Red Hat Summit

Connect

Integración de tareas de calidad de datos con IA Generativa en MLOps

Marlon Cárdenas, PhD.

Área de Data & IA

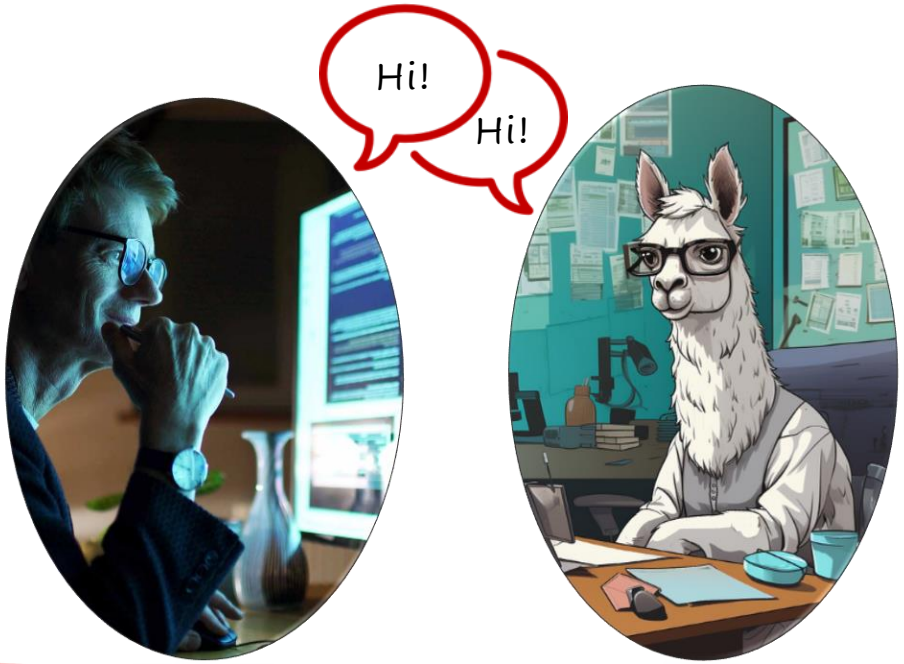
sopra steria

Red Hat

Agenda

- La Era de “Conversar” con los datos
- **Apuesta de la compañía**: Plataforma Inteligente
- **Nuevo objetivo**: usar, entrenar y evaluar LLMs
- **Nuevos retos**: generar datos sintéticos en contextos sensibles
- **Discusión**: problemas con el uso de datos sintéticos

La Era de “conversar” con los datos



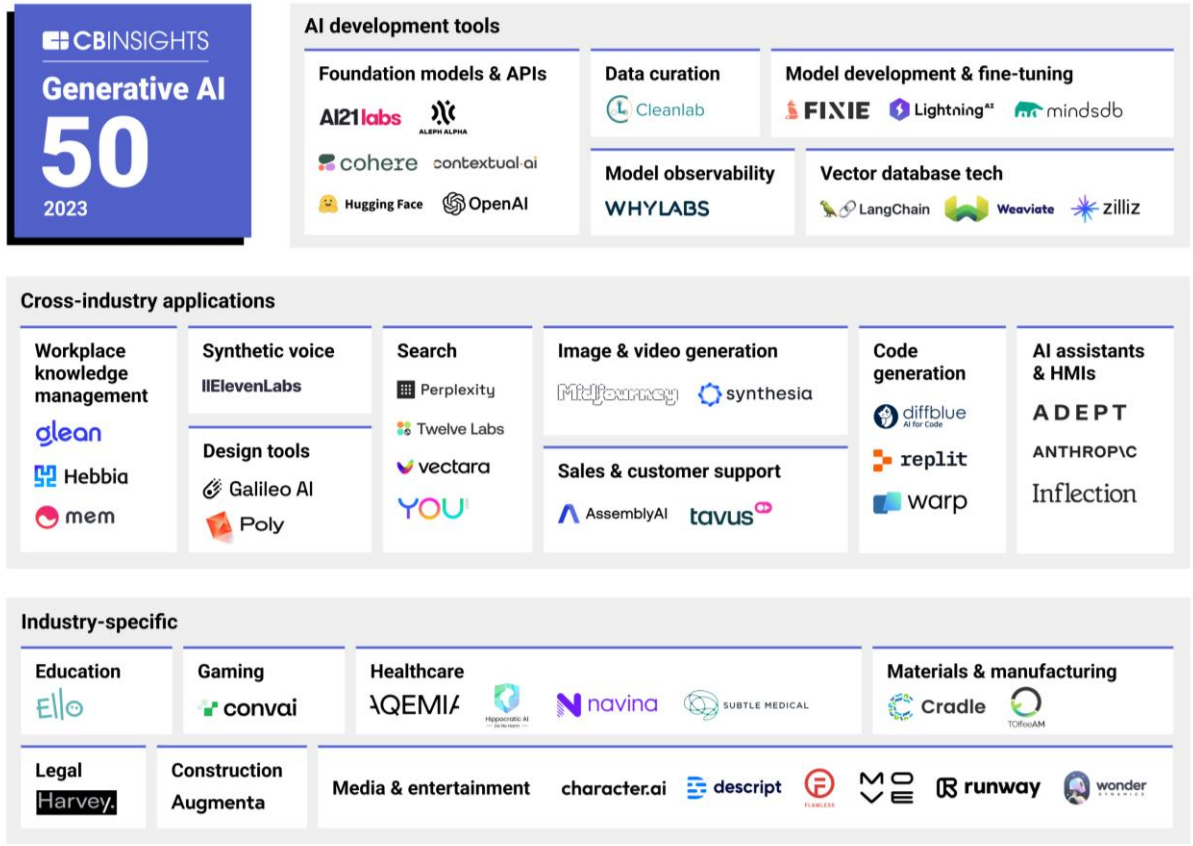


Imagen “Llama” extraída de ejemplos de en midjourney.com/

Red Hat

Nuestra apuesta

Plataforma Inteligente

- Buscamos que los **prototipos de IA y sus resultados prometedores evolucionen**.
- El paso hacia la escala de **soluciones de IA industriales y confiables** es esencial para ofrecer valor a nuestros clientes y convertir la IA en una ventaja competitiva.
- Sopra Steria ha desarrollado la **Plataforma InnerData AI (basada en Red Hat OpenShift y el operador Open Data Hub)**, para **analistas** y **científicos de datos** capaces de producir y desplegar eficientemente modelos de IA.

- Facilitar el acceso y uso sencillo de un entorno completo orientado a datos e IA, desde la **ingestión de datos hasta la exposición compatible con la industria**.
- Servir como una base de software para una **arquitectura genérica de datahub automatizado** que permita la rápida realización de pruebas de concepto (POV) y servir como un núcleo básico para proyectos clave de clientes.
- Facilitar la creación progresiva de un **marketplace** de modelos de IA.

Cookbooks

Recetas digitales para resolver problemas específicos de IA con un ejemplo empresarial

Model Zoo

Modelos preentrenados con Datasets orientados a negocios

Assets & Tools

Mostrar proyectos de IA realizados y permitir la capitalización de nuestros éxitos.

Red Hat

OPEN DATA HUB
AI Platform powered by Open Source

Reference Architecture for AI on OpenShift

sopra steria

Confiance ai

Artificial Intelligence & Machine Learning

Model Training
Tensorflow Training (TFjob)
PyTorch Training
Spark *Katib*

Model Serving
Seldon Core
Tensorflow Serving
KFServing

Interactive Notebooks
JupyterHub
Notebook Server (Kubeflow)
Elyra *JupyterLab*

ML Applications
Open Data Hub AI Library

Business Intelligence
Superset

Data Scientist

Business Analyst

Data Analysis

Big Data Processing
Spark *Spark SQL Thrift* *Presto*

Streaming
Kafka Streams *Elasticsearch*

Data Exploration
Hue *Kibana*

Metadata Management

Hive Metastore *Metadata (Kubeflow)*

Storage

Data Lake
Red Hat® Ceph Storage (Ceph)
Red Hat® Openshift Container Storage (rook-ceph)

In-Memory
Red Hat® Data Grid (Infinispan)

Relational Databases
PostgreSQL *MySQL*

Data Engineer

Data in Motion

Red Hat® AMQ Streams (Kafka Strimzi) *Red Hat® Ceph S3 API* *Red Hat® Fuse (Camel K)* *Kafka Connect* *Logstash* *Fluentd* *rsyslog*

Security & Governance

Red Hat® OpenShift OAuth

Red Hat® Single Sign-On (Keycloak)

Red Hat® Ceph Object Gateway

Red Hat® 3scale

Data Steward

Monitoring & Orchestration

Prometheus

Grafana

Kubeflow Pipelines

Elyra

Kubeflow Experiments

Airflow

Argo Workflows

Openshift Pipelines (Tekton)

Argo CD

DevOps Engineer

Red Hat OpenShift

kubernetes

Red Hat Enterprise Linux

Red Hat Hybrid Cloud Management

mongoDB · ceph · neo4j — Versatile Data Storage

jupyter · H2O.ai — Dev environments

argo · Apache Airflow — Pipelines & CI/CD

Prometheus · Grafana — Perf. & Stat. Monitoring

ELK — Log monitoring & analysis

Inner Source_

Reuse to build

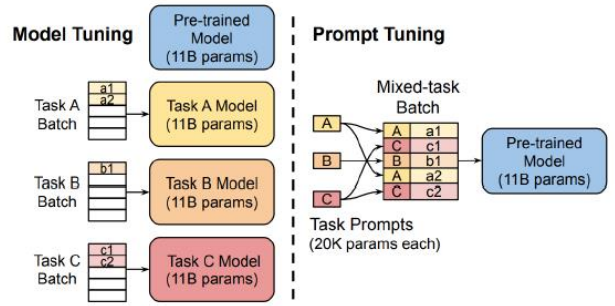
Red Hat

Nuevo objetivo de la plataforma:

Usar, entrenar, y evaluar LLMs

1 **Datos**
Obtención de conjuntos de datos para entrenar LLMs, su contenido y problemas.

2 **Entrenamiento**
Entrenamiento de una LLM desde 0, y fine-tuning para tareas específicas.



3 **Evaluación**
Evaluación de LLMs y benchmarks utilizados en el estado del arte.

4 **Inferencia**
Inferencia con LLMs, quantization, prompt-engineering, y casos de uso.

Plataforma Inteligente

sopra steria

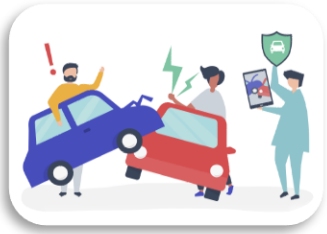
Red Hat OpenShift

Red Hat OpenShift Data Science

Red Hat

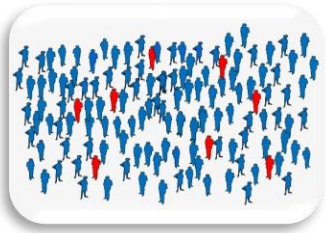
Nuevos retos en los casos de uso:

Generar datos de entrenamiento



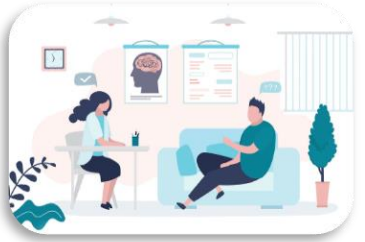
Seguros

Crear perfiles de clientes simulados para predecir reclamaciones y calcular tarifas.



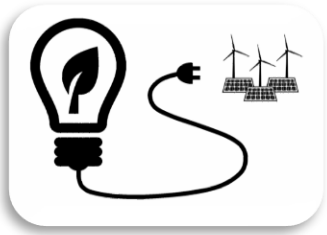
Sociales y Demografía

Generar poblaciones virtuales y estudiar tendencias demográficas y la planificación urbana.



Salud

Crear conjuntos de datos de pacientes simulados sin comprometer la privacidad de los pacientes.



Energía y Medio Ambiente

Simulación de datos de sensores y redes eléctricas.

Finanzas

Imitar el comportamiento de transacciones financieras para la detección de anomalías.

Otros...

Red Hat

Los LLMs no resuelven los viejos problemas:

Problemas con los datos sintéticos

Preocupaciones éticas y de seguridad: uso indebido (generación de contenido engañoso o dañino).

Sesgo: reflejar y amplificar los sesgos presentes en sus datos de entrenamiento.

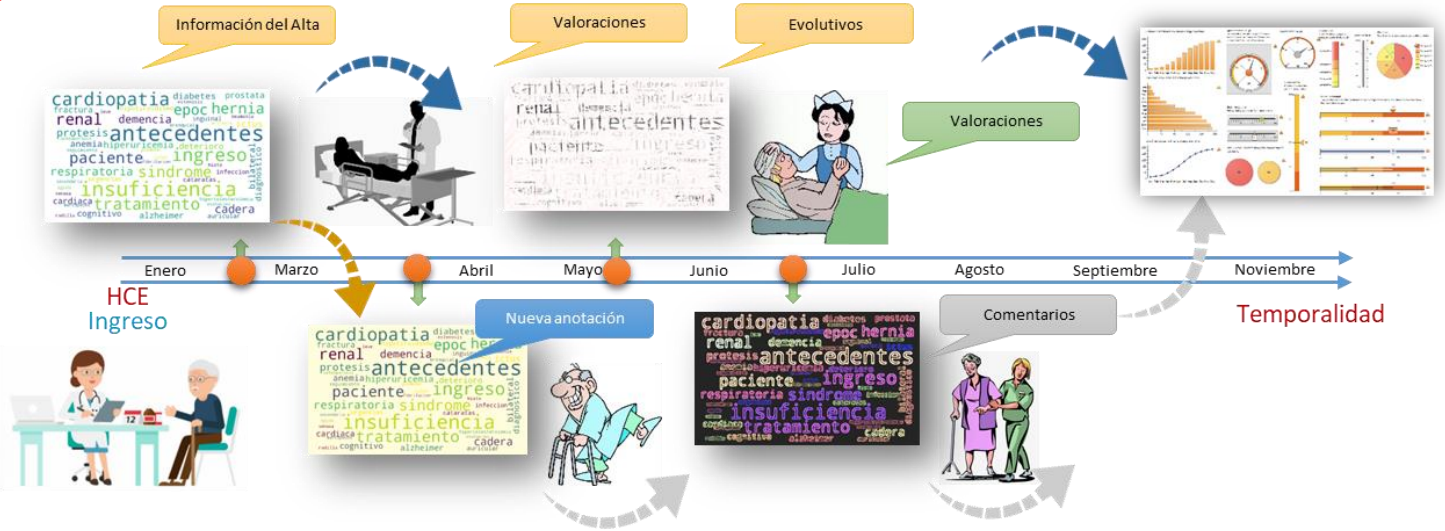
Interpretabilidad y transparencia: entre más grandes y complejos, más difícil es entender cómo toman sus decisiones, un problema para tareas donde la transparencia es importante.

Generalización y solidez: pueden tener dificultades con tareas que son ligeramente diferentes o con entradas que no han visto durante el entrenamiento.

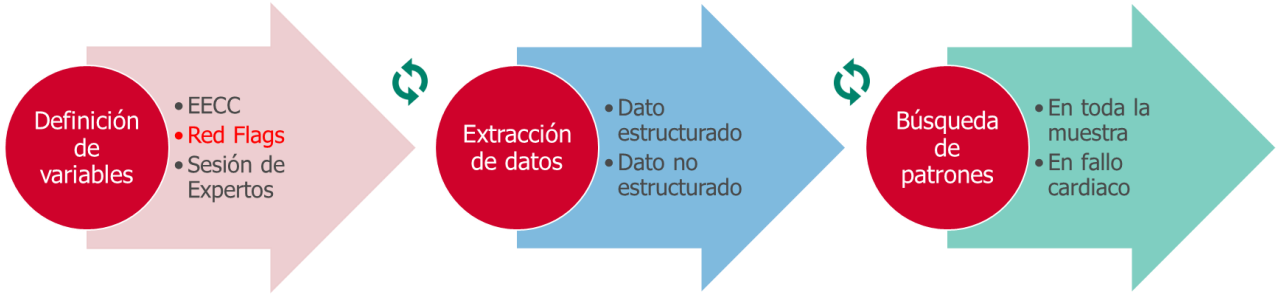
Red Hat

Problemas con el uso de datos sintéticos en contextos sensibles

1 Contexto del dato...



2 Estructuración de los datos...



Imágenes autor: ponente

Uso de técnicas de etiquetado débil sobre datos sintéticos

Problema...
La necesidad de cantidades sustanciales de datos de entrenamiento etiquetados manualmente.

Contexto...
La llegada del transfer learning ha aliviado enormemente este requisito, pero aún sigue siendo necesario generar más ejemplos etiquetados.

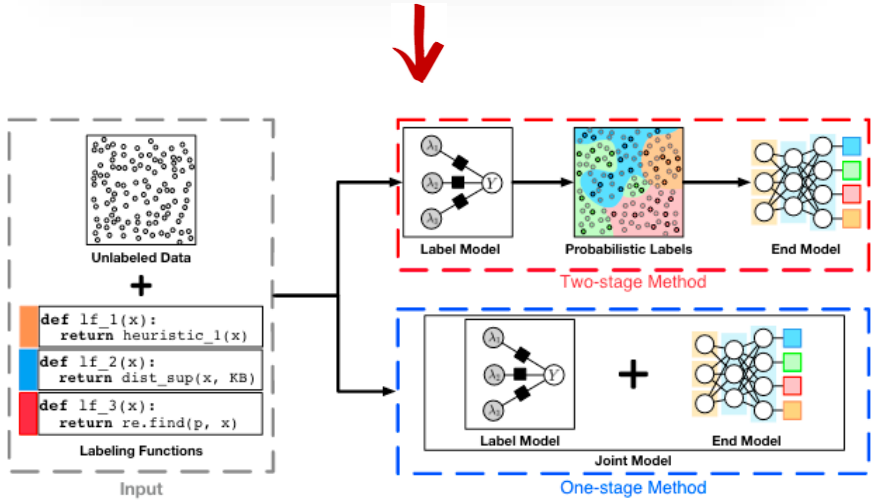


Imagen autor: https://arxiv.org/abs/2109.11377

Weak Supervision

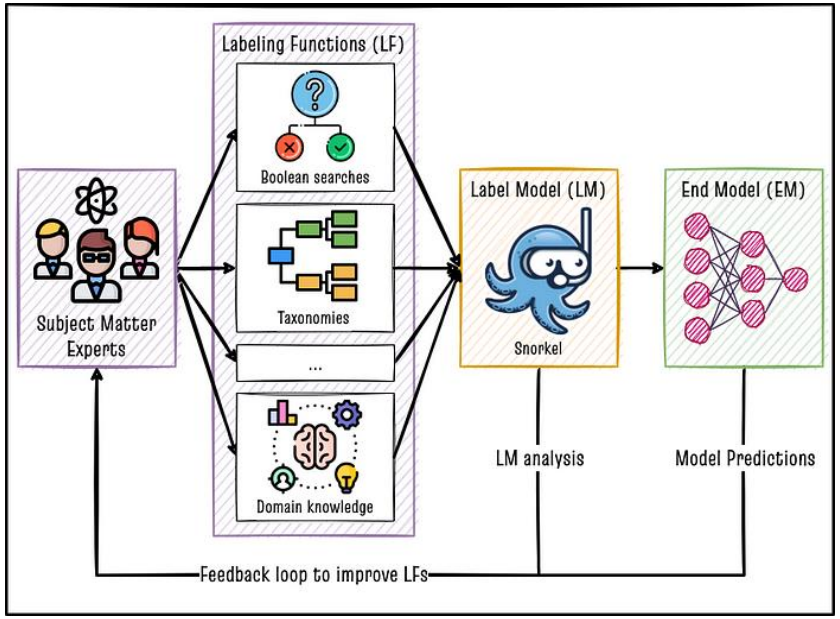


Imagen autor: linkedin.com/in/marie-stephen-leo

Formalización clásica de la generación de datos sintéticos

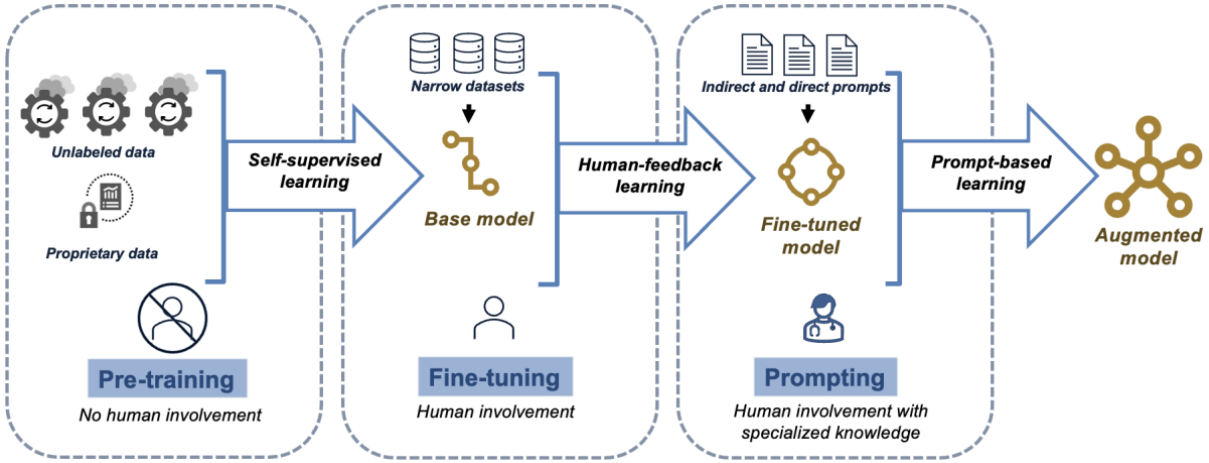


Imagen Autor: Jesutofunmi & Gui, Haiwen & Rezaei, Shawheen & Zou, James & Daneshjou, Roxana. (2023)

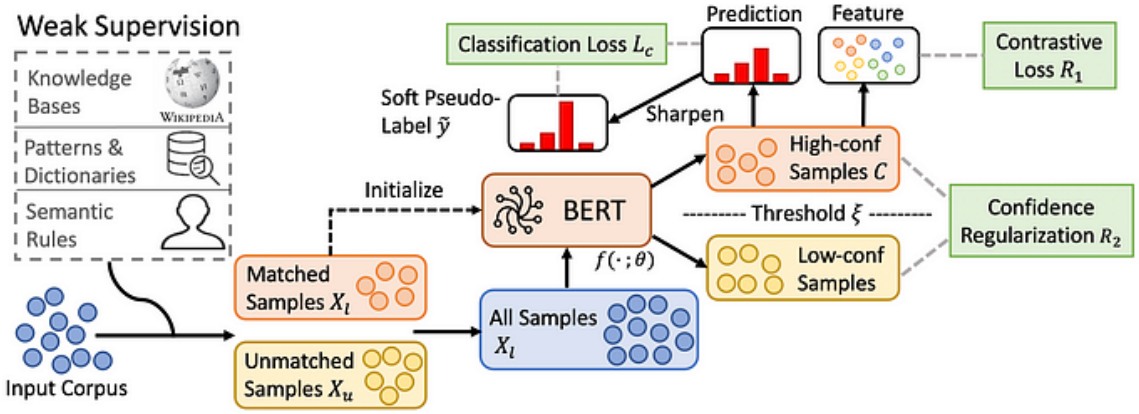
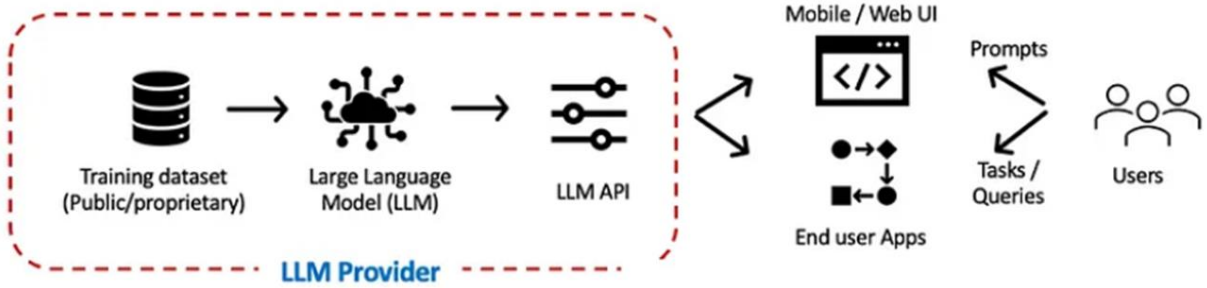


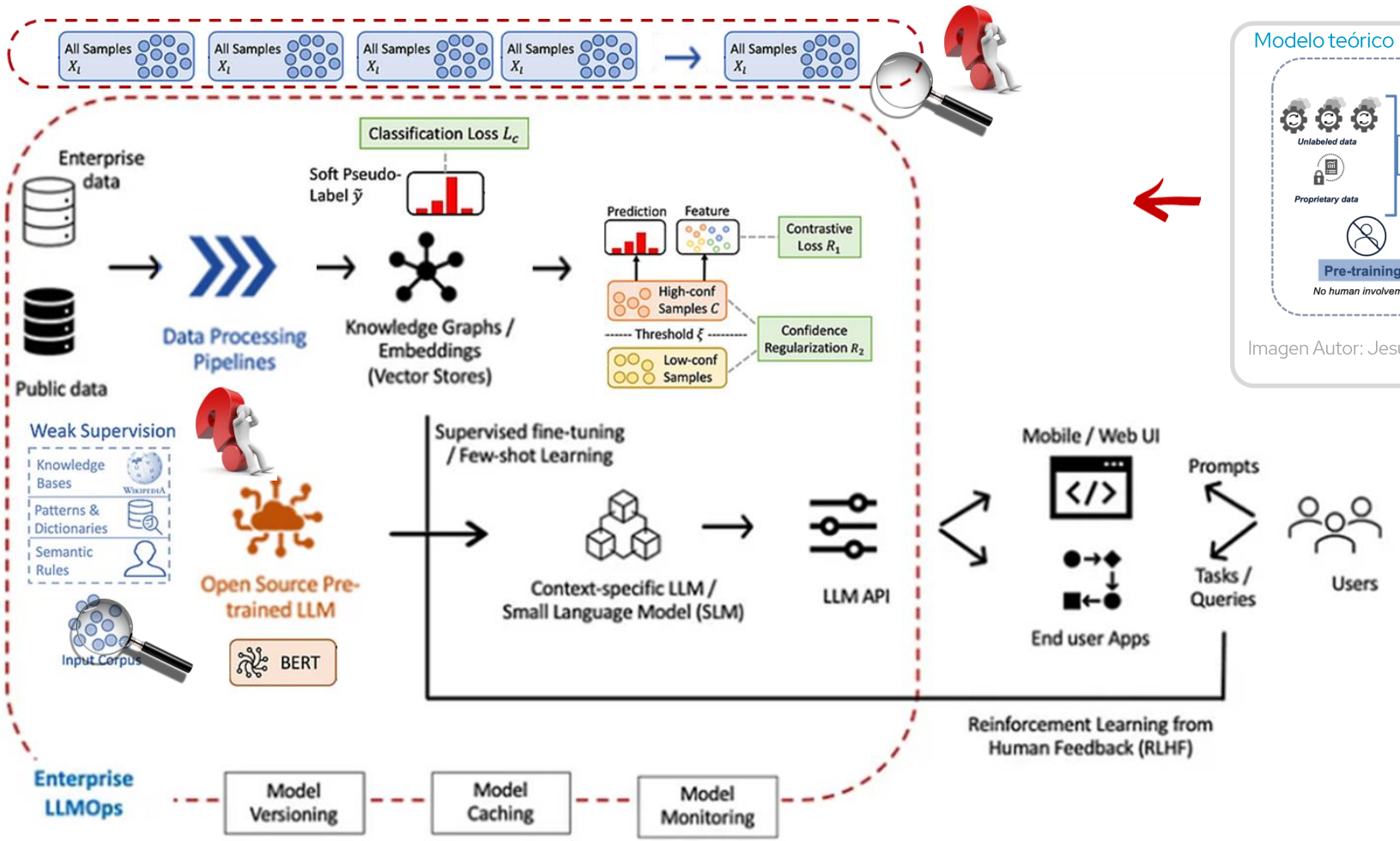
Imagen autor: https://aclanthology.org/2021.naacl-main.84/

Modelo Caja Negra



Tomado del artículo de Debmalyas Biswar

LLMOps y el versionado de datos sintéticos



Inspirado en el artículo de Debmalyas Biswar

Modelo teórico

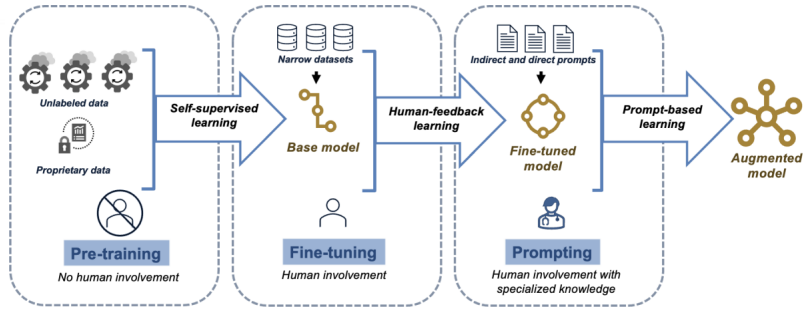
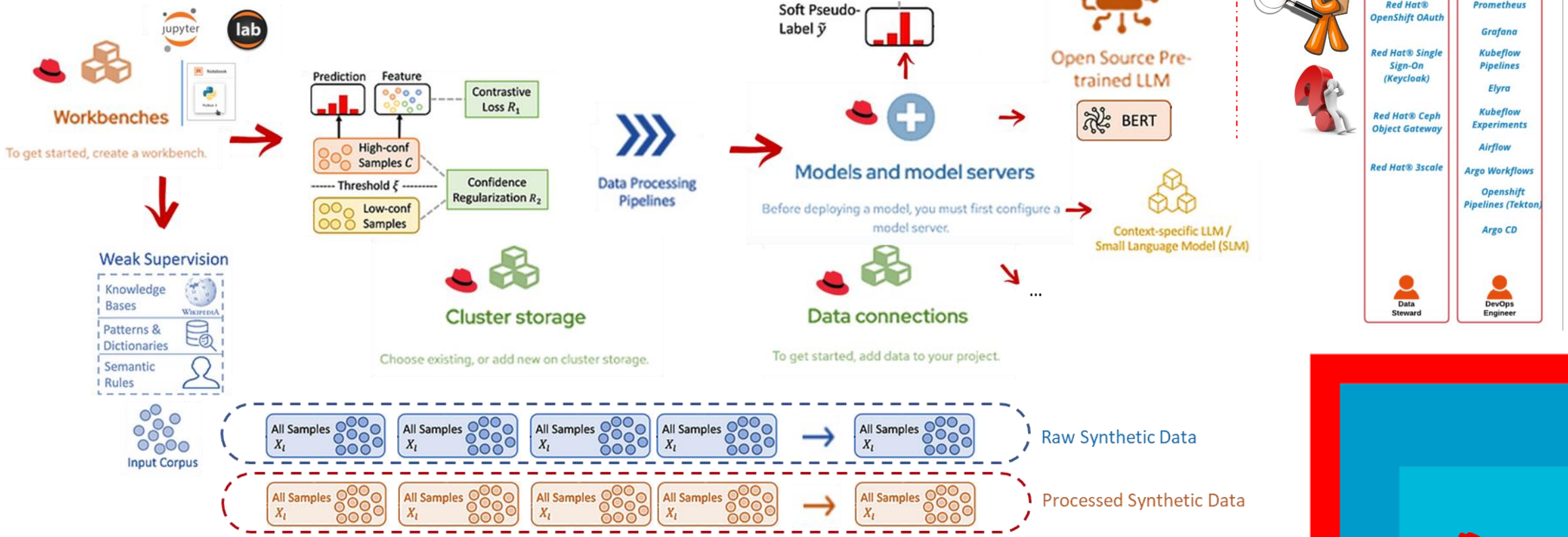


Imagen Autor: Jesutofunmi & Gui, Haiwen & Rezaei, Shawheen & Zou, James & Daneshjou, Roxana. (2023)

Laboratorios para el versionado de datos sintéticos

Red Hat OpenShift Data Science



Red Hat

Red Hat Summit

Connect

Thank you

 linkedin.com/company/soprasteria

 facebook.com/SopraSteriaEspana/

 youtube.com/@SopraSteriaGroup

 twitter.com/SopraSteria_ES

Red Hat