




Privacy SIG Webinar
**How to address privacy
needs across the AI tech
stack**




September 19, 2023





Chat Bots == Data Leaks

 **REUTERS®** [World](#) [Business](#) [Markets](#) [Sustainability](#) [Legal](#) [Breakingviews](#) [Technology](#) [Investigations](#)




Disrupted

Google, one of AI's biggest backers, warns own staff about chatbots

By Jeffrey Dastin and Anna Tong

June 15, 2023 12:47 PM MDT · Updated a day ago



Regulations Evolving Quickly



news | analysis | podcasts | the magazine | newsletters | fp live | events | fp analytics

WORLD BRIEF

FP's flagship evening newsletter guiding you through the most important world stories of the day. Delivered weekdays.

EU Lawmakers Pass Landmark AI Regulation Bill

The AI Act instills greater privacy standards, stricter transparency laws, and steeper fines for failing to cooperate.

Bloomberg

US Edition ▾ Sign In [Subscribe](#)

• Live Now Markets Economics Industries Technology **Politics** Wealth Pursuits Opinion Businessweek Equality Green CityLab Crypto More ⋮

Politics

Biden to Meet AI Experts as He Pushes for Privacy Safeguards

- White House calls AI top priority as its popular use explodes
- Administration says companies working on privacy, security



• LIVE ON BLOOMBERG

[Watch Live TV >](#)

[Listen to Live Radio >](#)

#ISSAPrivacySIG

Regulations Evolving Quickly



ews | analysis | podcasts | the magazine | newsletters | fp live | events | fp analytics

WORLD BRIEF

FP's flagship evening newsletter guiding you through the most important world stories of the day. Delivered weekdays.

EU Lawmakers Pass Landmark AI Regulation Bill

The AI Act instills greater privacy standards, stricter transparency laws, and steeper fines for failing to cooperate.

g US Edition ▼ Sign In Subscribe 🔍
ets Economics Industries Technology Politics Wealth Pursuits Opinion Businessweek Equality Green CityLab Crypto More ⋮

Biden to Meet AI Experts as He Pushes for Privacy Safeguards

- White House calls AI top priority as its popular use explodes
- Administration says companies working on privacy, security


● LIVE ON BLOOMBERG
Watch Live TV >
Listen to Live Radio >

Make your AI systems secure and private by design and by default

Make sure AI systems you use are secure and private by design as well.

#ISSAPrivacySIG

Privacy Lawsuits Flying



REUTERS®




World ▾ Business ▾ Markets ▾ Sustainability ▾ Legal ▾ Breakingviews Technology ▾ Investigations More ▾


Litigation | Data Privacy | Data Privacy | Litigation | Intellectual Property

OpenAI, Microsoft hit with new US consumer privacy class action

By **Blake Brittain**

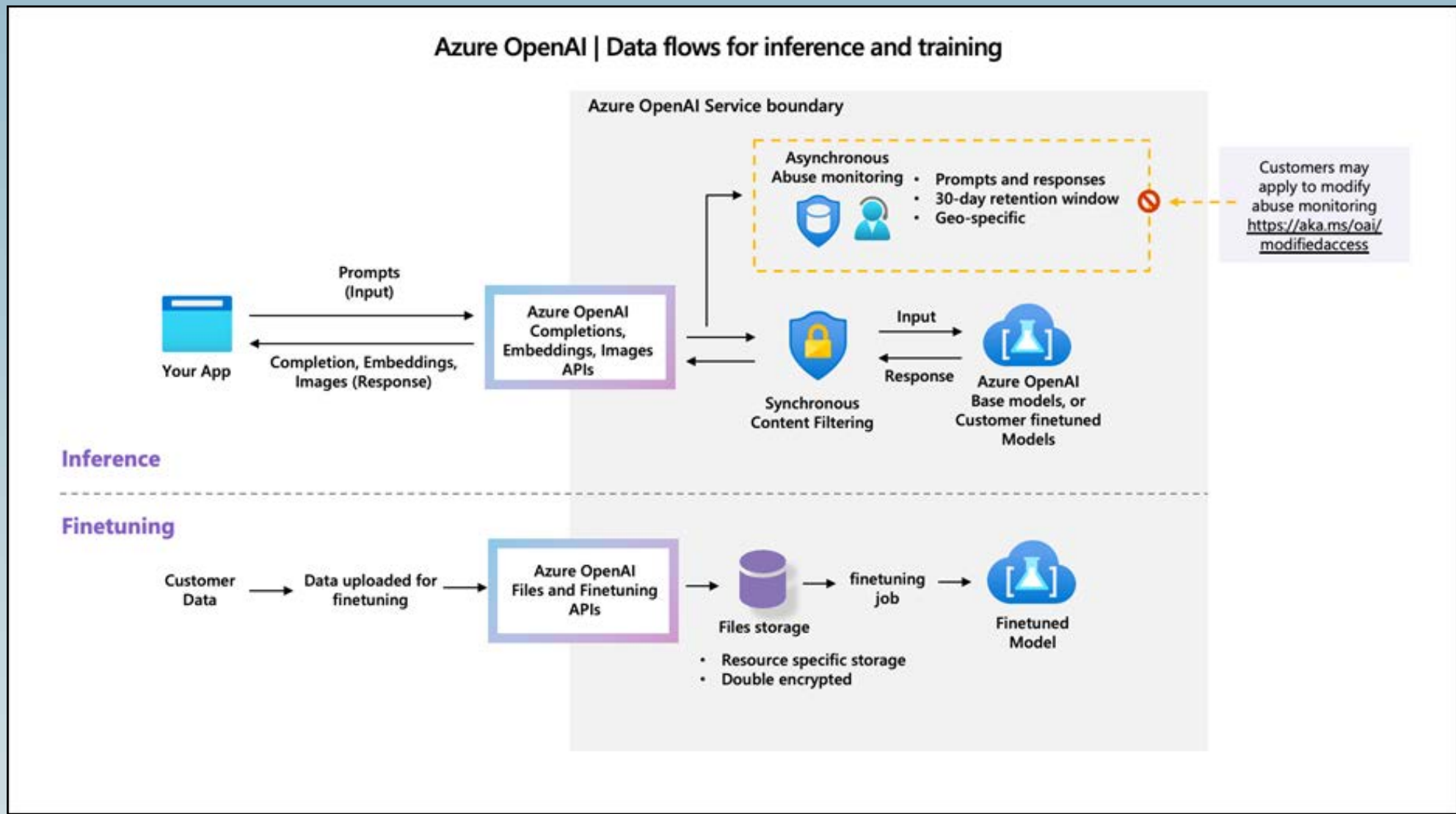
September 6, 2023 1:52 PM MDT · Updated 7 days ago





OpenAI logo and rising stock graph are seen in this illustration taken, February 3, 2023. REUTERS/Dado Ruvic/Illustration/File Photo *Acquire*

Non-obvious Data Flows



Non-obvious Data Flows



Research ▾ API ▾ ChatGPT ▾ Safety Company ▾

Search

Log in ↗

Get started ↗

How does OpenAI handle data retention and monitoring for API usage?

OpenAI may securely retain API inputs and outputs for up to 30 days to identify abuse. You can also request zero data retention (ZDR) for eligible endpoints if you have a qualifying use-case. For details on data handling, visit our Platform Docs page.

Who can view stored API inputs, outputs, and fine-tuning data?

Access to API business data stored on our systems is limited to (1) authorized employees that require access for engineering support, investigating potential platform abuse, and legal compliance and (2) specialized third-party contractors who are bound by confidentiality and security obligations, solely to review for abuse and misuse.

Where's the data?

- As with any privacy control, you need to know **what data** flows **where** and **who can and does see** that data.
- You probably also need to be **able to remove specific data** by request or at the end of a retention period.
- And you need data protection controls.

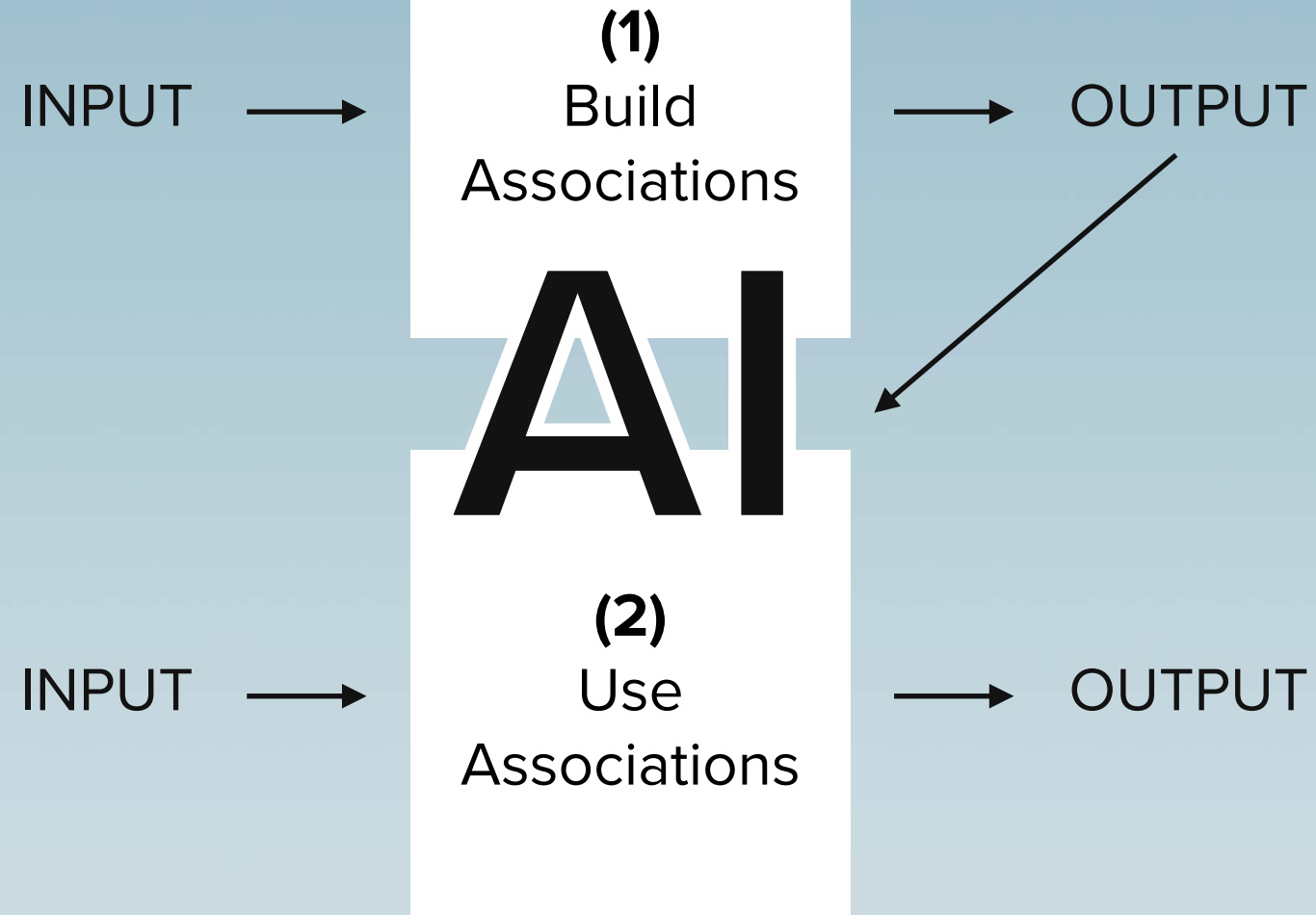
Artificial Intelligence

INPUT →

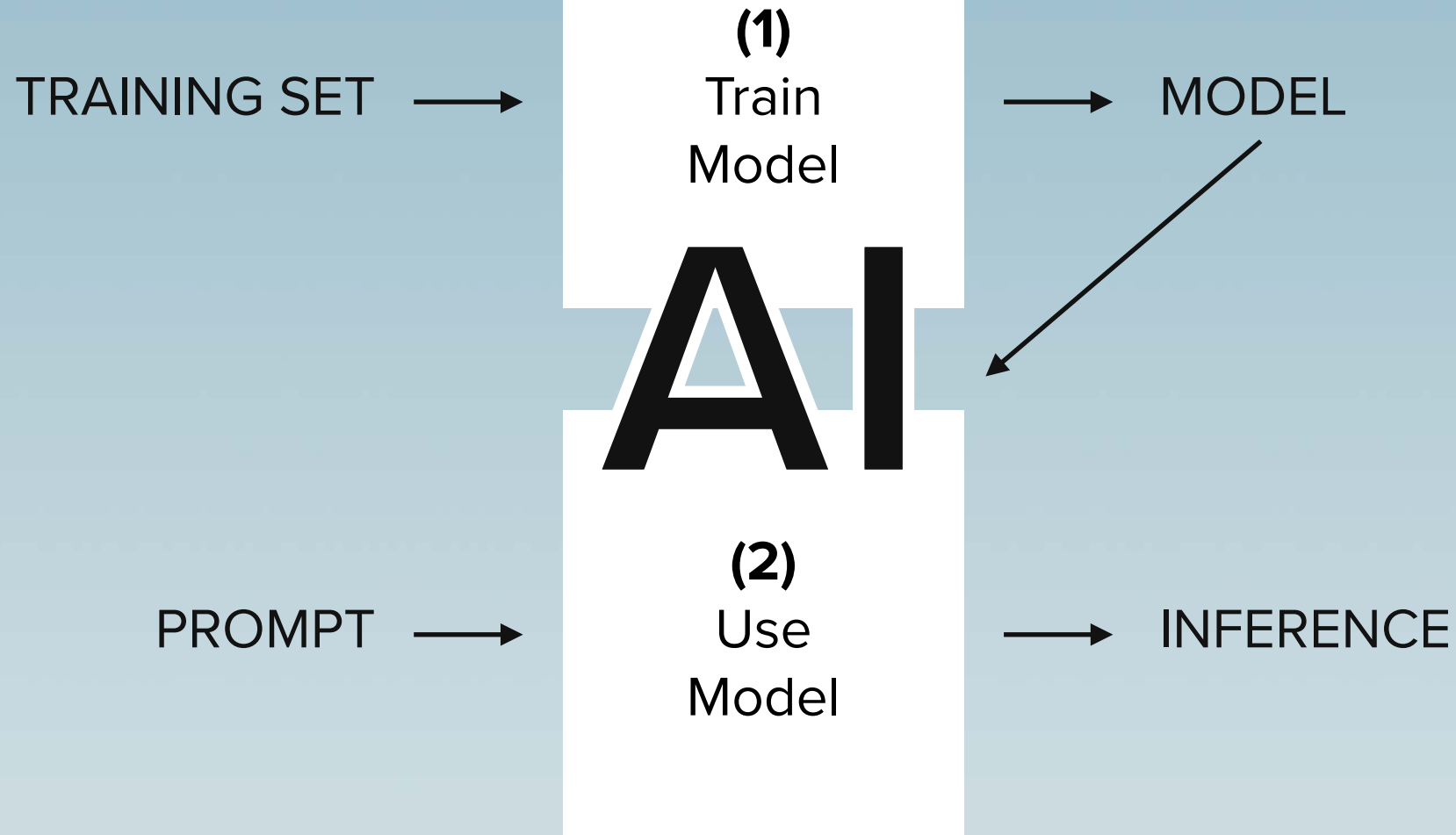
AI

→ OUTPUT

Artificial Intelligence



Artificial Intelligence



aka Machine Learning

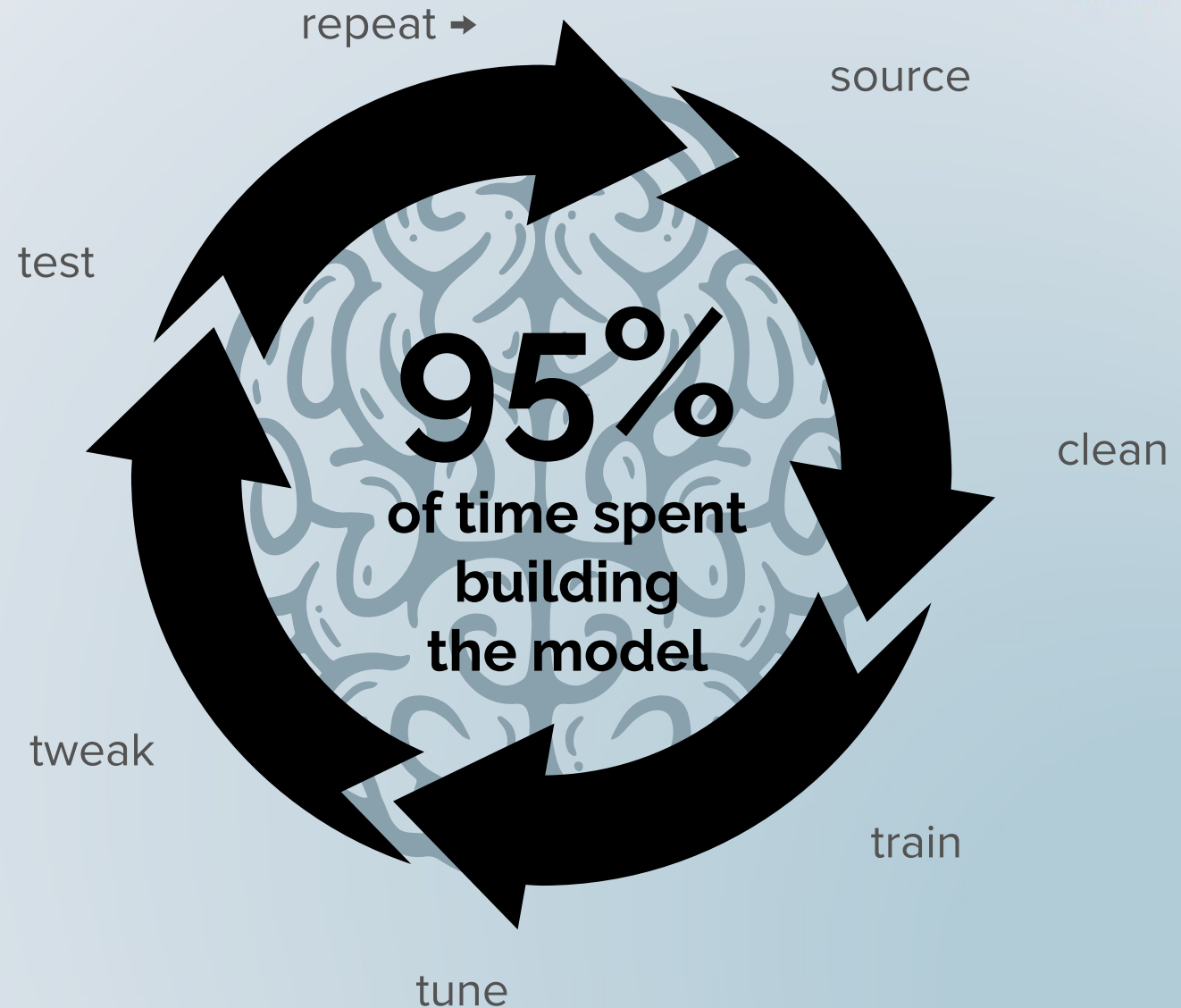
- 1943: Discovery of the "artificial neuron"
- 1950's: First AI uses simple algorithms and statistical methods
- 1960's: Use of bayesian methods for probabilistic inference
- 1980's: Recurrent neural networks and reinforcement learning
- 1990's: Support vector machines and Deep Blue Chess
- 2000's: Support vector clustering
- 2017: Transformer architecture
- 2021+: Breakthroughs in generative AI and large generally intelligent models



I've been a part of ~5 machine learning projects since 2000.

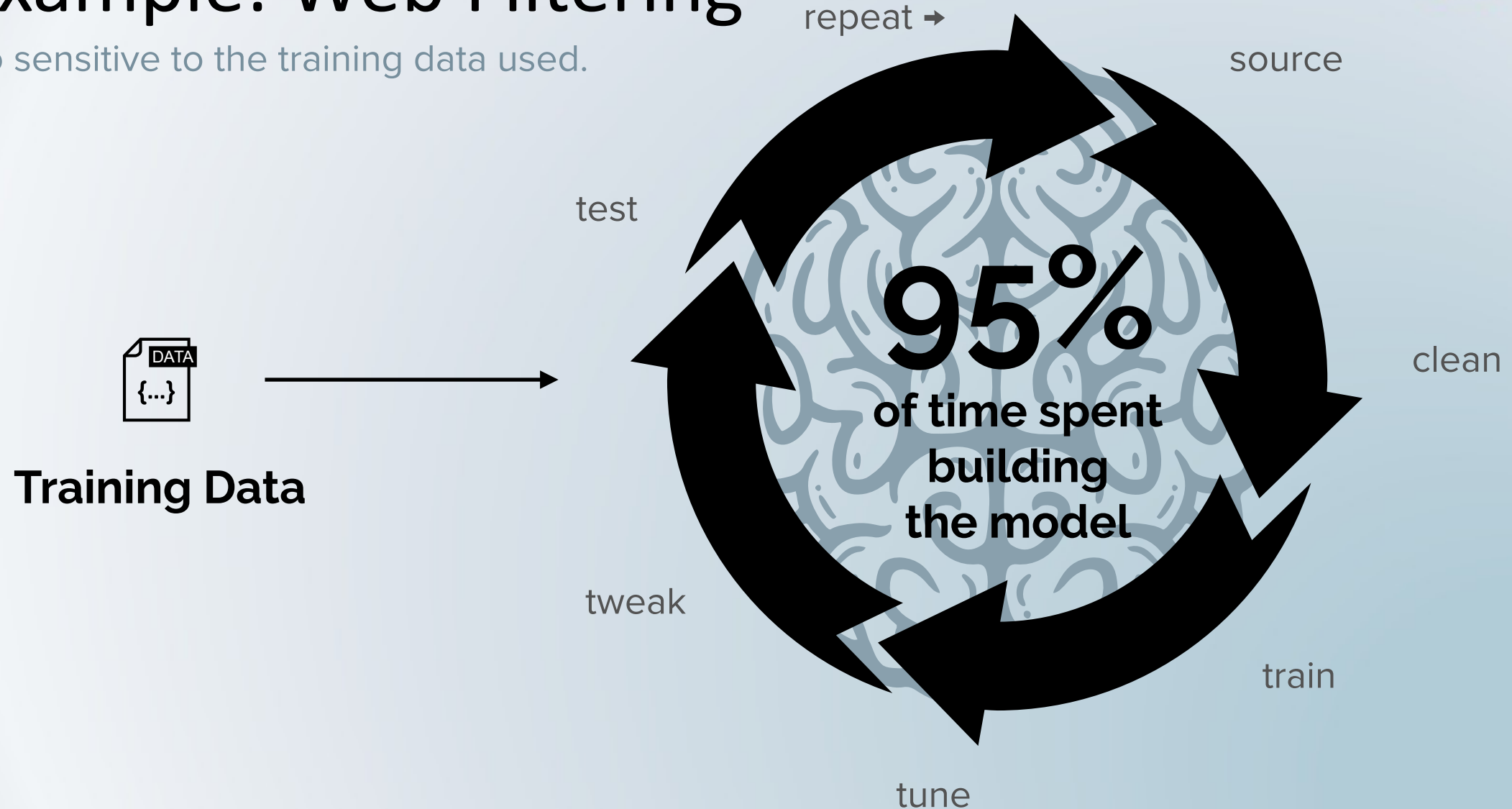
AI Model

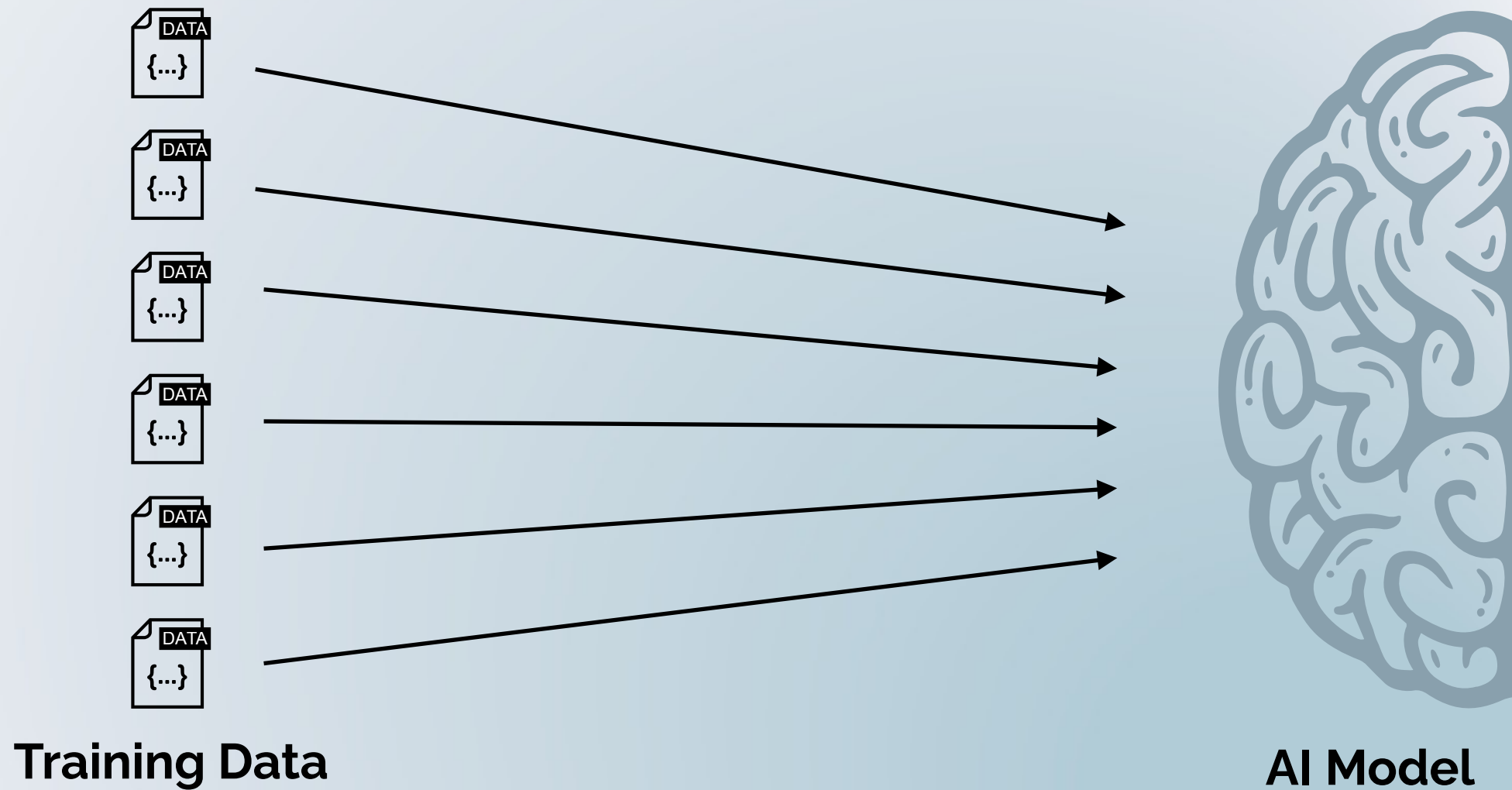
The knowledge base and
decision center of AI

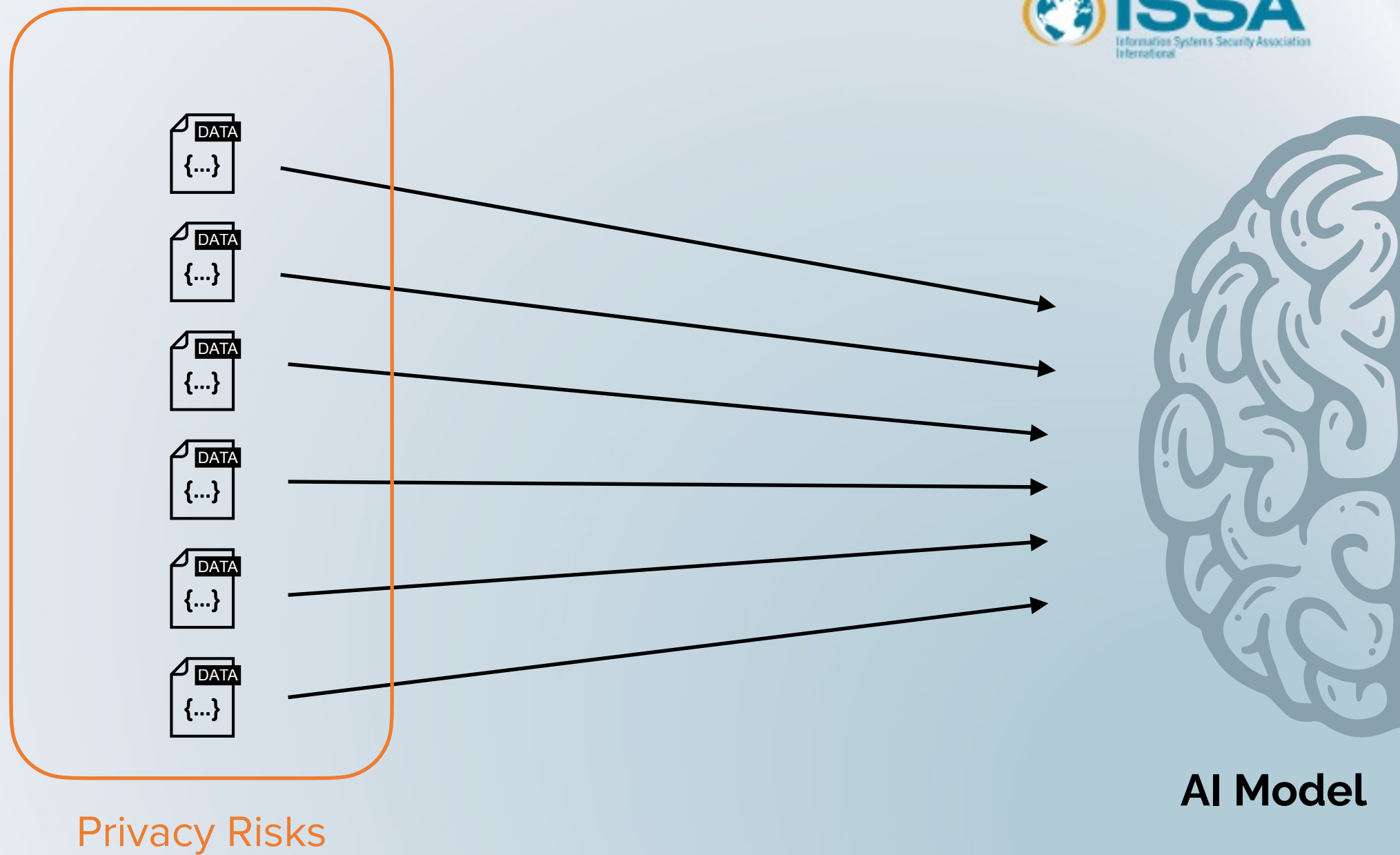


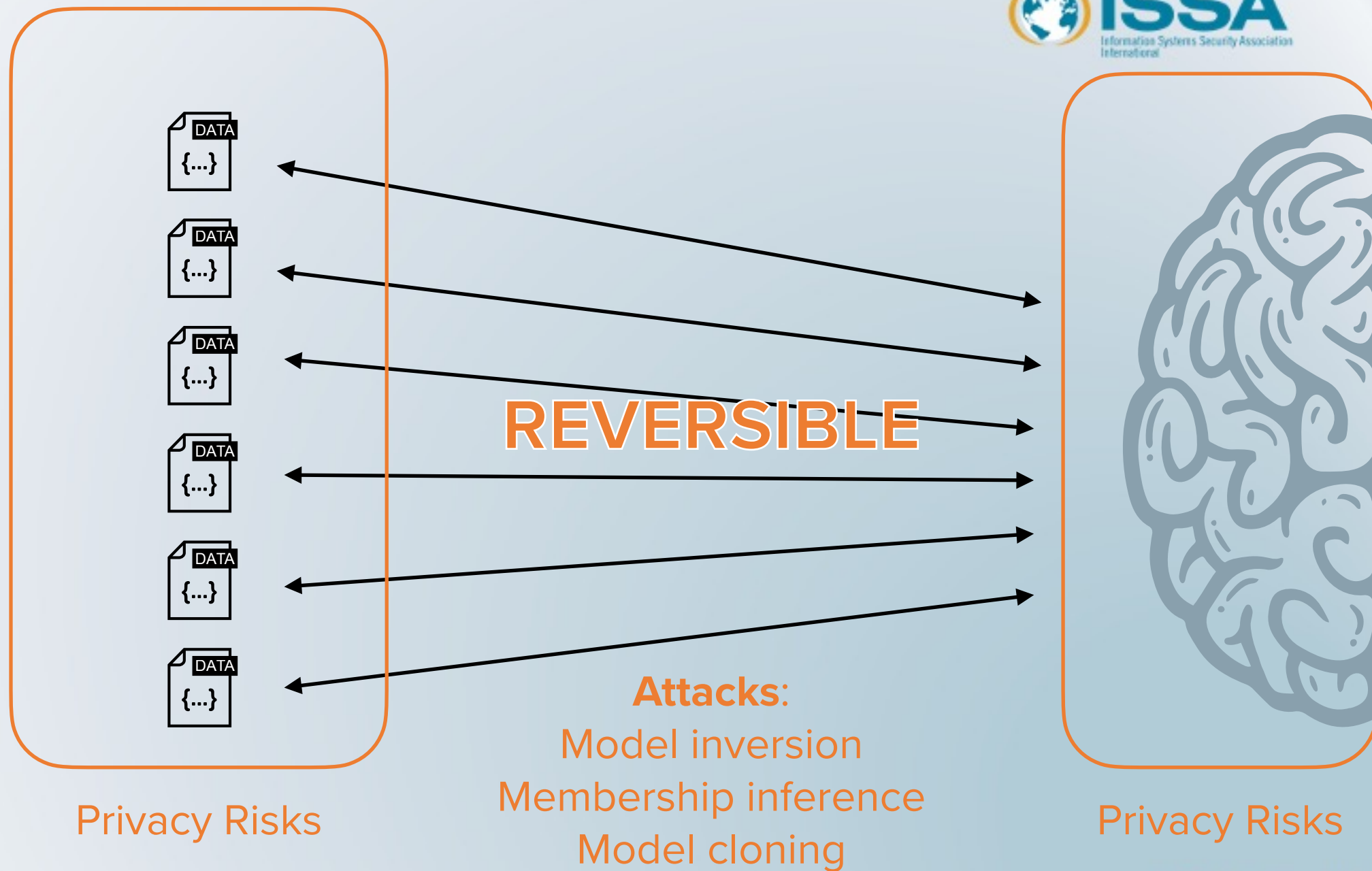
Example: Web Filtering

So sensitive to the training data used.







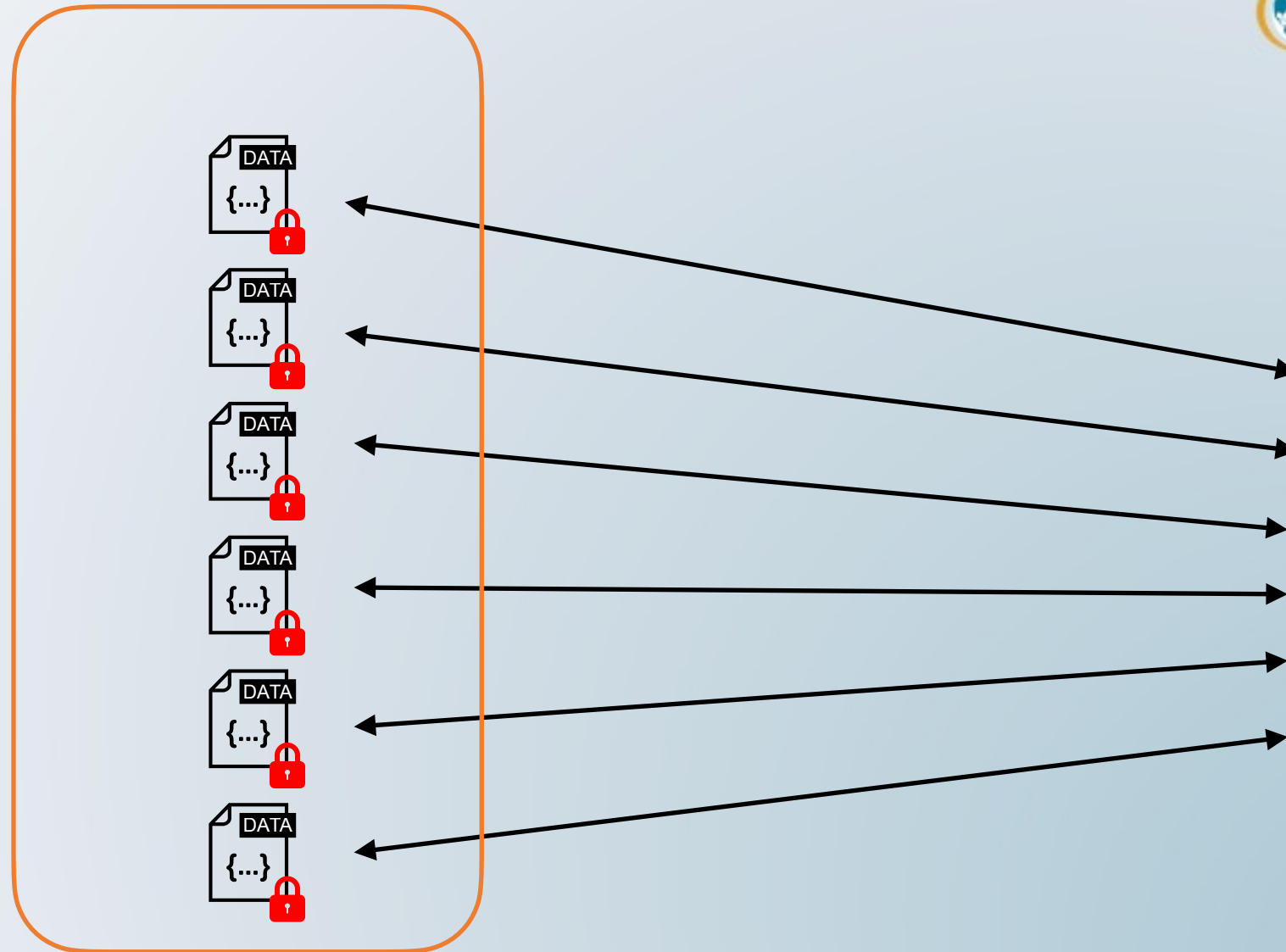


Solution Approaches

- Encryption
- PII redaction
- Anonymization
- Diff. privacy
- Synthetic data

Sample Co's

- Mostly AI
- Synthesis AI
- Thales
- Assembly AI
- Private-AI
- Protopia AI
- DynamoFL



Privacy Risks



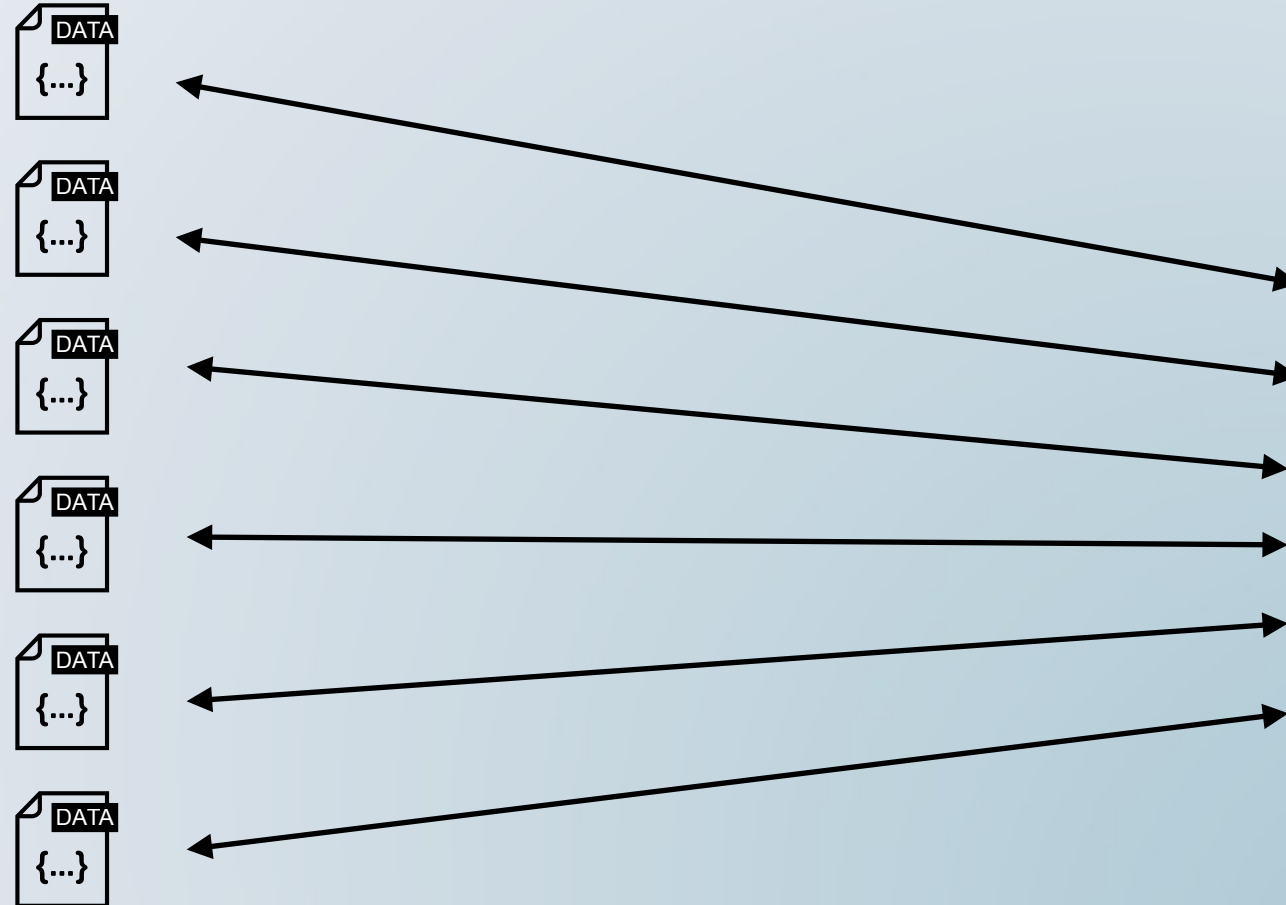
AI Model

Solution Approaches

- Query filtering
- Query rate limiting
- Attack simul
- Compliance reporting
- Code analysis
- PII checking
- Fully homomorphic encryption

Sample Co's

- Robust Intellig.
- Protect AI
- Cranium
- Adversa
- Advai
- Mindgard AI
- Enveil



Privacy Risks

Big Changes

- That was then; this is now.
- New tools and capabilities
- Why build your own models if you don't need to?

Artificial **General** Intelligence (AGI)

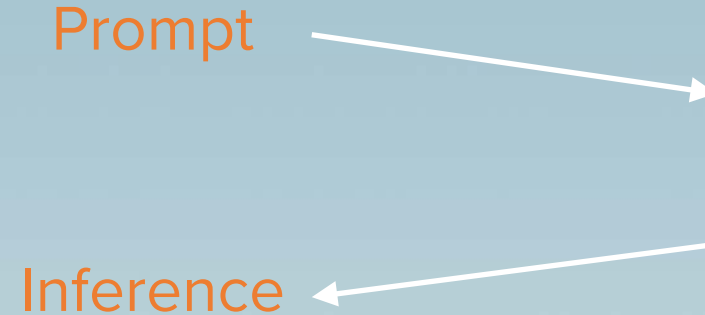
- Large, shared, general-purpose models
 - Large language models (LLMs)
 - GPT3/GPT4
 - Llama2
 - Large vision models



Artificial General Intelligence

- Large, shared, general-purpose models
 - Large language models (LLMs)
 - GPT3/GPT4
 - Llama2
 - Large vision models

Maybe 90% of use cases covered by general purpose models now.



Someone else's AGI model

Artificial General Intelligence

Default (GPT-3.5)

OP

Given the following eleven categories: 1. Adult content, 2. Business content, 3. Tech content, 4. Family content, 5. News, 6. Travel, 7. Sports, 8. Entertainment, 9. Shopping, 10. Health, 11. Finance, use a category to label the following text:

The days of rock-bottom interest rates have long been over — and consumers are facing money dilemmas they haven't had to debate for years.

At its July meeting, the Federal Reserve lifted interest rates by a quarter of a percentage point, bringing its key benchmark interest rate to a new target range of 5.25-5.5 percent. That's the highest level since early 2001.

The category that best fits the following text is "Finance."

Prompt

Inference



LLM

Personalize?

- Our (private) data matters still...



Option 1: Fine Tuning



Add data on top of a pre-trained large model.

Option 2: Add Memory

MODEL



MEMORY

Option 2: Add Memory

MODEL



MEMORY

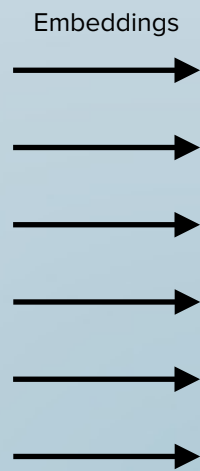
Remembered snapshots of documents, images, chats, histories, faces, voices, etc.



Confidential Data



"Embedding Model"



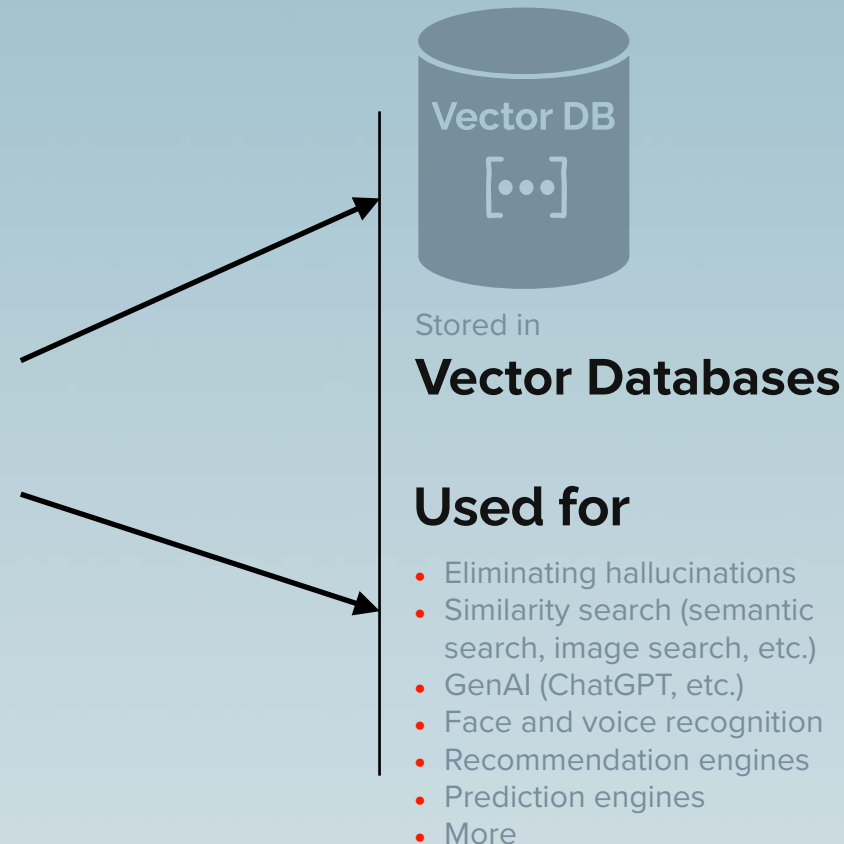
Memory



AI Memory

High fidelity inferences stored for later use.

aka Vector Embeddings

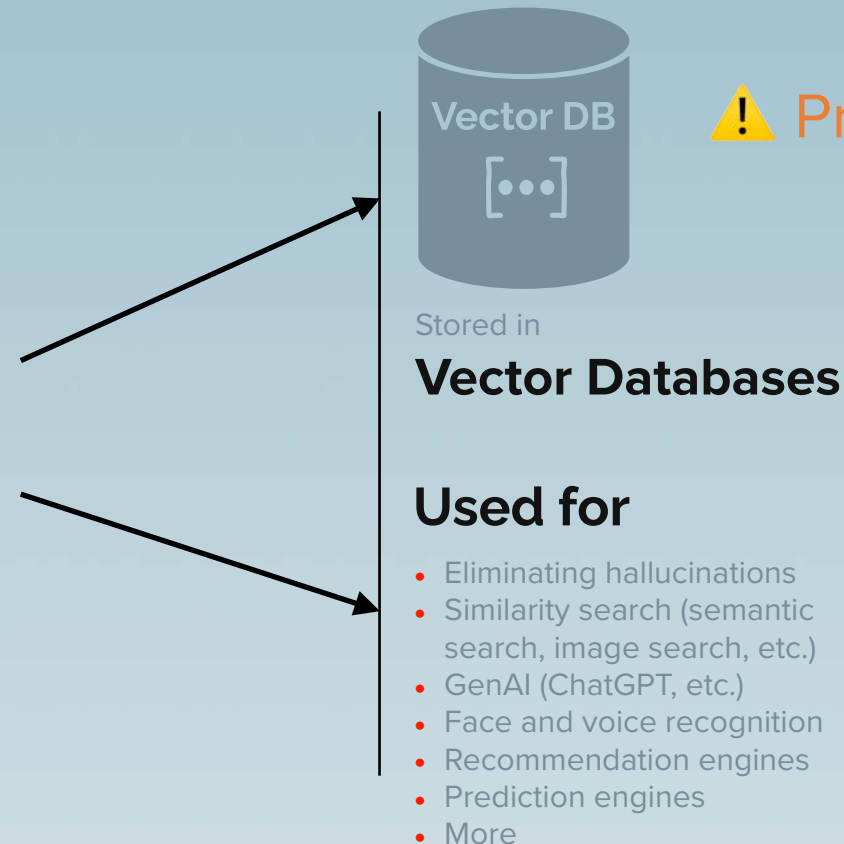


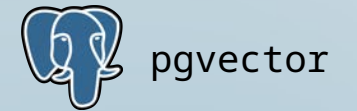
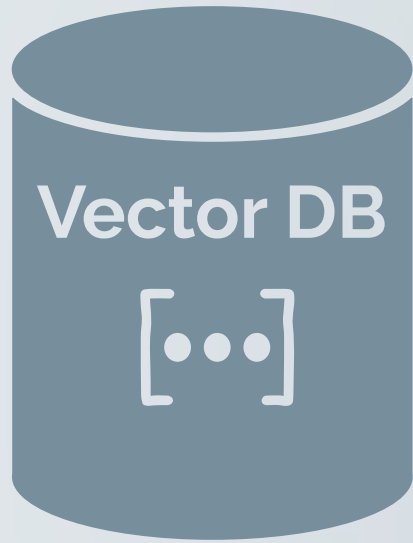


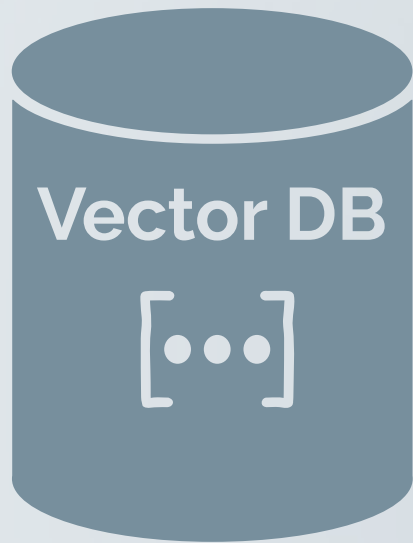
AI Memory

High fidelity inferences stored for later use.

aka Vector Embeddings







Weaviate



drant



Chroma



vespa



Pinecone

KX



milvus



LanceDB



ISSA
Information Systems Security Association
International

OpenSearch

redis

elastic



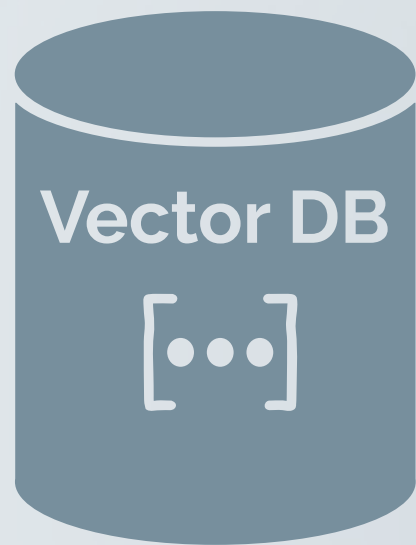
pgvector

SQLite-vss

MongoDB.

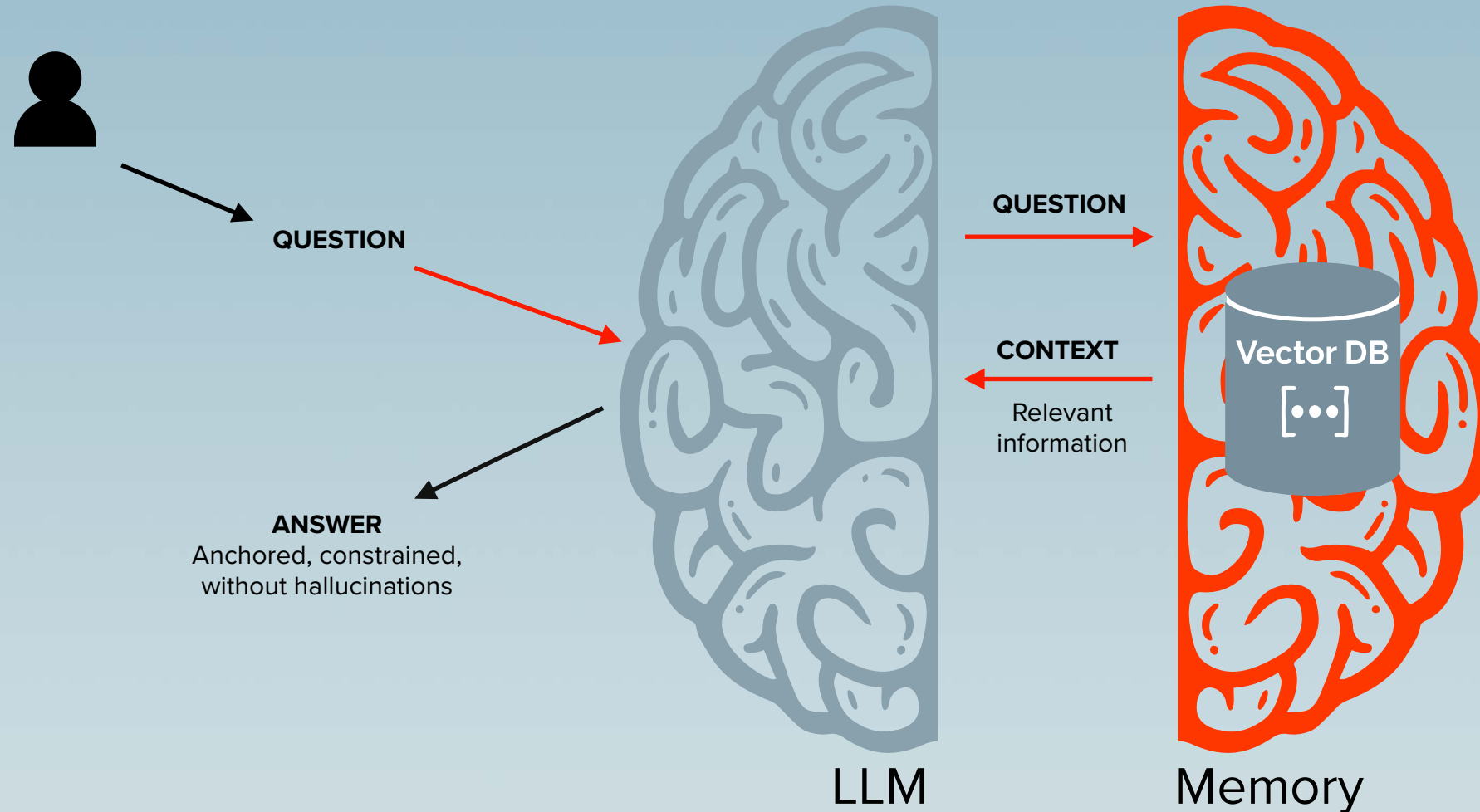
DEDICATED

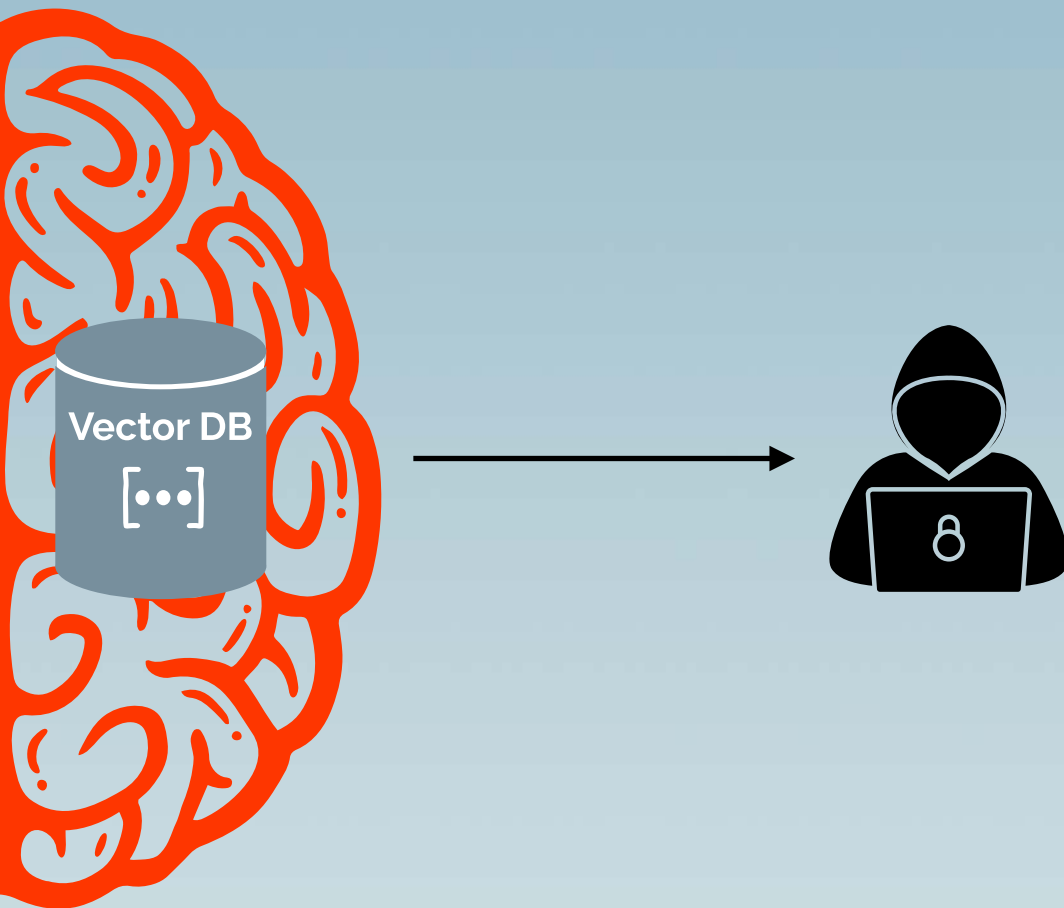
#ISSAPrivacySIG



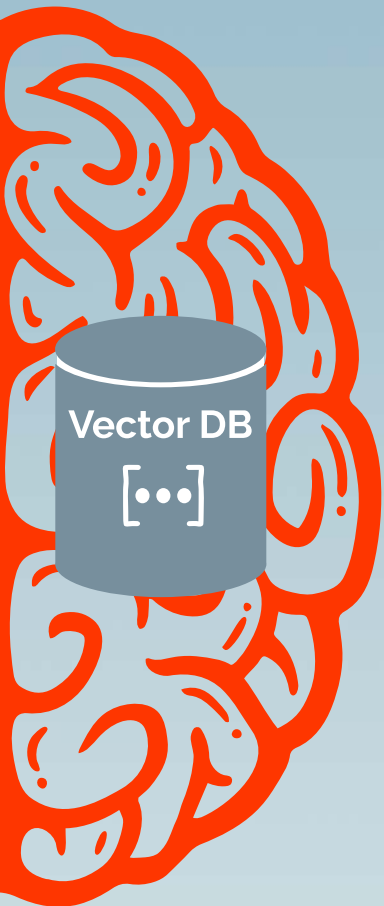
AUGMENTED

Retrieval Augmented Generation (RAG)





- **Embedding inversion attack**
- Membership inference attack
- Better results finding sensitive info
- Source data and metadata



! Privacy Risks

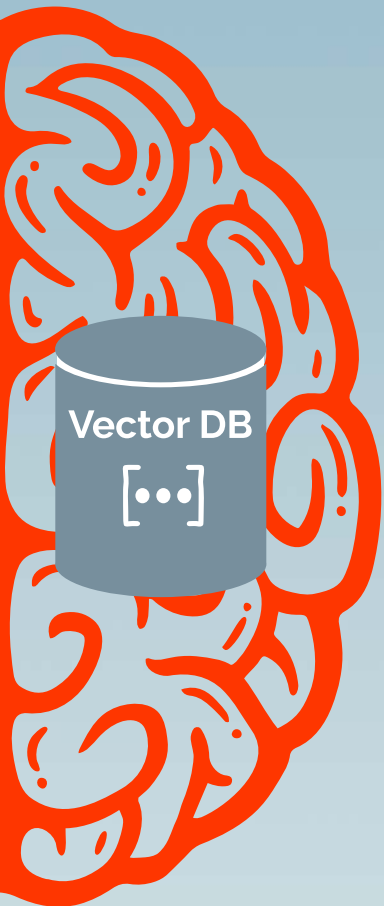


- **Embedding inversion attack**
- Membership inference attack
- Better results finding sensitive info
- Source data and metadata

Embeddings are equivalent to their source data.

Memory

So Much Private Data



Memory

[0.123, -0.987, 1.234, ..., 0.004]



```
{  
  name: "Patrick Walsh",  
  gender: "Male",  
  title: "CEO",  
  company: "IronCore Labs",  
  birthday: 19xx-xx-xx  
}
```

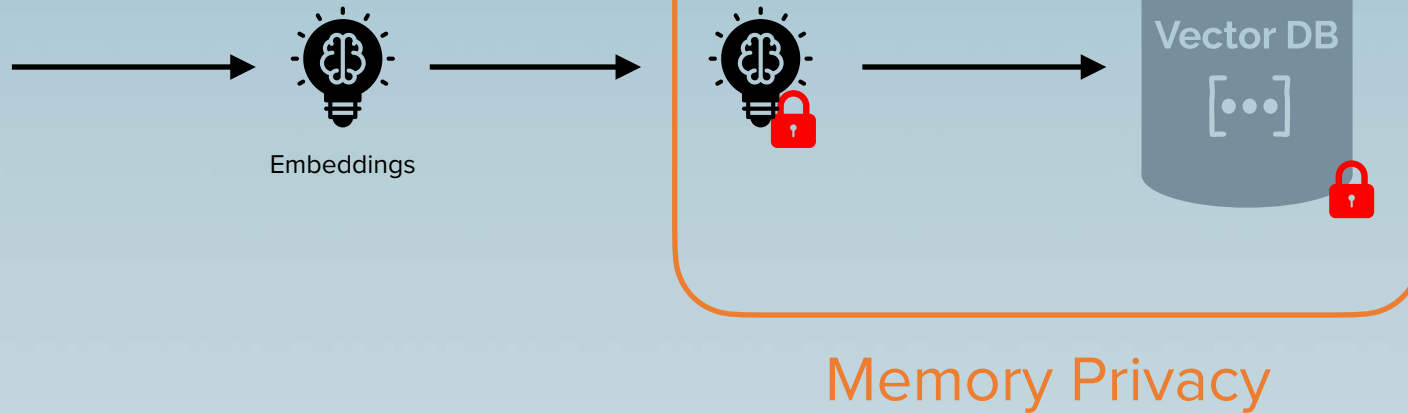
! Privacy Risks

Artificial **General** Intelligence

- New AI use cases
- New approaches to old AI use cases
- New AI tech stacks
- New AI data



AI Memory Protection



Approaches

- Property-preserving, data-in-use encryption

Sample Co's

- IronCore Labs

Pipelines, Services, Deployment Options



Pipelines, Services, Deployment Options



Pipelines, Services, Deployment Options



Managed service



Cloud infra



On-prem

 **Easiest**

Most Complex

Less Private

More Private 

Pipelines, Services, Deployment Options



Managed service



Cloud infra



On-prem

 **Easiest**



Most Complex



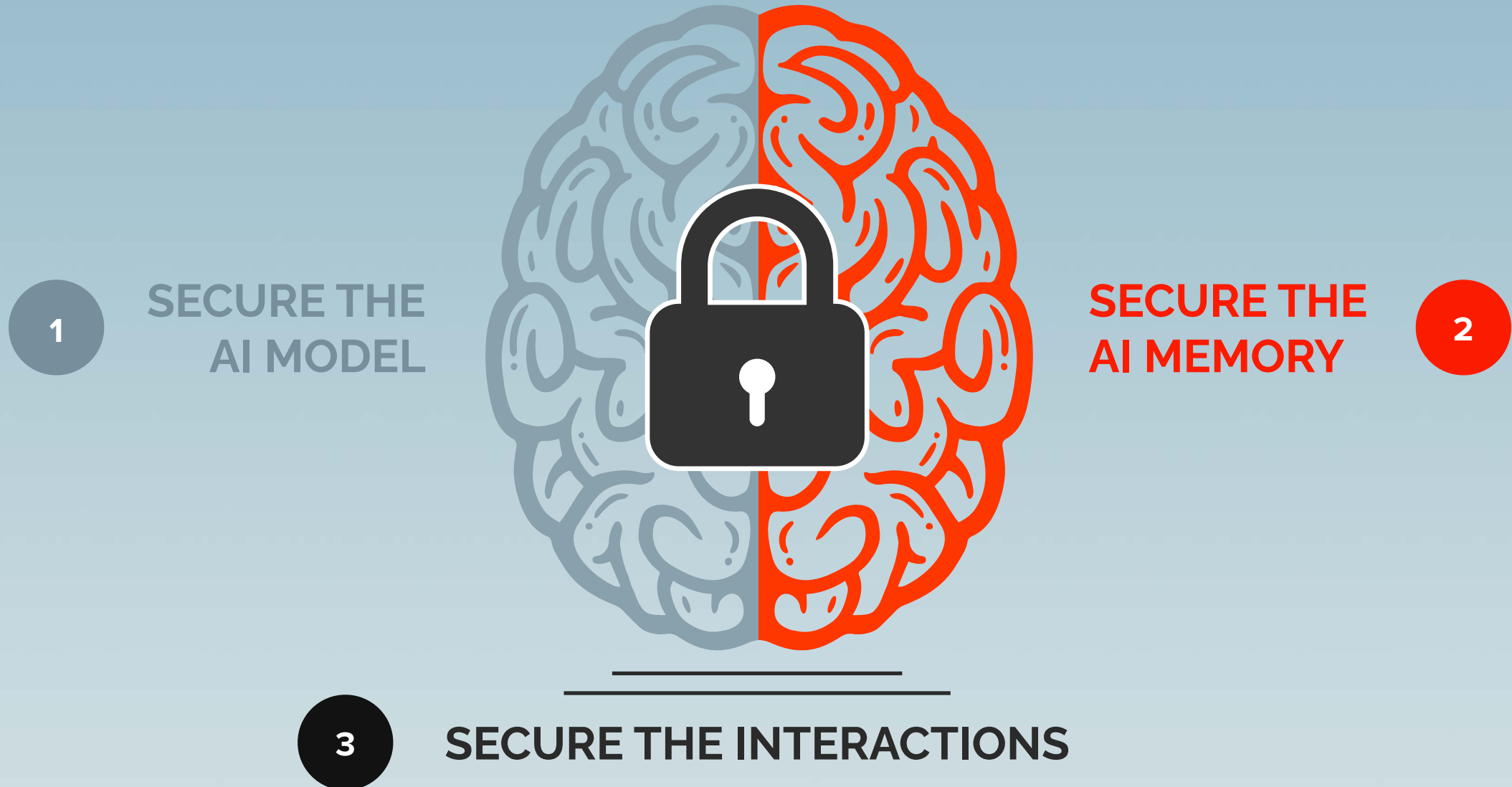
Encrypted

Most Private 



Private

How To Secure AI Systems



Further materials

- Video tutorials at <https://youtube.com/@ironcorelabs>
 - More aspects of the Security of AI Landscape
 - More in depth on Protecting Sensitive Data in GenAI Systems
- Explainers at <https://ironcorelabs.com/resources/>
 - Details on embedding inversion attacks
 - Go deeper on data protection concepts
- Reach out to me directly:
 - patrick.walsh@ironcorelabs.com / @zmre / @zmre@mastodon.social



ISSA
Information Systems Security Association
International

www.issa.org

QUESTIONS?