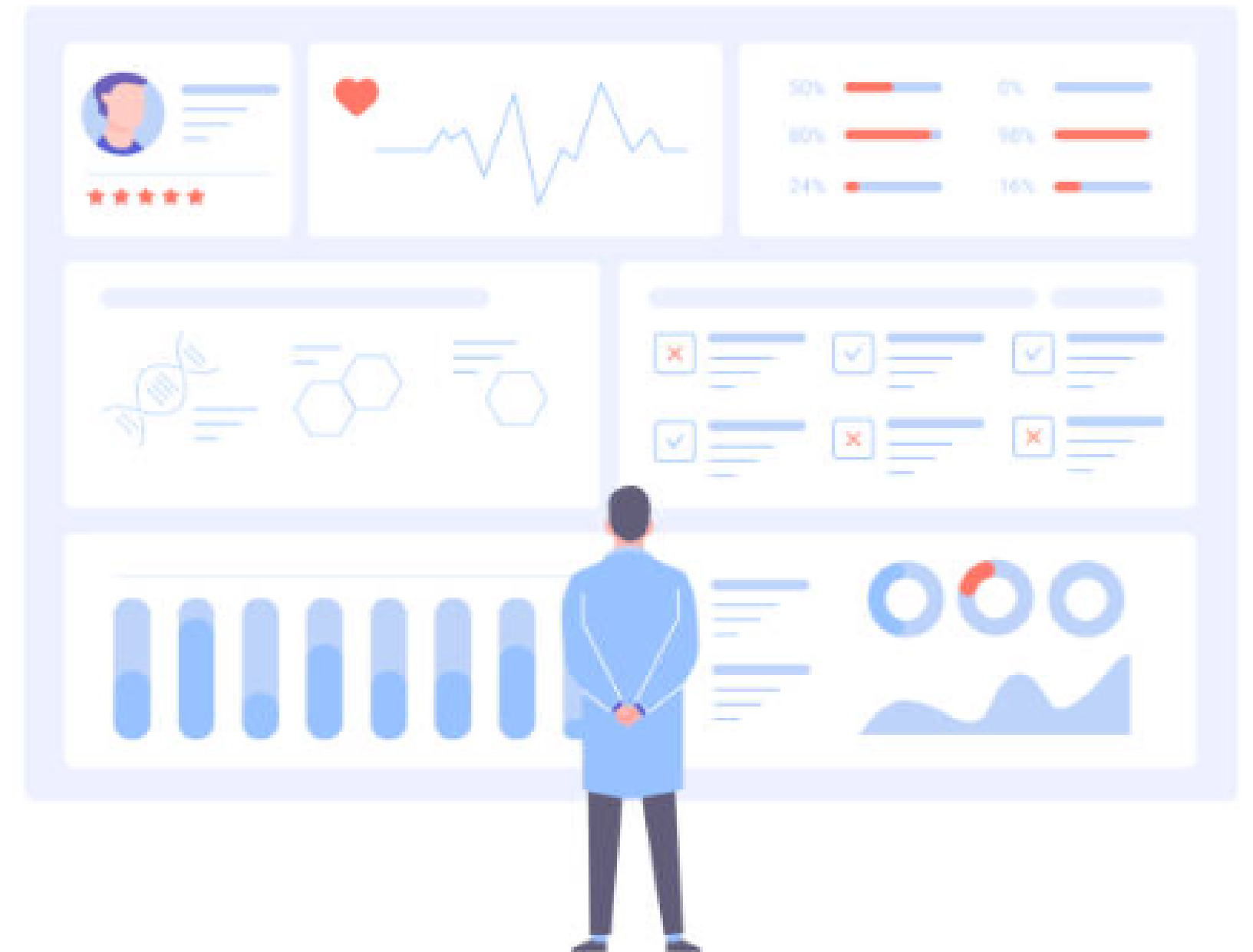# HiLabs™

Problem Statement

You are a PM at HiLabs and are tasked with building a unified platform that serves as a Single Source of Truth (SSOT) for provider data across the U.S. healthcare ecosystem. The platform should aggregate, reconcile, and standardize data from multiple federal and state sources to ensure accurate, up-to-date provider information for payers and health plans.
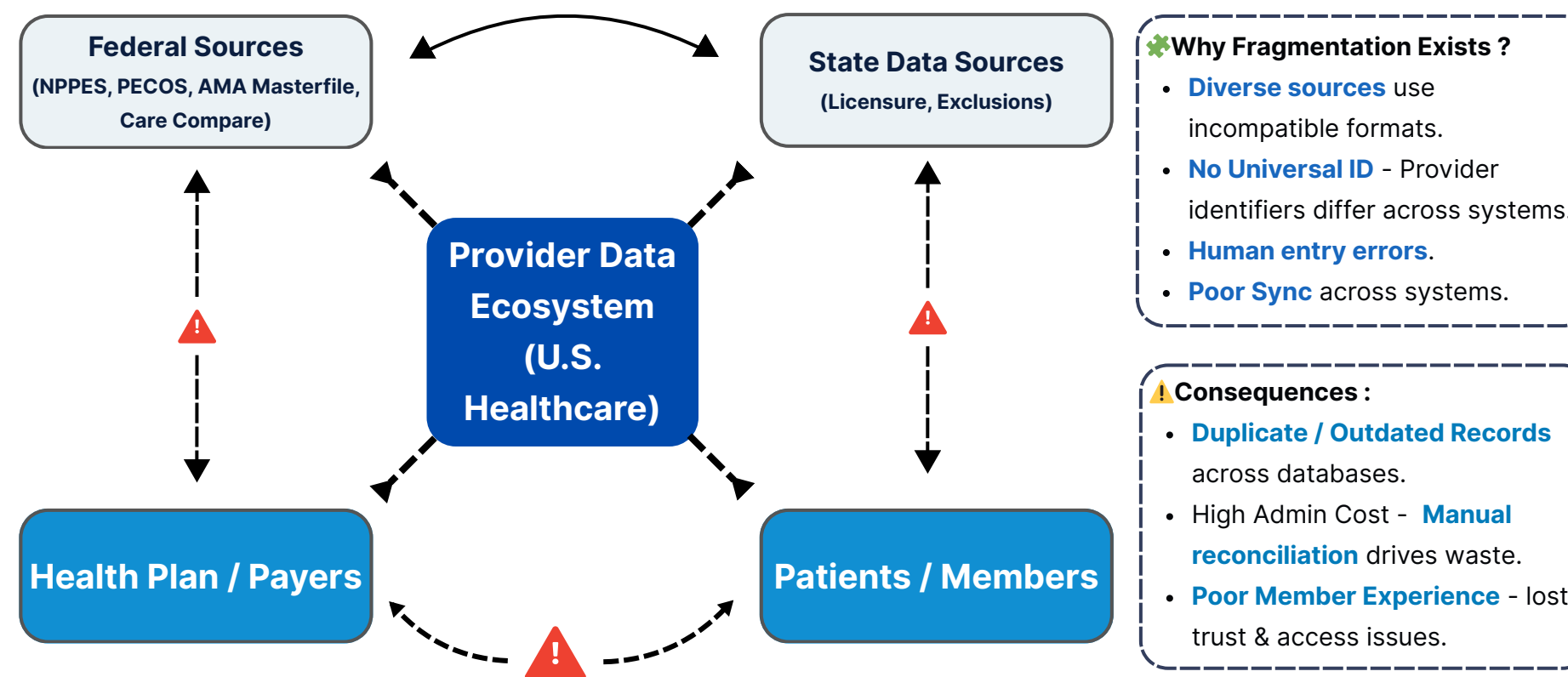
Preet Sharaf (IIT-BHU Varanasi)

preet.sharaf.phe22@itbhu.ac.in
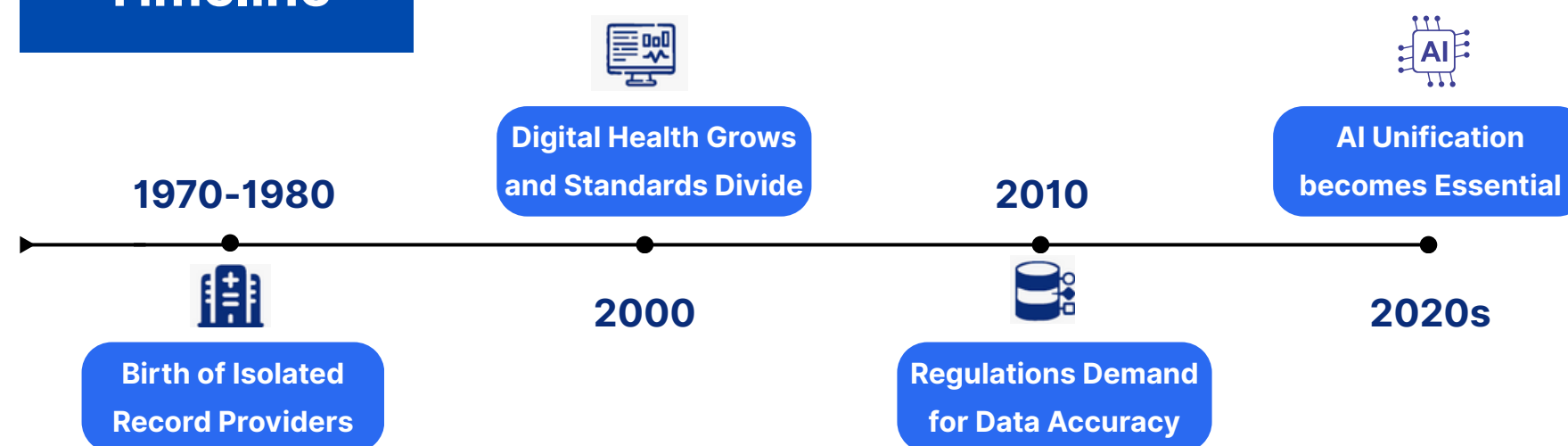
22165049

+91 7435095010

Understanding the Current Process

Federal Sources
(NPPES, PECOS, AMA Masterfile, Care Compare)

State Data Sources
(Licensure, Exclusions)

Provider Data Ecosystem (U.S. Healthcare)

Health Plan / Payers

Patients / Members

✵Why Fragmentation Exists ?
- **Diverse sources** use incompatible formats.
- **No Universal ID** - Provider identifiers differ across systems.
- **Human entry errors**.
- **Poor Sync** across systems.

⚠Consequences :
- **Duplicate / Outdated Records** across databases.
- High Admin Cost - **Manual reconciliation** drives waste.
- **Poor Member Experience** - lost trust & access issues.

Provider data in the *U.S. healthcare ecosystem* **flows** through **fragmented federal and state registries**. Each maintains different schemas, update cycles, and validation rules, creating **discrepancies** when health plans consolidate the data into their systems. This fragmented flow results in **outdated or duplicate provider profiles, regulatory non-compliance,** and **inaccurate directories** that ultimately impact patient access and payer efficiency.

Timeline

1970-1980 — **Birth of Isolated Record Providers**

2000 — **Digital Health Grows and Standards Divide**

2010 — **Regulations Demand for Data Accuracy**

2020s — **AI Unification becomes Essential**

Potential Impacts of Solving it

$400–600 M
annual saving for payers
Through **reduction** in redundant verification, manual data entry, reconciliation overhead.

>95%
↑ data accuracy from ~70% baseline across registries
Achieved via **AI-led deduplication** and schema normalization.

60-70 %
faster updates to provider records
Real-time sync across sources eliminates lag from quarterly refresh cycles.

30 % ↑
improvement in compliance scores with CMS and state audits
Consistent provider data reduces directory error penalties .

Business & Operational Impact

- **Cost Efficiency:** Automation of data curation reduces verification & onboarding costs by > 40%.
- **Regulatory Confidence:** Single Source of Truth minimizes CMS audit errors and compliance penalties.
- **Scalable Integration:** Unified provider database supports faster partner onboarding and downstream analytics.
- **Revenue Protection:** Fewer claim denials from provider mismatches and outdated records.

Impact on the Healthcare Ecosystem

- **Improved Decision-Making:** Reliable data enables evidence-based insights for providers and payers.
- **Enhanced Care Access:** Accurate directories connect patients with active, credentialed providers faster.
- **Trust & Transparency:** A unified SSOT builds confidence among members, providers, and regulators.
- **AI-Led Innovation:** Clean, standardized data forms the foundation for predictive analytics and population health models.

Assumptions

- **Trusted Data Sources:** Provider data is assumed to be accessible from verified entities such as NPPES, PECOS, and state medical boards, ensuring baseline accuracy and completeness for integration.
- **Scalable Technical Foundation:** HiLabs is equipped with a secure, cloud-native infrastructure supporting data ingestion, transformation, and ML-driven reconciliation using ETL pipelines and metadata management.
- **Regulatory & Privacy Compliance:** All workflows adhere to HIPAA and CMS standards, with built-in de-identification, encryption, and audit trails to maintain patient privacy and data integrity.
- **Secure Data Management:** Provider and member data are safeguarded through robust access controls and continuous monitoring, ensuring confidentiality, reliability, and compliance throughout the data lifecycle.

Sources - 1)HCLTech 2)CAQH 3)LexisNexis 4)Wipro

Payers / Health Plans

For now, we will focus on **large and mid-sized payers** who face heavy regulatory scrutiny, data management overhead, and CMS penalties for provider data inaccuracies. They have the **financial capability** and **highest urgency** to adopt automated, AI-driven data unification solution.

EHR & Data Providers

Targeting Academic Medical Centers **(AMCs)**, **Clinical Trial Databases**, and **Government Public Datasets** ensures scalable, reliable, and diverse data streams for aggregation and model training. These sources form the **backbone** for interoperability and real-time data verification.

PAYERS / HEALTH PLANS

	Large Tier (≈ 50) ✓	Medium Tier (≈ 200) ✓	Small Tier (≈ 1500)
Scale of the Company			
Avg Annual Revenue	~ $40–50 Billion	~ $10 Billion	~$1 Billion
R&D + Data Ops Budget (~ 1-2%)	~ $400–600 Million	~ $100 Million	~ $10 Million
Market Penetration Potential	50%	40%	20%
Total Addressable Market (TAM)	50×500×0.5 ≈ **$12.5 Billion**	200×100×0.4 ≈ **$8 Billion**	1500×10×0.2 ≈ **$3 Billion**

EHR / DATA PROVIDER

Provider Type	Academic Medical Centers (AMCs) ✓	Genomic / Specialty Databases	Clinical Trial Databases ✓	Govt. Public Databases ✓
Data Type / Value	Deep clinical + research data	Complex disease-level biology	Structured outcome data from trials	National registries (e.g., NPPES, PECOS)
Integration Cost (yearly)	~$500K	~$1M / db	~$250K / trial	Free / low cost

PAYER / BUYER PERSONA ☑

Regional Health Plan

(Chicago, IL) │ ~5M members │ ~$80M Data Ops budget

Pain Points:
- Fragmented EHR inputs cause directory inaccuracy.
- CMS penalties for outdated provider data.
- Manual record matching inflates cost.

Needs:
- Centralized, validated provider database.
- AI Automation - Smart deduplication, sync, & compliance checks.

EHR/DATA PARTNER PERSONA ☑

Academic Medical Network

(Boston, MA) │ 8 hospitals │ 10K+ providers

Pain Points:
- Multiple EHRs with incompatible schemas.
- Delayed sync slows research & billing cycles.
- Limited interoperability with payer systems.

Needs:
- Secure data exchange with payers/regulators.
- Smart QA Layer - AI-based validation to flag incomplete or mismatched records.

Sources - 1)CMS 2)CAQH 3)Atlas Systems 4)Availity

Introducing HiSync - The Single Source of Truth (S.S.O.T.)

I.) Fragmented Provider Data Sources

HiSync **aggregates provider data** from federal, state, and clinical registries (e.g., NPPES, PECOS, AMC trials). Each source has unique schemas, terminologies, and refresh cycles, **causing duplicates and inconsistencies**.

II.) AI-Driven Data Processing

HiSync's **AI Engine automates** data transformation via :
1.) **De-identification** – Masks personal identifiers.
2.) **Data Cleaning** – Detects/fixes errors via AI framework.
3.) **Entity Resolution** – Merges duplicates, aligns IDs.
4.) **Schema Mapping** – Normalizes data into a unified structure.
5.) **Confidence Scoring** – Assigns trust scores to all records.

User Feedback & Model Retraining

IV.) Real-Time Access & Continuous Learning

Standardized data is shared via secure **APIs and dashboards** for payers, providers, and regulators. A **feedback loop retrains AI models** to improve accuracy, detect schema drift, and enhance reliability over time.

III.) Unified Provider Repository (The SSOT)

Validated data is stored in HiSync's **central repository** — a single, continuously updated source of truth. Every provider record is **version-controlled, traceable,** and **HIPAA-compliant**, ensuring accuracy and regulatory readiness.

Product Vision

- To build a **seamless, intelligent foundation** for U.S. healthcare — where provider data is accurate, compliant, and consistent.
- HiSync turns scattered registries into a **living network of verified truth**, helping payers, providers, and regulators make faster, more confident decisions.

HiSync is an **AI-driven data integrity platform** that **unifies** provider information across **federal, state,** and **clinical sources**. It automates **de-identification, cleaning, deduplication, and schema mapping** to create continuously validated provider profiles and establishes an **unified SSOT database.**

Strategic Impact

- The U.S. **loses $2B+ annually** to poor-quality provider data.
- Existing tools check data — **HiSync fixes it**.
- By unifying and learning continuously, HiSync cuts **reconciliation time** by **60%**.
- A **self-improving SSOT** that learns from every correction — that's the HiSync edge.

Sources - 1)CAQH Index 2)SalesForce 3)OliverWyman

1. Provider Data Collection, Preprocesing and Database Creation
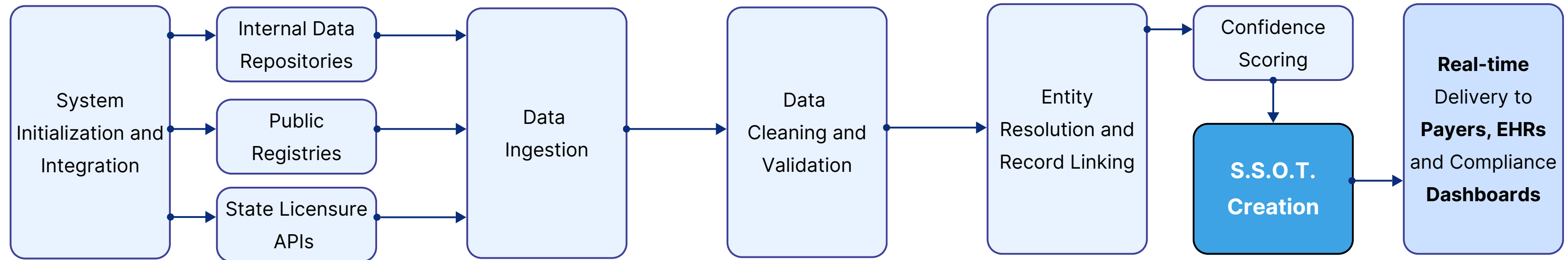
System Initialization and Integration

HiSync **securely connects** to payer databases, public registries, and state licensure APIs to **establish continuous data pipelines**.

Data Ingestion

The platform **automatically fetches provider information** from integrated sources, verifying file formats, schemas, and metadata for **compatibility**.

Data Cleaning and Validation

All incoming data is **standardized, corrected, and validated** to ensure consistency across identifiers, addresses, and regulatory fields.

Entity Resolution and Record Linking

HiSync detects and **merges duplicate or fragmented provider records**, consolidating them into unified, traceable profiles.
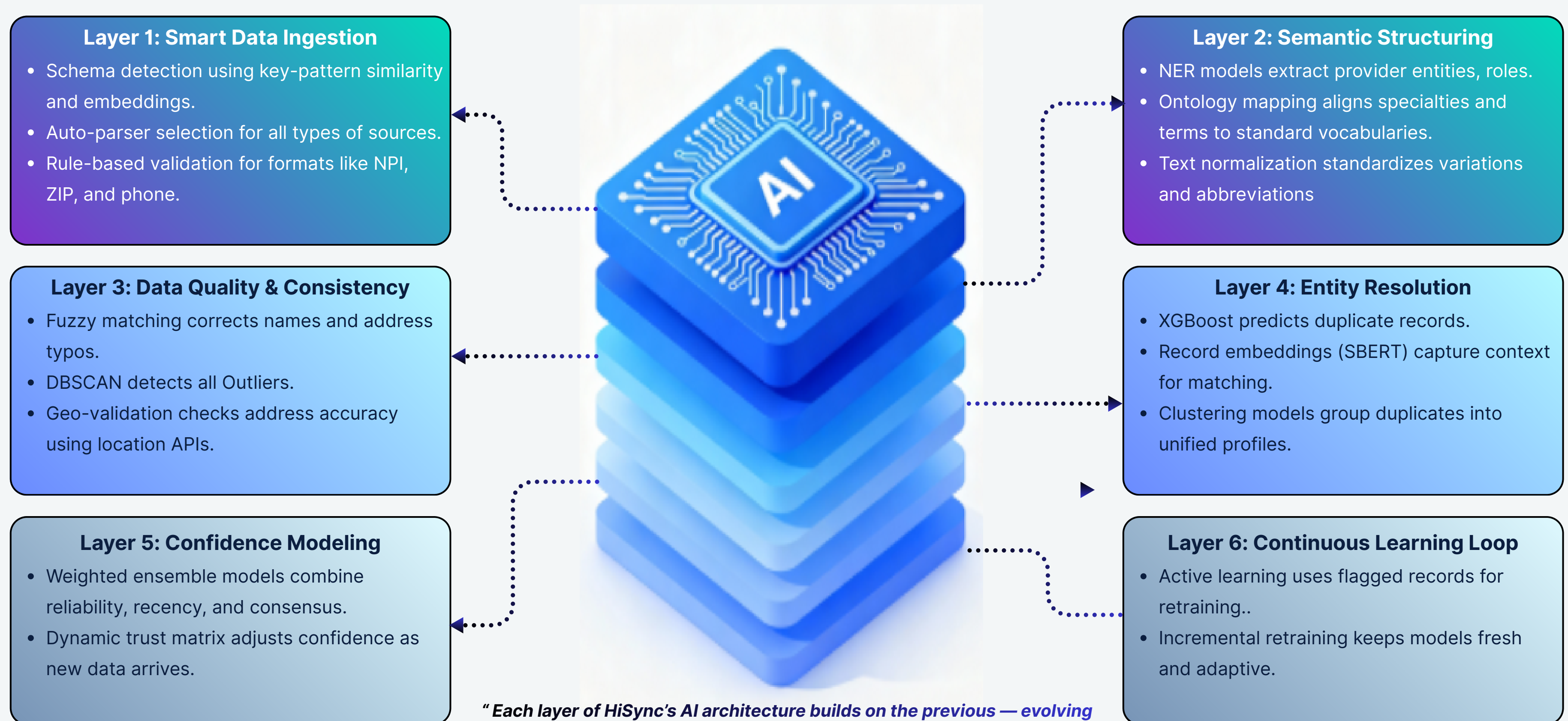
Confidence Scoring

Each attribute within a provider profile is **ranked by reliability and recency**, producing a **field-level confidence score**.

S.S.O.T. Creation

Finalized, high-confidence profiles are stored in **HiSync's Single Source of Truth database for real-time delivery** to payers, EHRs, and compliance dashboards.

## 2. The Layered AI/ML Framework - Architecture Behind Automation

Layer 1: Smart Data Ingestion

- Schema detection using key-pattern similarity and embeddings.
- Auto-parser selection for all types of sources.
- Rule-based validation for formats like NPI, ZIP, and phone.

Layer 2: Semantic Structuring

- NER models extract provider entities, roles.
- Ontology mapping aligns specialties and terms to standard vocabularies.
- Text normalization standardizes variations and abbreviations

Layer 3: Data Quality & Consistency

- Fuzzy matching corrects names and address typos.
- DBSCAN detects all Outliers.
- Geo-validation checks address accuracy using location APIs.

Layer 4: Entity Resolution

- XGBoost predicts duplicate records.
- Record embeddings (SBERT) capture context for matching.
- Clustering models group duplicates into unified profiles.

Layer 5: Confidence Modeling

- Weighted ensemble models combine reliability, recency, and consensus.
- Dynamic trust matrix adjusts confidence as new data arrives.

Layer 6: Continuous Learning Loop

- Active learning uses flagged records for retraining..
- Incremental retraining keeps models fresh and adaptive.

" Each layer of HiSync's AI architecture builds on the previous — evolving raw provider data into self-learning, decision-grade intelligence "

Sources - 1)AWS ML Blogs 2)IBM

User WorkFlow
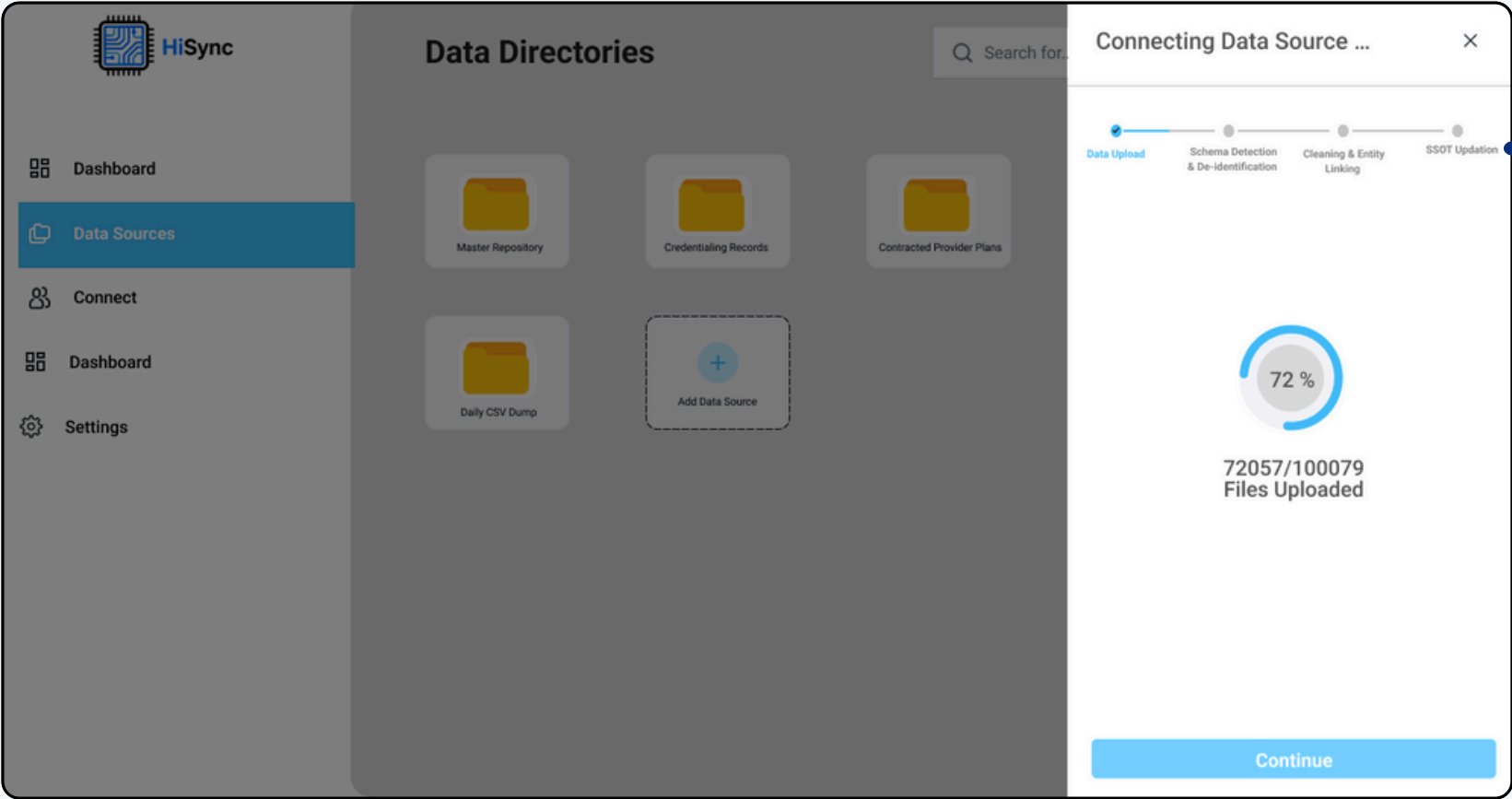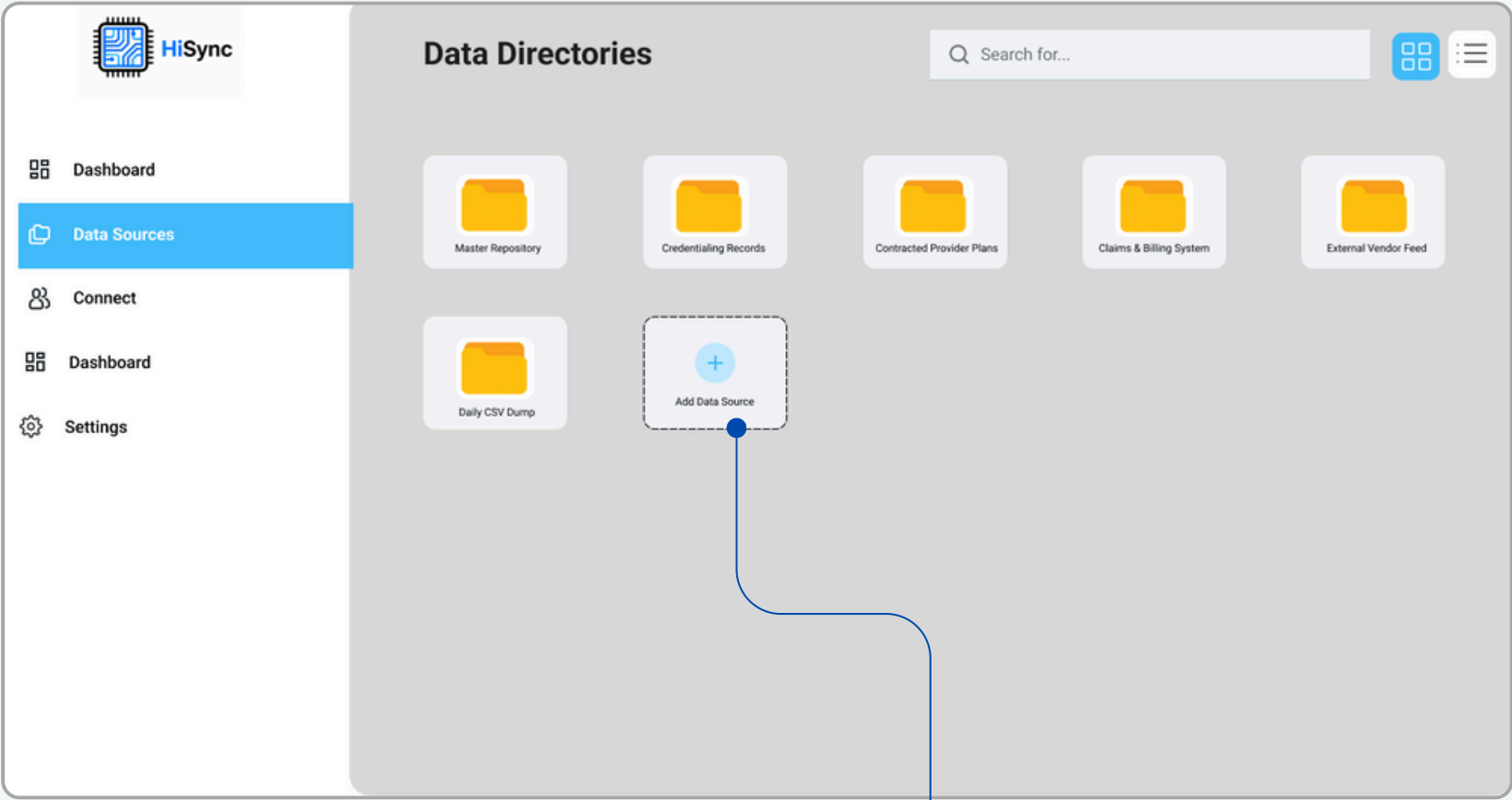
Data Onboarding & Standardization

Users upload a data source (EHR exports, registries, claims files) into HiSync. The platform automatically ingests the file, detects schema formats, standardizes fields, removes duplicates, and merges fragmented provider records. ML handles de-identification, validation, and entity linking in the background. The refined output becomes part of HiSync's Single Source of Truth (SSOT) — a continuously updated, clean, and trusted dataset for the organization.
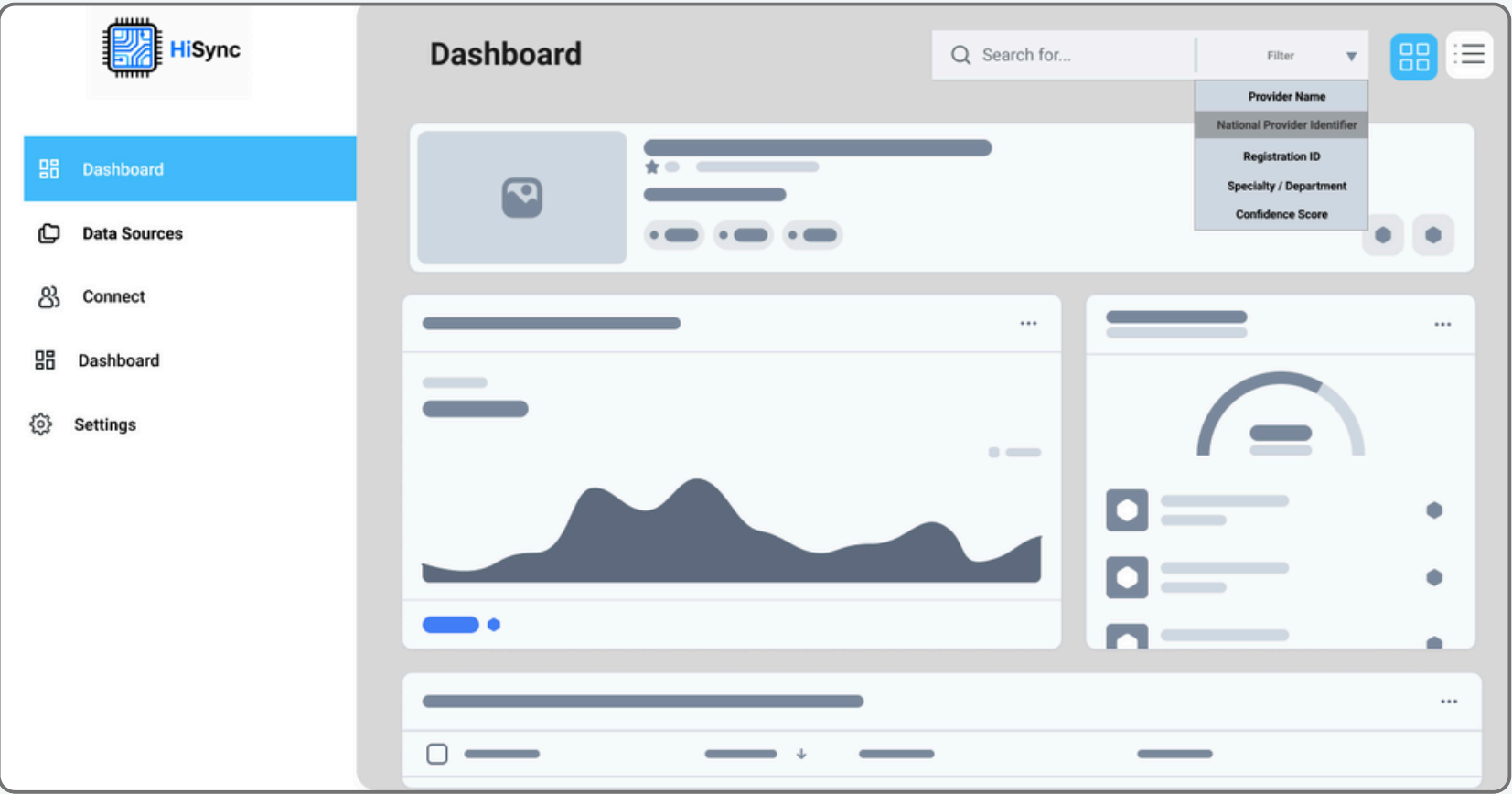
Unified Insights Dashboard

Once processing finishes, users land on the dashboard where reconciled provider profiles are ready for analysis. They can filter and search by various parameters The dashboard displays KPIs and trends with verified records, and the SSOT dataset becomes available to downstream systems via APIs — ensuring teams always use accurate, real-time data.

GTM Plan : Land → Expand → Scale

Land (Pilot: 4–6 weeks)	Expand (Multi-team adoption)	Scale (Org-wide SSOT + integrations)
Start with one use case: provider data cleanup for a single unit. HiSync generates SSOT, deduplication & confidence scores.	After value is proven, onboard more teams (Network Adequacy, Provider Ops, Compliance). Unlock dashboards, NPI search, confidence scoring.	Integrate HiSync via real-time APIs / exports into EHR, Snowflake, CRM, BI tools → becomes the default data source.
Low effort pilot — no IT dependency, minimal training.	Teams rely on HiSync daily → internal stickiness increases.	Renewals become inevitable due to data dependency + high switching costs.
Success: Reduced duplicates + higher data confidence.	Success: Higher completeness + faster onboarding.	Success: SSOT powering downstream systems.

Pitfalls and Mitigations

Metrics

AI Errors in Merging Provider Records
Automated entity resolution may incorrectly merge two different providers or **fail to merge duplicates.**

Mitigation
Low-confidence matches are flagged for manual review, and every merge decision shows field-level lineage to avoid incorrect consolidation..

Limited Adoption from Data Teams
Data/operations teams may **hesitate to trust an automated system due to lack of transparency** or fear of losing control over data governance.

Mitigation
Provide full transparency through **source-level traceability and allow manual overrides**, ensuring teams retain control over the final data..

Integration Barriers With Existing Systems
Even if HiSync cleans the data, organizations may **struggle to integrate the SSOT output** into their infrastructure

Mitigation
Expose the **SSOT via standard APIs and scheduled exports**, enabling easy integration without dependency on IT modernization.

North Star Metric:
% of Actionable Provider Records in SSOT

KPIs:
Avg. Confidence Score, Duplicate Reduction Rate, Profile Completeness Rate

Time to Value (TTV):
Time from first data connection → to first delivery of actionable SSOT records.