treelogic
A TREE TECHNOLOGY S.A. BRAND

IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA PROTECCIÓN DE DATOS: POLÍTICAS Y DESAFÍOS

We Deal With Data

José Barranquero Tolosa

Sr. Solution Architect – Treelogic
jose.barranquero@treelogic.com

ANTECEDENTES

IECPI INCIBE

La IECPI es el resultado de la reflexión realizada al respecto y de su consenso con otros actores públicos del ecosistema de la ciberseguridad y pretende realizar un conjunto de **actuaciones dirigidas a impulsar la I+D+i y la creación de productos y soluciones en el ámbito de la ciberseguridad**.

IECPI RETOS SELECCIONADOS

treelogic

PRIMERA CONVOCATORIA

- Gestión de identidades
- Sistemas para el seguimiento de criptotransacciones
- Protección de datos e información

SEGUNDA CONVOCATORIA

- Soluciones para la seguridad de los datos y prevenir su uso malicioso
- Sistemas innovadores para la evaluación, cumplimiento normativo y certificación

TERCERA CONVOCATORIA

- Ciberseguridad en el vehículo conectado

PRIMERA CONVOCATORIA

RETO 4

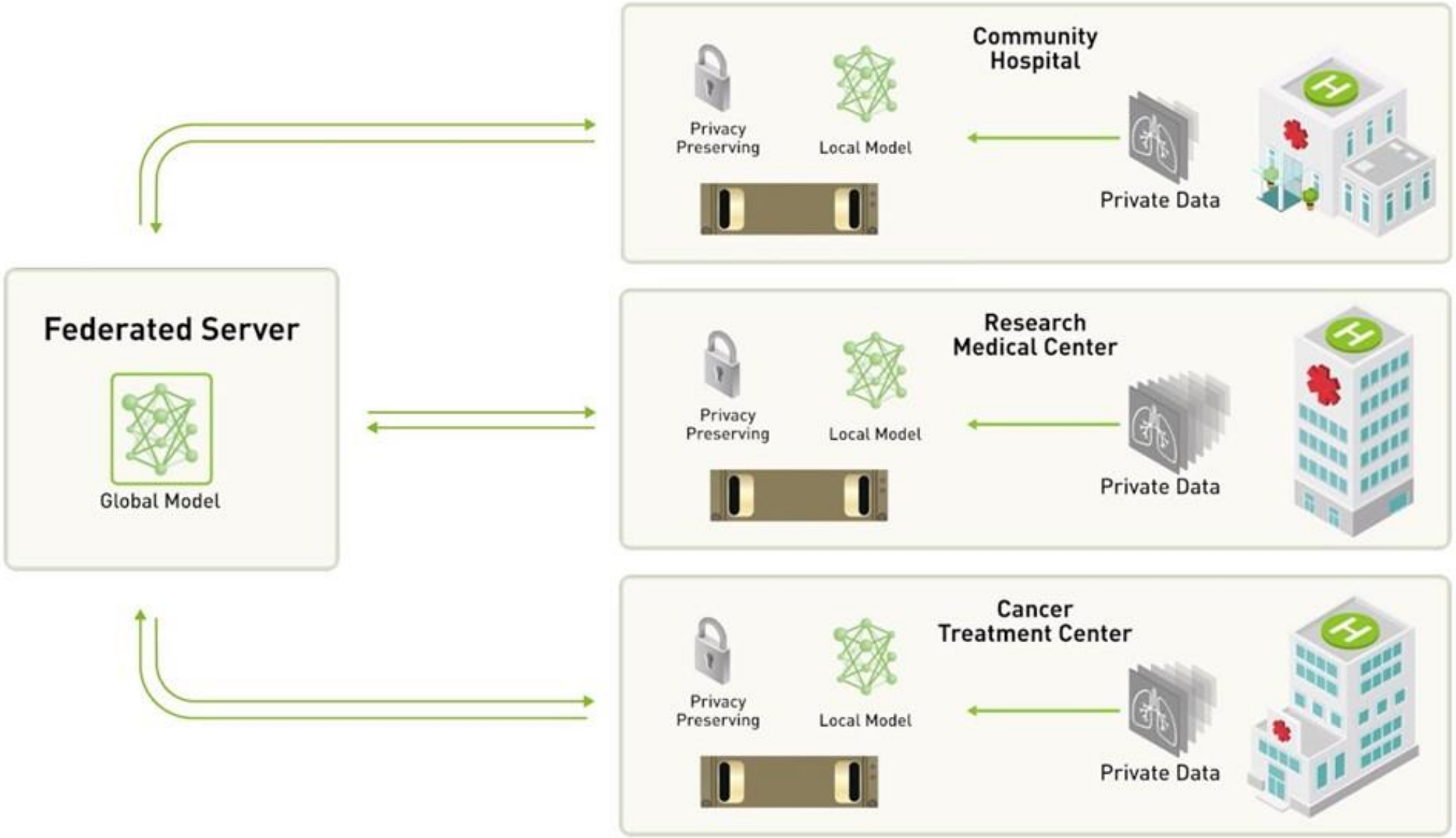
Protección de datos e información.

SOLUCIÓN

APRENDE Aseguramiento de la protección y la privacidad de datos en entornos federados.

Objetivo:	Desarrollar y validar un esquema de **aprendizaje automático federado** (FML) con diferentes niveles de seguridad y modos de operación que pueda adaptarse a las particularidades de diferentes casos de uso.
Presupuesto:	1.701.459,00 €
Duración:	36 meses.

Federated Machine Learning (FML)



SEGUNDA CONVOCATORIA

RETO 4

Evaluación, cumplimiento normativo y certificación.

SOLUCIÓN

SECAP Sistema para la evaluación y el cumplimiento normativo en la Administración Pública.

Objetivo:	Creación de una solución innovadora en el ámbito de las Administraciones Públicas (AA.PP.), que brinde soporte y apoyo automatizado para evaluar el cumplimiento de la normativa de **protección de datos** y otras que puedan resultar aplicables, así como para prevenir y detectar posibles **filtraciones**.
Presupuesto:	1.335.852,2 €
Duración:	36 meses.

Anonimización de datos

Fases del proceso de anonimización

1. Definición del equipo de trabajo
2. Evaluación de riesgos de reidentificación
3. Definición de objetivos y finalidad de la información anonimizada
4. Viabilidad del proceso
5. **Preanonimización:** definición de variables de identificación
6. Eliminación/reducción de variables
7. Selección de técnicas de anonimización
   - Algoritmo de hash
   - Algoritmo de cifrado
   - Sello de tiempo
   - Capas de anonimización
   - Perturbación de datos
   - Reducción de datos
8. Segregación de la información
9. Proyecto Piloto
10. Anonimización
11. Formación e información al personal implicado
12. Garantías jurídicas
13. Auditoría del proceso de anonimización

Singularización (singling out): Identificar a un individuo concreto.

Vinculabilidad (linkability): Asociar uno o varios registros de un individuo en uno o varios conjuntos de datos.

Inferencia (inference): Deducir atributos de un individuo a partir de otros atributos.

SEGUNDA CONVOCATORIA

treelogic

RETO 3

Seguridad de los datos y prevención de su uso malicioso.

SOLUCIÓN

CONFIA Herramienta para la detección de datos sesgados, erróneos o fraudulentos mediante inteligencia artificial.

Objetivo:	Desarrollar una solución que aumente tanto la **seguridad** como la **confiabilidad de los datos** almacenados en un *Data Lake*, que, posteriormente, serán utilizados por diferentes modelos y procesos de inteligencia artificial.
Presupuesto:	1.502.661,77 €
Duración:	36 meses.

Gobierno de datos para IA

Data is an asset

Definir los propietarios

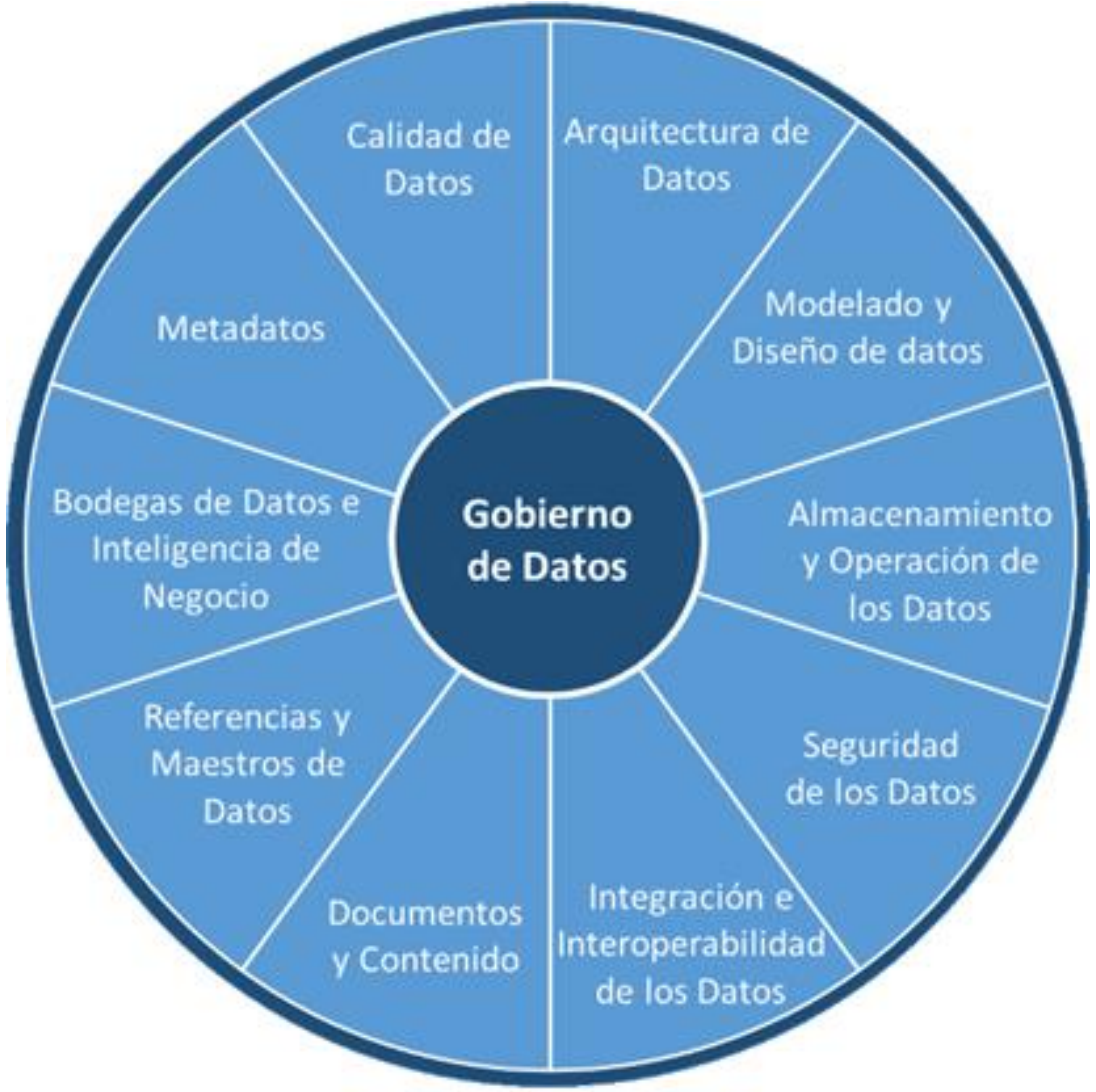
Políticas y procesos

Normas y procedimientos

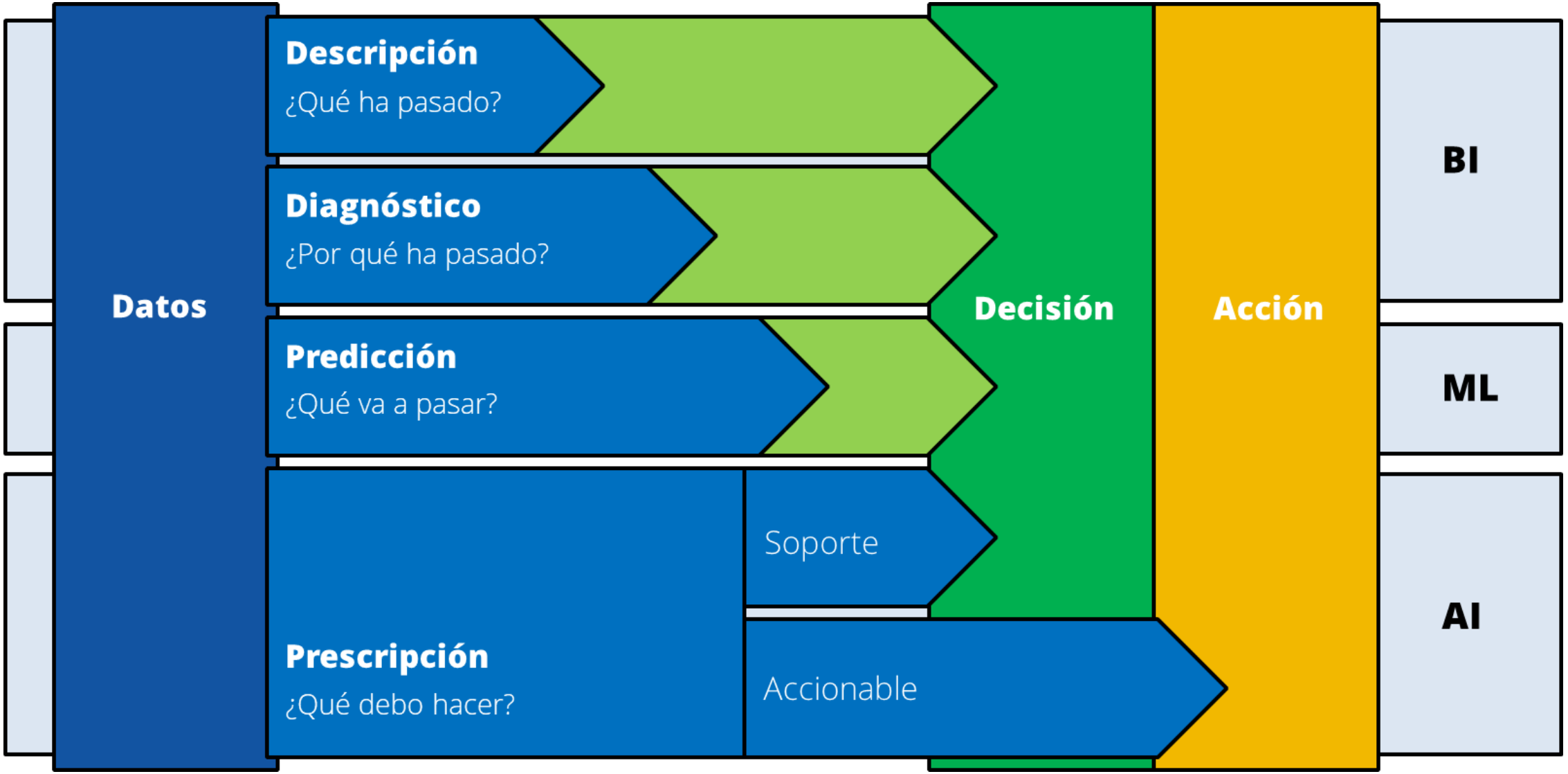
Control y auditoría

Ataques adversarios

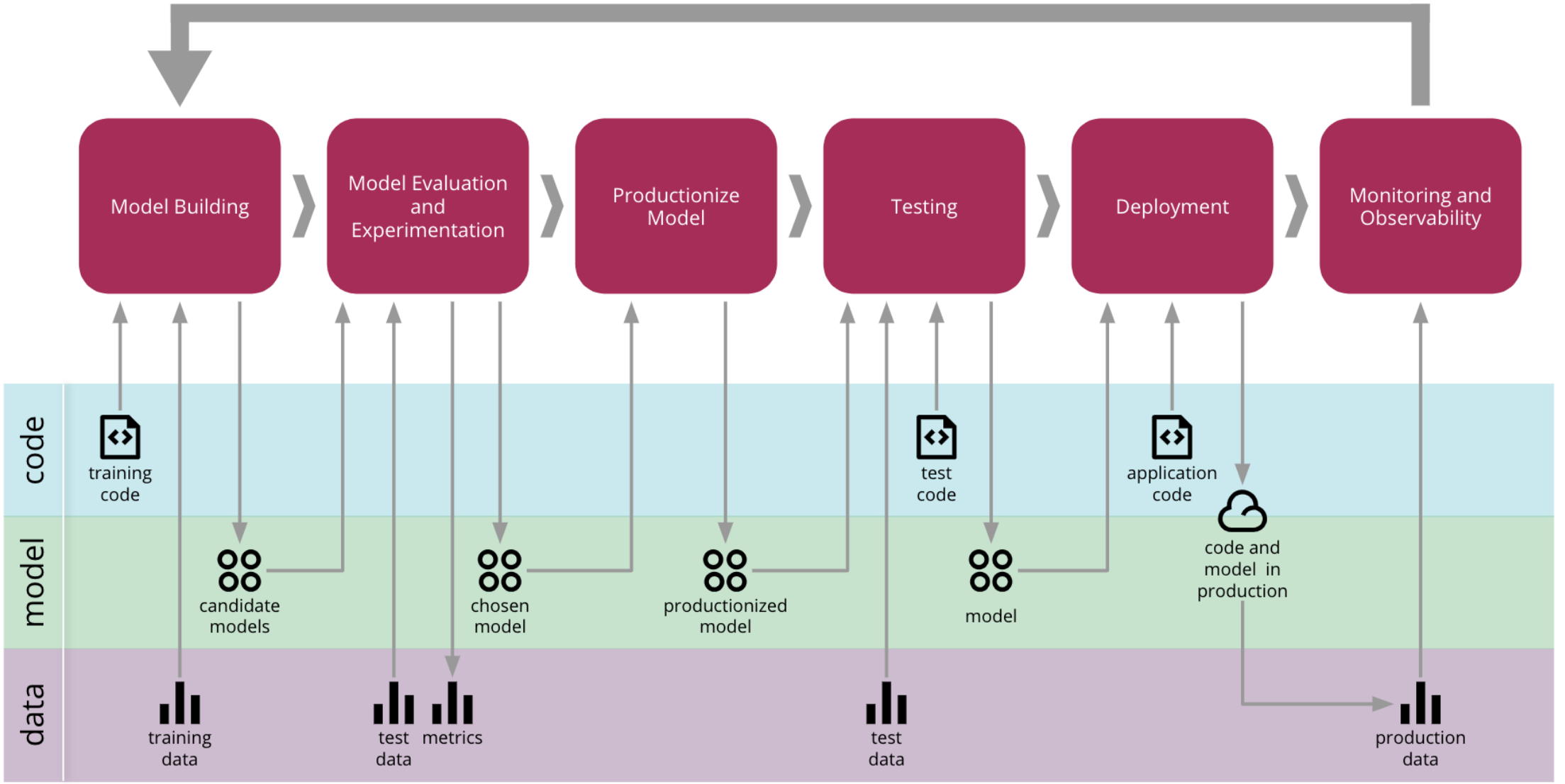
Mitigación de sesgos



Data Science



Productización de modelos (MLOps)



https://martinfowler.com/articles/cd4ml.html

LARGE LENGUAGE MODELS (LLM)

Privacidad y riesgos LLMs

treelogic

Tres riesgos principales a tener en cuenta:

- Compartir información sensible con un proveedor externo
- Seguridad/privacidad del propio modelo LLM
- Acceso no autorizado a los datos de entrenamiento

https://www.forbes.com/sites/forbestechcouncil/2023/11/06/how-to-mitigate-the-enterprise-security-risks-of-llms/

https://machine-learning-made-simple.medium.com/7-methods-to-secure-llm-apps-from-prompt-injections-and-jailbreaks-11987b274012

Privacidad y riesgos LLMs

Extracting Training Data from ChatGPT

AUTHORS

Milad Nasr*[1], Nicholas Carlini*[1], Jon Hayase[1,2], Matthew Jagielski[1], A. Feder Cooper[3], Daphne Ippolito[1,4], Christopher A. Choquette-Choo[1], Eric Wallace[5], Florian Tramèr[6], Katherine Lee+[1,3]

[1]Google DeepMind, [2] University of Washington, [3]Cornell, [4]CMU, [5]UC Berkeley, [6]ETH Zurich. * Joint first author, +Senior author.

PUBLISHED

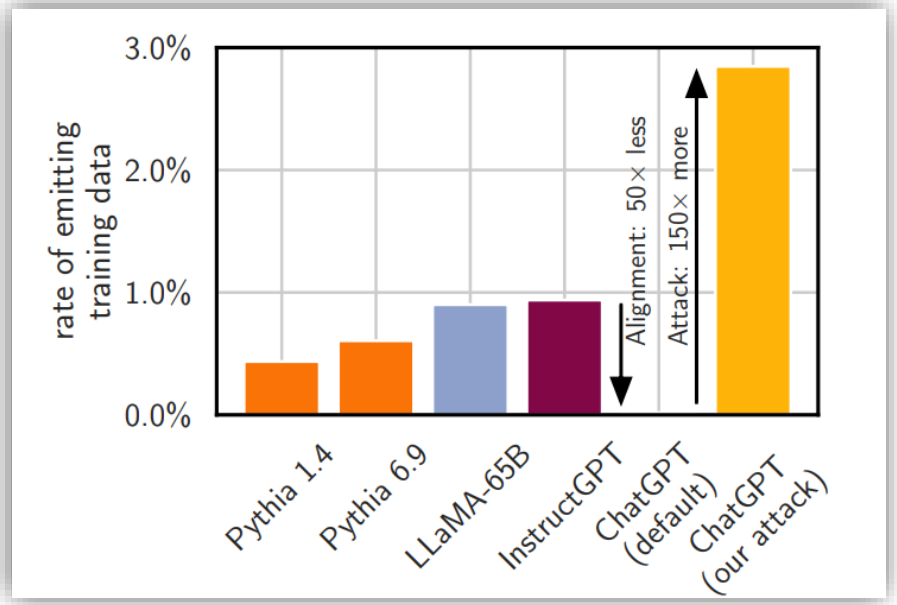
November 28, 2023

READ:

[arxiv]

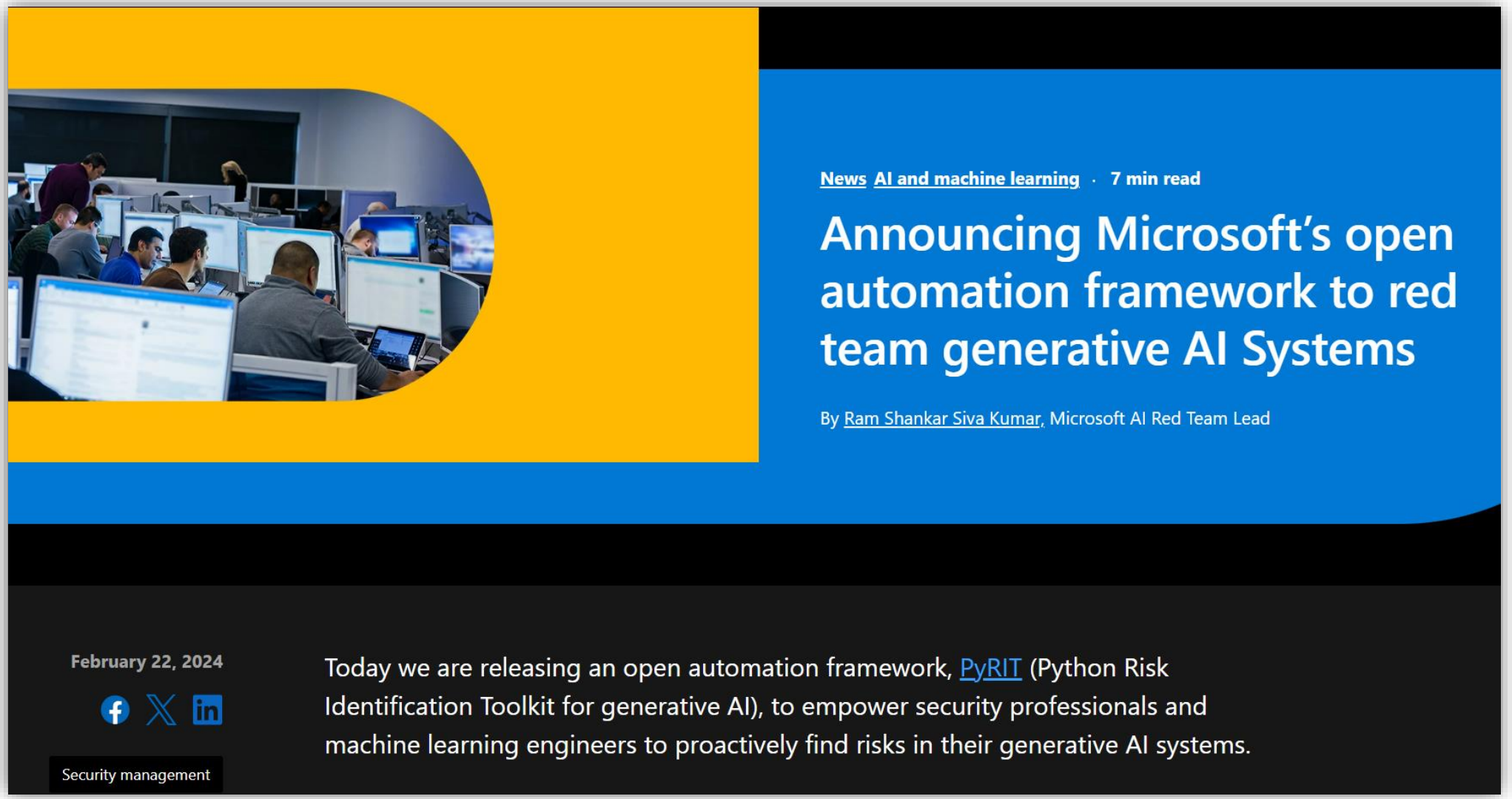
We have just released a paper that allows us to extract several megabytes of ChatGPT's training data for about two hundred dollars. (Language models, like ChatGPT, are trained on data taken from the public internet. Our attack shows that, by querying the model, we can actually extract some of the exact data it was trained on.) We estimate that it would be possible to extract ~a gigabyte of ChatGPT's training dataset from the model by spending more money querying the model.

Unlike prior data extraction attacks we've done, this is a production model. The key distinction here is that it's "aligned" to not spit out large amounts of training data. But, by developing an attack, we can do exactly this.



https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html
https://medium.datadriveninvestor.com/google-extracted-chatgpts-training-data-using-a-silly-trick-5544b1dada71

Privacidad y riesgos LLMs



News AI and machine learning · 7 min read

Announcing Microsoft's open automation framework to red team generative AI Systems

By Ram Shankar Siva Kumar, Microsoft AI Red Team Lead

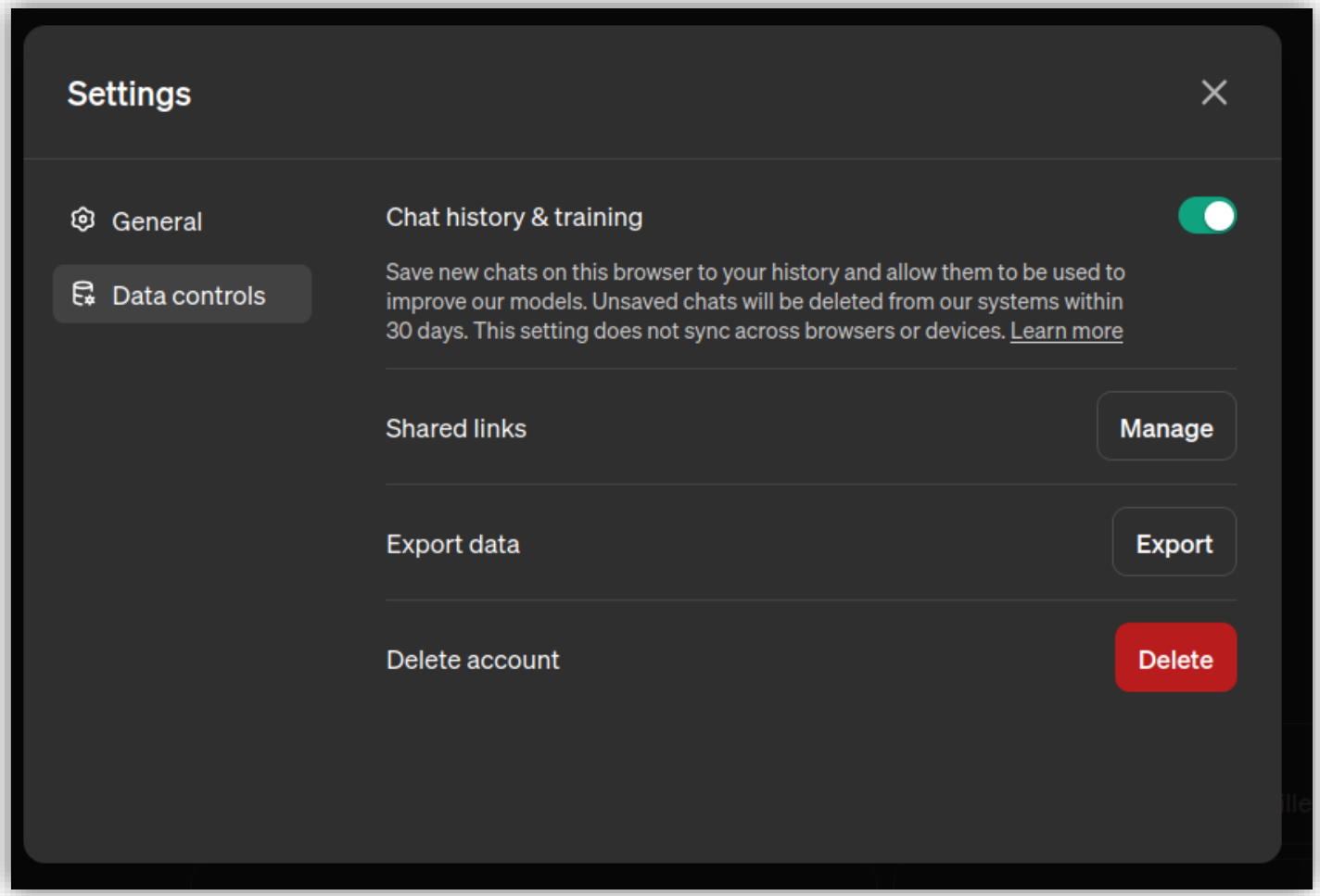
February 22, 2024

Today we are releasing an open automation framework, PyRIT (Python Risk Identification Toolkit for generative AI), to empower security professionals and machine learning engineers to proactively find risks in their generative AI systems.

Security management

https://www.microsoft.com/en-us/security/blog/2024/02/22/announcing-microsofts-open-automation-framework-to-red-team-generative-ai-systems/

Privacidad y riesgos LLMs



https://openai.com/enterprise-privacy
https://learn.microsoft.com/en-us/azure/ai-services/openai/chatgpt-quickstart

Privacidad y riesgos LLMs

Large language model

Llama 2: open source, free for research and commercial use

We're unlocking the power of these large language models. Our latest version of Llama – Llama 2 – is now accessible to individuals, creators, researchers, and businesses so they can experiment, innovate, and scale their ideas responsibly.

Download the model

https://llama.meta.com/llama2
https://mistral.ai/news/mixtral-of-experts/

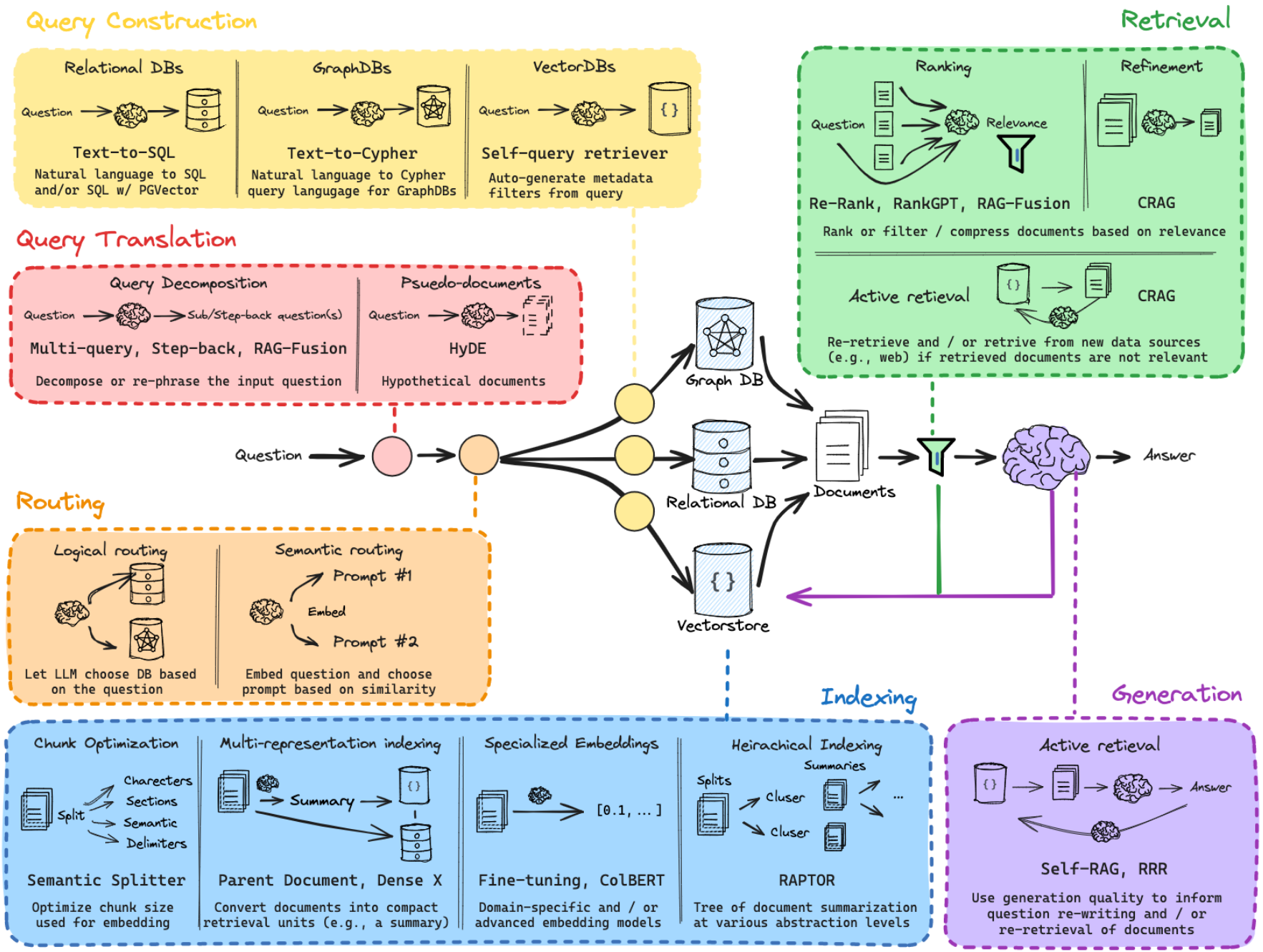
Privacidad y riesgos LLMs

treelogic

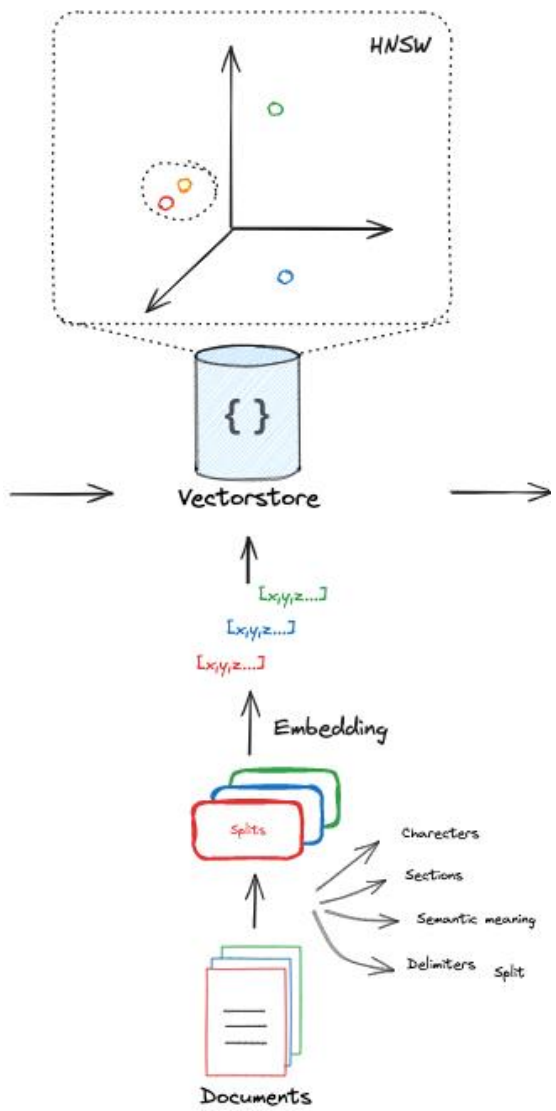
RAG (Retrieval Augmented Generation)

Have you ever asked an AI language model like ChatGPT about the latest developments on a certain topic, only to receive this response: *'I apologize, but as an AI language model, I do not have real-time data or access to current news. My knowledge was last updated in September 2021, and I cannot provide you with the latest developments on the topic beyond that point.'*? If so, you've encountered a fundamental limitation of large language models. They are, in essence, time capsules of knowledge, frozen at the point of their last training. They can't 'learn' and 'remember' new information without undergoing a retraining process, which is both computationally intensive and time-consuming.

https://medium.com/@amodwrites/understanding-retrieval-augmented-generation-a-simple-guide-d638ac92c123
https://github.com/langchain-ai/rag-from-scratch


Query Construction
Relational DBs
Question
Text-to-SQL
Natural language to SQL and/or SQL w/ PGVector
GraphDBs
Question
Text-to-Cypher
Natural language to Cypher query langugage for GraphDBs
VectorDBs
Question
{ }
Self-query retriever
Auto-generate metadata filters from query
Retrieval
Ranking
Question
Relevance
Re-Rank, RankGPT, RAG-Fusion
Refinement
CRAG
Rank or filter / compress documents based on relevance
Active retieval
{ }
CRAG
Re-retrieve and / or retrive from new data sources (e.g., web) if retrieved documents are not relevant
Query Translation
Query Decomposition
Question
Sub/Step-back question(s)
Multi-query, Step-back, RAG-Fusion
Decompose or re-phrase the input question
Psuedo-documents
Question
HyDE
Hypothetical documents
Graph DB
Relational DB
{ }
Vectorstore
Question
Documents
Answer
Routing
Logical routing
Let LLM choose DB based on the question
Semantic routing
Prompt #1
Embed
Prompt #2
Embed question and choose prompt based on similarity
Indexing
Chunk Optimization
Split
Charecters
Sections
Semantic
Delimiters
Semantic Splitter
Optimize chunk size used for embedding
Multi-representation indexing
Summary
{ }
Parent Document, Dense X
Convert documents into compact retrieval units (e.g., a summary)
Specialized Embeddings
[0.1, ...]
Fine-tuning, ColBERT
Domain-specific and / or advanced embedding models
Heirachical Indexing
Summaries
Splits
Cluser
Cluser
...
RAPTOR
Tree of document summarization at various abstraction levels
Generation
Active retieval
{ }
Answer
Self-RAG, RRR
Use generation quality to inform question re-writing and / or re-retrieval of documents

treelogic


HNSW
{ }
Vectorstore
[x,y,z...]
[x,y,z...]
[x,y,z...]
Embedding
Splits
Charecters
Sections
Semantic meaning
Delimiters
Split
Documents

treelogic
A TREE TECHNOLOGY S.A. BRAND

IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA PROTECCIÓN DE DATOS: POLÍTICAS Y DESAFÍOS

We Deal With Data

José Barranquero Tolosa

Sr. Solution Architect – Treelogic
jose.barranquero@treelogic.com