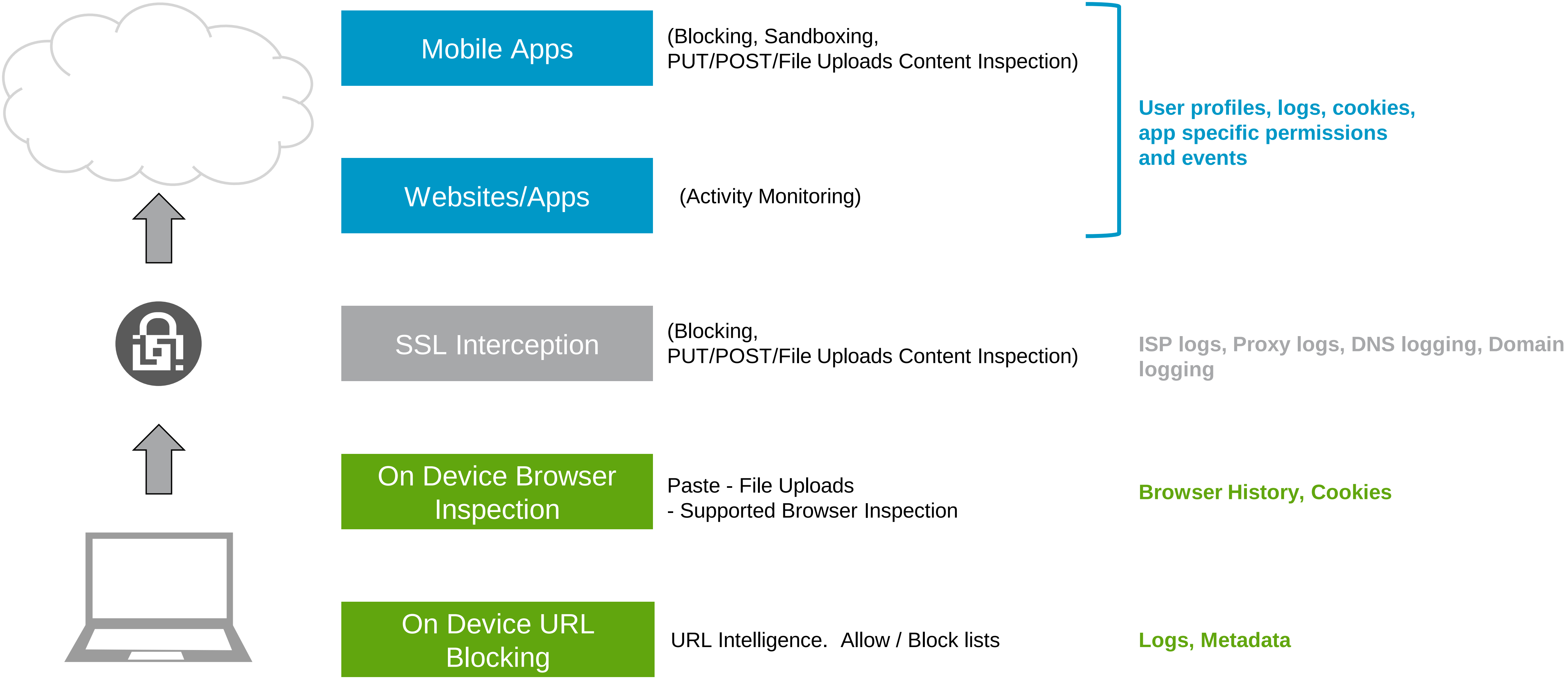
- Healthcare Sector Cybersecurity
- Health Equity and Accountability Act (draft)
- HIPAA **expect updates to HIPAA Security Rules**
    - LLM use could require collection and storage of excessive quantities of PHI
    - Opt-in to data sharing? Must be defined and limited use
    - Review data access to and use of PHI within the models only for specific limited purposes
    - **BAA required with the vendor of the AI**

Generative AI: Economic Concerns

- Many AI integrations bypass website “pay walls” impacting the revenue streams for content creators. AI aggregators/summaries may:
  - Deny “referral” credit to websites
  - Avoid redirect and notification options
  - Avoid ads, cookies/tracking, subscriptions
  - Deny affiliate link capabilities

- Multiple AI integrations allow for video summaries (such as from YouTube)
  - May not register the “view”, affecting the creator’s algorithm statistics
  - Disrupts ad revenues commonly shown during videos
  - Limits discovery of additional content normally advertised and “subscribe” or “follow”
- Content creators/providers that rely on internet revenue methods loose valuable income streams. Many of these are small businesses or independents.

Layered Data Inspection and Collection



Mobile Apps

(Blocking, Sandboxing,
PUT/POST/File Uploads Content Inspection)

Websites/Apps

(Activity Monitoring)

User profiles, logs, cookies, app specific permissions and events

SSL Interception

(Blocking,
PUT/POST/File Uploads Content Inspection)

ISP logs, Proxy logs, DNS logging, Domain logging

On Device Browser Inspection

Paste - File Uploads
- Supported Browser Inspection

Browser History, Cookies

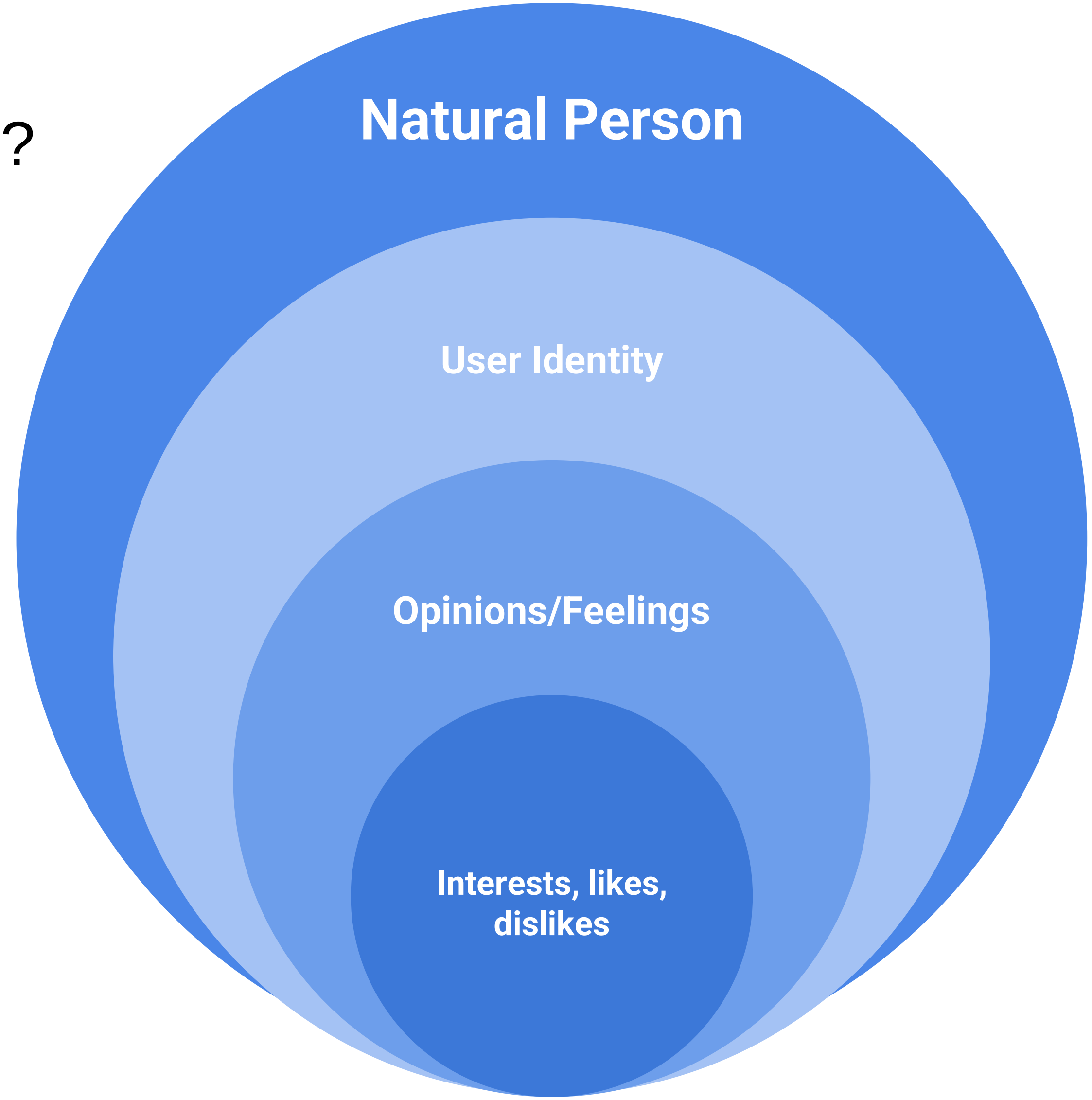
On Device URL Blocking

URL Intelligence. Allow / Block lists

Logs, Metadata

Generative AI: Human Rights

What does all that collected data represent?



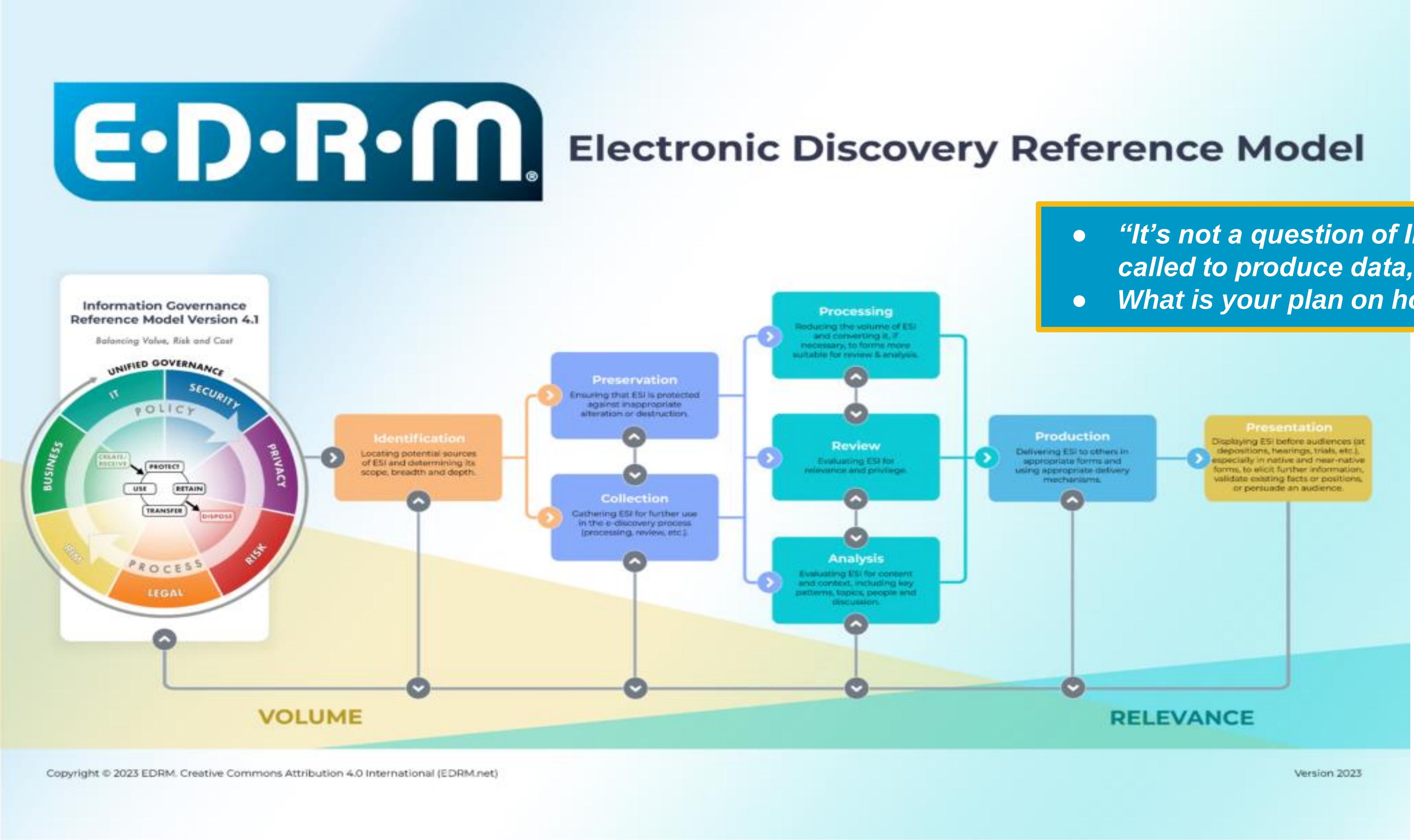
Natural Person

User Identity

Opinions/Feelings

Interests, likes, dislikes

Collecting and Retaining Data is a LIABILITY



Identification: Locating potential sources of ESI and determining its scope, breadth and depth.

Preservation: Ensuring that ESI is protected against inappropriate alteration or destruction.

Collection: Gathering ESI for further use in the e-discovery process (processing, review, etc.).

Processing: Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review & analysis.

Review: Evaluating ESI for relevance and privilege.

Analysis: Evaluating ESI for content and context, including key patterns, topics, people and discussion.

Production: Delivering ESI to others in appropriate forms and using appropriate delivery mechanisms.

Presentation: Displaying ESI before audiences (at depositions, hearings, trials, etc.), especially in native and near-native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

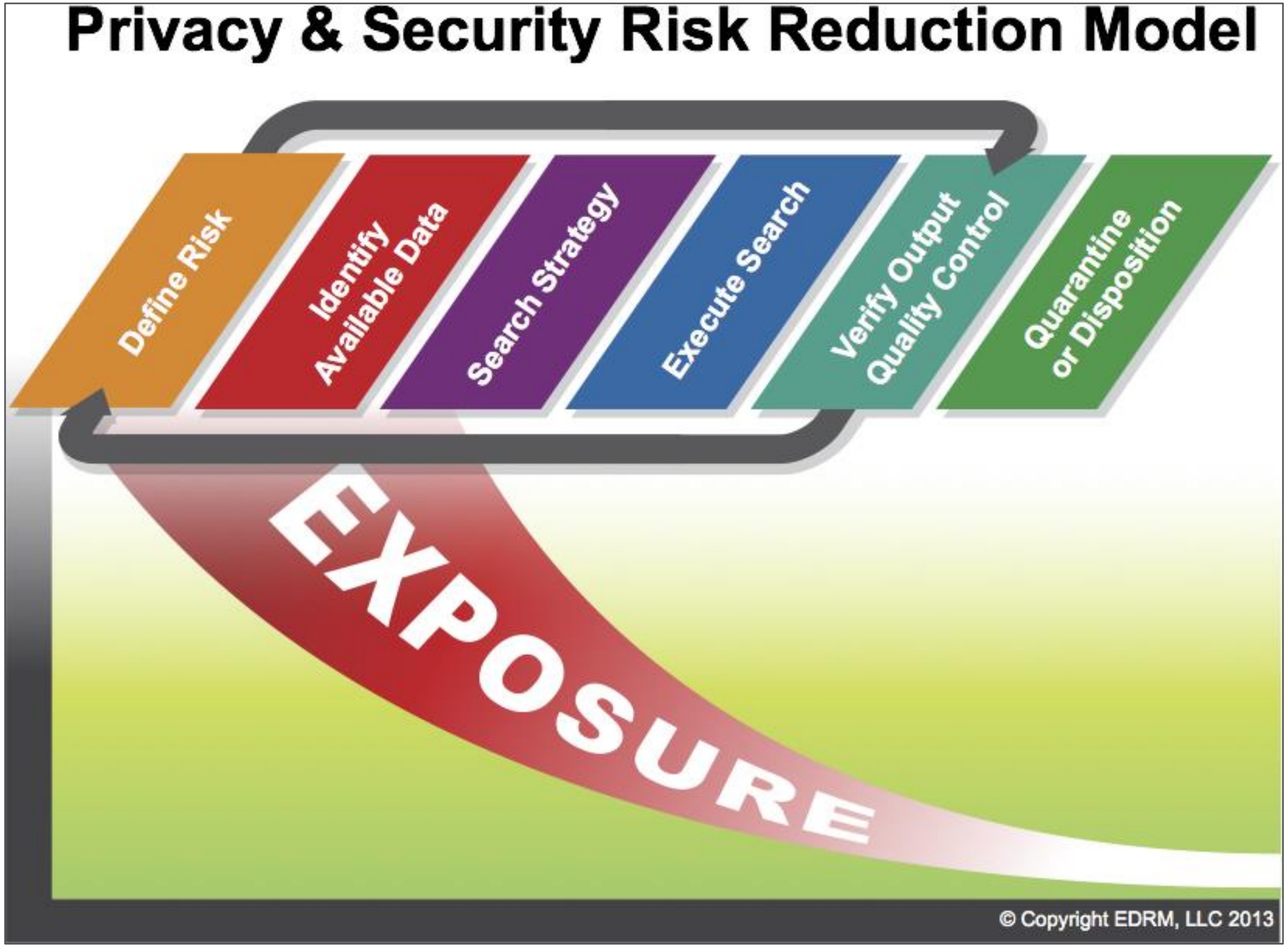
Version 2023

- *“It’s not a question of IF you will be called to produce data, but WHEN”*
- *What is your plan on how to respond?*

Symantec. by Broadcom

What can we do? Don’t recreate the wheel!

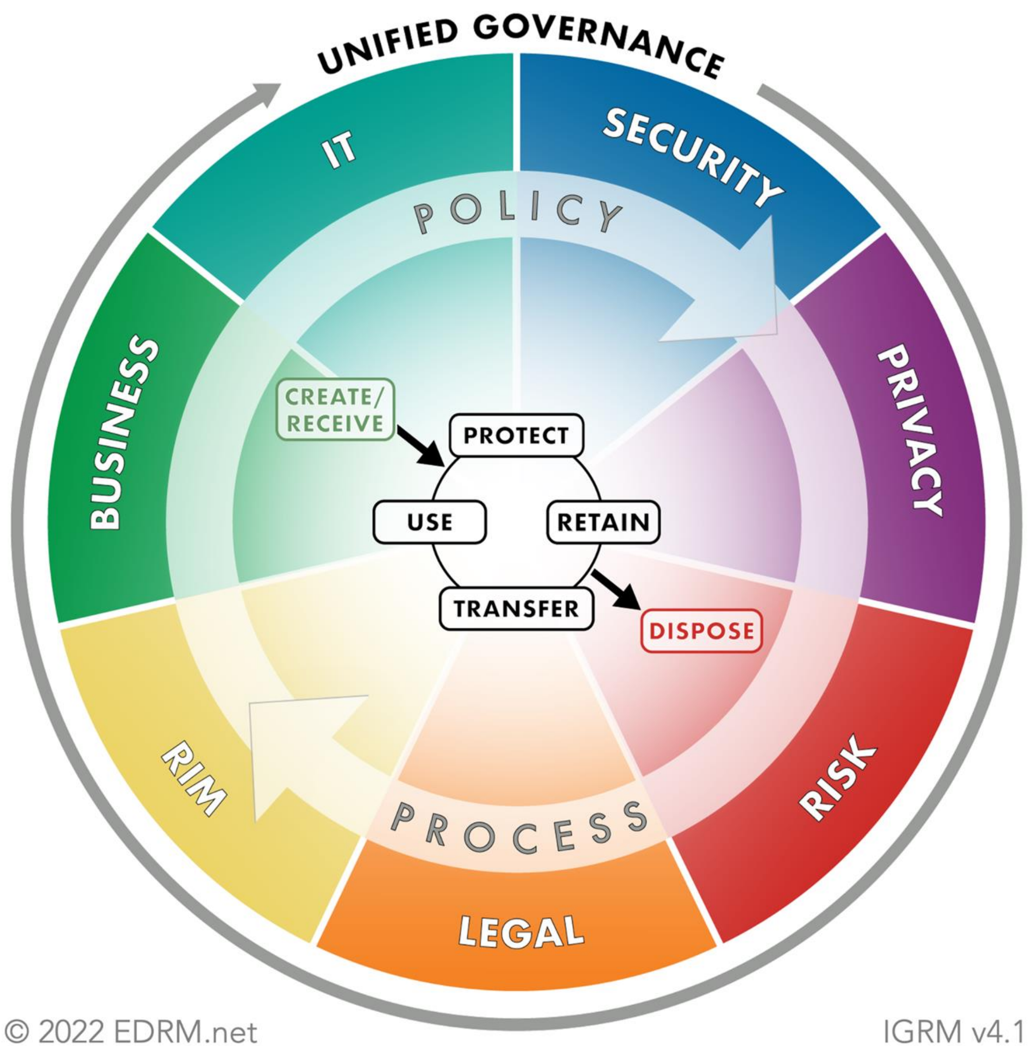
https://edrm.net/resources/frameworks-and-standards/privacy-and-security-risk-reduction-model/


Privacy & Security Risk Reduction Model
Define Risk
Identify Available Data
Search Strategy
Execute Search
Verify Output Quality Control
Quarantine or Disposition
EXPOSURE
© Copyright EDRM, LLC 2013

What can we do? Help to define the data lifecycle

Information Governance Reference Model (IGRM)

Balancing Value, Risk and Cost



© 2022 EDRM.net

IGRM v4.1

Enabling proper Information Governance and DATA lifecycle management is imperative.

The Information Governance Reference Model is also part of the EDRM and helps organizations to frame the data they are collecting, using and retaining.

As Global partners and industry leaders, we need synchronization

GDPR

- Data Protection
- Regulation Compliance
- Global Impact
- Standardization
- Risk Assessments
- Accountability Principle

EDRM - IGRM - PSRRM

- Data Protection
- Regulation Compliance
- Global Impact
- Standardization
- Risk Assessments
- Accountability Principles

EU AI Act

- Data Protection
- Regulation Compliance
- Global Impact
- Standardization
- Risk Assessments
- Accountability Principles

Thank You

Julya Rebstock

Information Governance Practice Manager

julya.rebstock@broadcom.com

+1 936-306-8518

https://www.linkedin.com/in/julya-rebstock

https://www.broadcom.com/products/cybersecurity