1. Team Details

Team Name: Data Guardians

Team Members:

- Pranay Jalan
- Sidak Malhotra
- Hrushikesh Kant
- Bhavya Meghnani

2. Problem Understanding and Scope

- **Problem Statement:**
  Organizations are struggling to leverage the power of AI and Large Language Models (LLMs) due to significant data privacy risks surrounding their most valuable assets: sensitive internal information, including Personally Identifiable Information (PII), Material Non-Public Information (MNPI), and proprietary company data. This presents a significant ethical and legal challenge, that is increased by a disconnect between business units who understand the data's context and IT teams who possess the technical expertise. This gap is widened because traditional methods for masking sensitive data are insufficient for the complexity of modern enterprise information. Simple regular expressions (regex) are too brittle, while standard machine learning models for Named Entity Recognition (NER) lack the deep contextual understanding needed to identify subtle or ambiguous sensitive information. The problem is especially acute with "linked visual data," as 30-60% of enterprise data is locked in complex scans and PDFs where information is visually and contextually connected, making simple extraction and redaction ineffective. Consequently, kick-starting critical AI proofs-of-concept (POCs) is a major hurdle. Promising initiatives are throttled by regulatory compliance fears and the immense difficulty of preparing safe, usable data for model training, leaving the transformative potential of AI largely untapped.

- **Target Documents & Formats:**
  Our solution will support a wide range of enterprise documents and data formats, including structured files like CSVs and JSON, unstructured data from databases and data lakes, and complex visual documents such as scans and multi-page PDFs. This includes financial statements, internal corporate records, and other documents containing sensitive firm information.

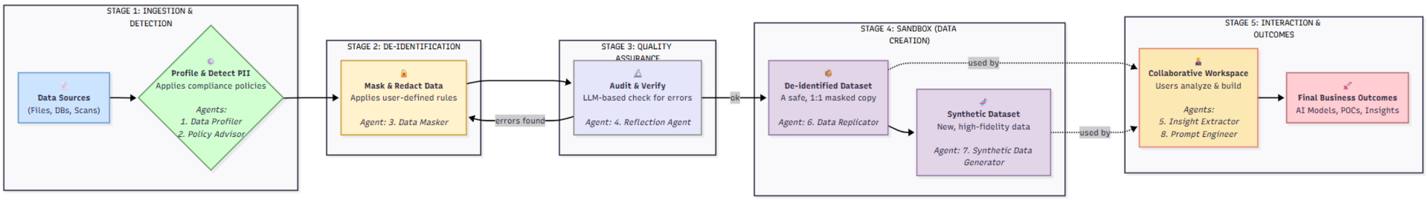
- **Types of Identifiable Data (PII) to Detect & Redact:**
  Our system is engineered to detect and redact a vast and nuanced spectrum of sensitive information, moving beyond simple keyword matching to a contextual, policy-driven approach. The detection logic is adaptable to specific regulatory frameworks (e.g., HIPAA, GDPR, GLBA, DPDPA) and recognizes the combinatorial power of quasi-identifiers. The platform identifies:
  - **Direct & Universal Identifiers:** Core PII that is a primary target in any context.

        - **Personal Identification:** Full names, Social Security Numbers (SSNs), driver's license numbers, Aadhar card, PAN card and Passport numbers
        - **Contact Information:** Full addresses (including any geographic subdivisions smaller than a state), email addresses, and telephone/fax numbers.
        - **Financial Information:** Bank account numbers, credit card purchase details, loan balances, and other certificate/license numbers.
    - **Quasi-Identifiers & Linkable Data:** Information that becomes sensitive when combined, leveraging the "linkability principle" to assess re-identification risk.
        - **Demographic Data:** ZIP codes, race, gender, and full dates of birth. A combination of just these three can uniquely identify a majority of citizens.
    - **Regulatory-Specific Data Categories:**
        - **Non-public Personal Information (NPI) under GLBA:** We identify any personally identifiable financial information collected by a financial institution, such as income details on a loan application or a simple list of a bank's depositors.
    - **India-Specific Identifiers (DPDPA Compliance):** The tool has dedicated recognizers for critical identifiers in the Indian context.
        - **Aadhaar Number:** The 12-digit unique ID from the world's largest biometric ID system.
        - **Permanent Account Number (PAN):** The 10-character alphanumeric code mandatory for most financial transactions in India.
        - Other common document IDs from driving licenses and voter cards.
    - **Visual & Biometric Identifiers:** Our vision pipeline is designed to detect sensitive information embedded in document images and scans.
        - **Photographs:** Full-face photographic images and any comparable images.
        - **Biometrics: Signatures**, fingerprints, voiceprints, and retinal scans.

- **User Personas / End Users:**
  The primary beneficiaries are enterprise teams looking to accelerate AI adoption safely. This includes:
    - **Business Teams/Analysts:** Who can experiment with data and generate insights without technical barriers.
    - **IT & Data Science Teams:** Who can quickly develop and validate AI POCs using safe, realistic data.
    - **Legal & Compliance Teams:** Who can leverage the platform to enforce data policies and ensure regulatory adherence.
    - **Information Security (InfoSec) Teams:** Who can enable innovation by providing business and IT teams with a secure, controlled sandbox for accessing and utilizing sensitive data.
    - **Executive Leadership (CIO, CDO, CTO):** Who can confidently sponsor and fund AI projects, assured that data privacy and compliance are fundamentally addressed.

3. Proposed Solution & Approach

- **High-Level Architecture (with Diagram):**



We propose an innovative, multi-agent pipeline called Infowise. The architecture is designed as a sequential workflow that guides data from initial ingestion to a variety of safe, usable outputs for AI development. This flow ensures robust detection, policy-driven redaction, and the creation of high-value, de-identified data assets.
Detailed Workflow:

1. Ingestion & Detection (Agents 1 & 2):

The process begins when a user connects a data source or uploads files (CSVs, JSONs, PDFs, etc.). For visual documents, an OCR (Optical Character Recognition) engine first extracts the text.

- **Agent 1 (The Data Profiler)** then scans this data to detect potential PII and sensitive information, producing an interactive Data Sensitivity Scorecard.
- **Agent 2 (The Policy Advisor)** enriches this detection by cross-referencing findings with regulations like DPDPA and GDPR, highlighting specific compliance risks.

2. Redaction & Masking (Agent 3):

This is the core transformation step.

- Based on the insights from the detection phase, **Agent 3 (The Data Masker)** executes the redaction using techniques like substitution and pseudonymization.

3. Quality Assurance (Agent 4): This is the crucial self-correcting loop.

- After masking, **Agent 4 (The Reflection Agent)** uses a Large Language Model (LLM) to audit the output from Agent 3. It checks for any missed PII (false negatives) or incorrectly masked data (false positives).

4. The Infowise Sandbox: Generation & Analysis (Agents 5, 6, 7, 8):

Once the data passes the quality assurance check, the final stage produces a variety of secure data assets within the Infowise sandbox.

- Teams can use **Agent 5 (The Insight Extractor)** on these safe datasets to find trends and build preliminary models.
- **Agent 6 (The Data Replicator)** creates a new, fully de-identified "clean" version of the original dataset for safe analysis.
- For projects requiring more data, **Agent 7 (The Synthetic Data Generator)** produces new, artificial data that mimics the statistical patterns of the original.
- Finally, business users can interact with all of these assets using **Agent 8 (The Prompt Engineer)**, which translates plain English questions into actions.

- **AI/ML Models Considered**

  Our solution, Infowise, is built on a modular, multi-layered pipeline that strategically combines various open-source AI models. Each layer is designed to handle a specific part of the de-identification process, from initial document ingestion to final quality assurance, ensuring both high accuracy and maximum security.

  1. Document Understanding: OCR & Layout Analysis

  Our first challenge is to accurately extract both text and its structural context from complex visual formats like scans and PDFs. Understanding the document's layout is critical for interpreting the data correctly.

  - **General-Purpose OCR:** For converting standard scanned documents and images into machine-readable text, we will use **Tesseract OCR** as a powerful and extensible baseline.
  - **Complex & Multilingual Documents:** For financial statements or documents with tables and varied languages, we will leverage **PaddleOCR** for its advanced capabilities in table detection and multilingual recognition.
  - **High-Accuracy Structural Analysis:** For form-like documents, we will use Visual Language Models like **Microsoft's LayoutLMv3** or **Donut**. These models fuse visual layout with textual information to understand the document's semantic structure (e.g., identifying which text belongs to a "Name" field versus an "Address" field).

  2. PII Detection & De-identification (NER)

  Once the document's content is digitized, this stage pinpoints and classifies sensitive information using a layered approach for maximum precision.

  - **Rule-Based First Pass:** We will integrate **Microsoft Presidio** as a highly accurate "first line of defense" for detecting structured PII that follows predictable patterns (e.g., credit card numbers, Social Security Numbers).
  - **Contextual ML-Based Detection:** To identify PII that doesn't follow a strict pattern (e.g., names), we will use transformer-based Named Entity Recognition (NER) models like **dslim/bert-base-NER**. For specific industries,

we can plug in domain-specific models like **obi/deid_roberta_i2b2** for healthcare data.

3. Visual Element Detection & Redaction

Because PII is not limited to text, this stage focuses on identifying and redacting visual elements like photos, signatures, and stamps.

- **Object Detection Models:** We will fine-tune open-source models like **YOLOv8** or **Detectron2** to specifically identify the bounding boxes of non-textual PII, allowing us to effectively redact these visual elements.

4. Advanced Verification & Quality Assurance (LLMs)

The final, most sophisticated layer ensures the highest level of quality and catches any subtle PII missed by the previous stages.

- **Open-Source LLMs:** We will leverage models like **Llama 3** and **Mistral 8x7B**. Unlike proprietary alternatives, these offer unparalleled transparency and auditability. This approach enables on-premises deployment, ensuring sensitive data remains within the organization’s controlled environment. The flexibility to fine-tune these models on specific datasets empowers us to achieve highly accurate verification tailored to an organization's unique policies, all while avoiding recurrent per-token costs and vendor lock-in.

- **Data Strategy:**

  Our strategy is built on a continuous cycle of data collection, de-identification, synthesis, and model improvement.

  - **Collect & Profile:** Connectors allow ingestion from existing enterprise data sources (lakes, databases) or manual uploads. **Agent 1** profiles this data to identify sensitivity.
  - **De-identify:** A suite of agents masks the identified sensitive data.
  - **Synthesize & Augment: Agent 6**, the Synthetic Data Generator, studies the masked data's statistical properties to generate new, artificial data at scale. This synthetic data is realistic and maintains relational integrity, which is crucial for training predictive models.
  - **Annotate & Fine-Tune:** To continuously improve our detection models, our platform will include a simple annotation interface. This allows a "human-in-the-loop" to review and correct the automated PII detection on a small subset of data. This expert-validated data is then used to fine-tune our NER and object detection models, improving their accuracy on an organization's specific document types.

- **Innovation / Unique Selling Point (USP):**
  Our core innovation is the **Multi-Agent Ensemble Sandbox**, which acts as a catalyst for AI adoption within the enterprise.
  - **Layered Defence Pipeline:** The hybrid approach combining rule-based, ML, and LLM agents ensures high accuracy and efficiency in de-identification.
  - **Self-Correcting Loop:** A "Reflection Agent" audits the masked output for errors and triggers re-processing, ensuring exceptional quality.
  - **High-Fidelity Synthetic Data Generation:** We don't just mask data; we create safe, statistically accurate synthetic datasets, solving the POC bottleneck.
  - **Collaborative Sandbox:** The solution is a dedicated workspace designed to bridge the gap between business and IT teams, fostering collaboration.

4. Proposed Solution & Approach (cont.)

- **Intended UI/UX Design (if applicable):**
  The user experience is centered around a web-based, collaborative workspace. It will feature:
  - **Interactive Dashboards:** For visualizing data sensitivity, tracking the masking process, and exploring insights from the safe data.
  - **Intuitive Controls:** A user-friendly interface for defining masking strategies without needing to code.
  - **Natural Language Interaction:** A "Prompt Engineer" agent will allow non-technical business users to query data and initiate tasks using plain English.

- **Input & Output Format Expectations:**
  - **Input:** The platform will support a wide array of formats including structured files (CSVs, JSON), direct connections to databases and data lakes, and complex documents like multi-page scans and PDFs.
  - **Output:** The system will produce fully de-identified/redacted datasets, high-fidelity synthetic data, and comprehensive logs and quality reports. Users can also export the AI models they build within the sandbox for production deployment.

- **Accessibility / Ease of Use Considerations:**
  The platform is explicitly designed for users across the technical spectrum.
  - **For Non-Tech Users:** The "Prompt Engineer" agent provides a natural language interface, and the "Policy Advisor" agent generates easy-to-understand compliance summaries, abstracting away complexity.
  - **For Multilingual Users:** The use of powerful models like Google Gemini 2.5 Pro will provide strong multilingual support for broader enterprise use.
  - **Collaboration-focused:** The entire sandbox environment is designed as a shared workspace to empower business and IT to work together seamlessly.
  - **For Low-Resource Settings:** The platform is designed with a client-server architecture. All heavy AI/ML processing is handled on the server-side, ensuring users can access the full power of Infowise from any standard web

browser without needing powerful local computers. Reports and outputs will also be available in lightweight formats for easy sharing in low-bandwidth environments.

A SCENARIO BASED USECASE OF OUR PRODUCT

Scenario: The Nexus Financial Group Challenge

The Problem: An AI Project Stalled

At Nexus Financial Group, a key AI project to personalize wealth management services was stalled. Business analyst Priya needed access to years of sensitive customer data, but IT security lead Raj had to block the request due to strict DPDPA compliance risks and the months-long, insecure process of manual de-identification. The result was a familiar stalemate: a high-impact business initiative was shelved, causing frustration and lost opportunity.

The Solution: Innovation with Infowise

Nexus introduced Infowise, a secure AI sandbox, to bridge the gap.

Together, Priya and Raj used the platform to instantly profile the data, identify all PII, and highlight the specific compliance risks. With a few clicks, they applied masking rules, which were then automatically audited for accuracy by a **Reflection Agent (Agent 4)**.

Within hours, the Infowise sandbox delivered two secure assets:

1. A fully **de-identified dataset** for analysis.
2. A larger, high-fidelity **synthetic dataset** for model training.

The outcome was transformative. Priya's team used the synthetic data to build their proof-of-concept in just two weeks. The stalled project became a competitive advantage, turning the relationship between business and IT from a roadblock into a true partnership.