

# SkillGenome X

Inferring a Nation's True Skill DNA from Fragmented Real-World Signals

An ML-based prototype for latent skill inference

Ahmed Arshadul Haque

# The Measurement Gap

Current methods for measuring national skills rely on degrees, certifications, surveys, and self-reported data. These metrics are static, incomplete, and fail to capture informal, emerging, and cross-domain skills.

Meanwhile, real skill signals exist across platforms like open-source work, gig activity, and online learning, but this data is fragmented, noisy, and partially adversarial.

## The Core Challenge

**Skills are latent and dynamic**, whilst current systems attempt to measure them directly and retrospectively.

Policymakers and institutions lack a reliable way to understand actual capabilities, regional strengths, evolving skills, and structural gaps.

# A Paradigm Shift in Skill Measurement



Fragmented Data

Latent Inference

Actionable Insight

SkillGenome X reframes skills as **latent variables that must be inferred rather than explicitly measured**. Using machine learning, the system learns compressed representations of skills from fragmented real-world behavioural data. These latent skill embeddings allow us to reconstruct underlying skill structures, identify hidden strengths, and analyse how skills distribute, evolve, and concentrate across regions over time.

# Why This Matters Now

## Credential Disconnect

Degrees and certificates do not reflect real capability or current proficiency in rapidly evolving fields.

## Invisible Talent

Informal and rural talent remains completely invisible to traditional measurement systems.

## Accelerating Change

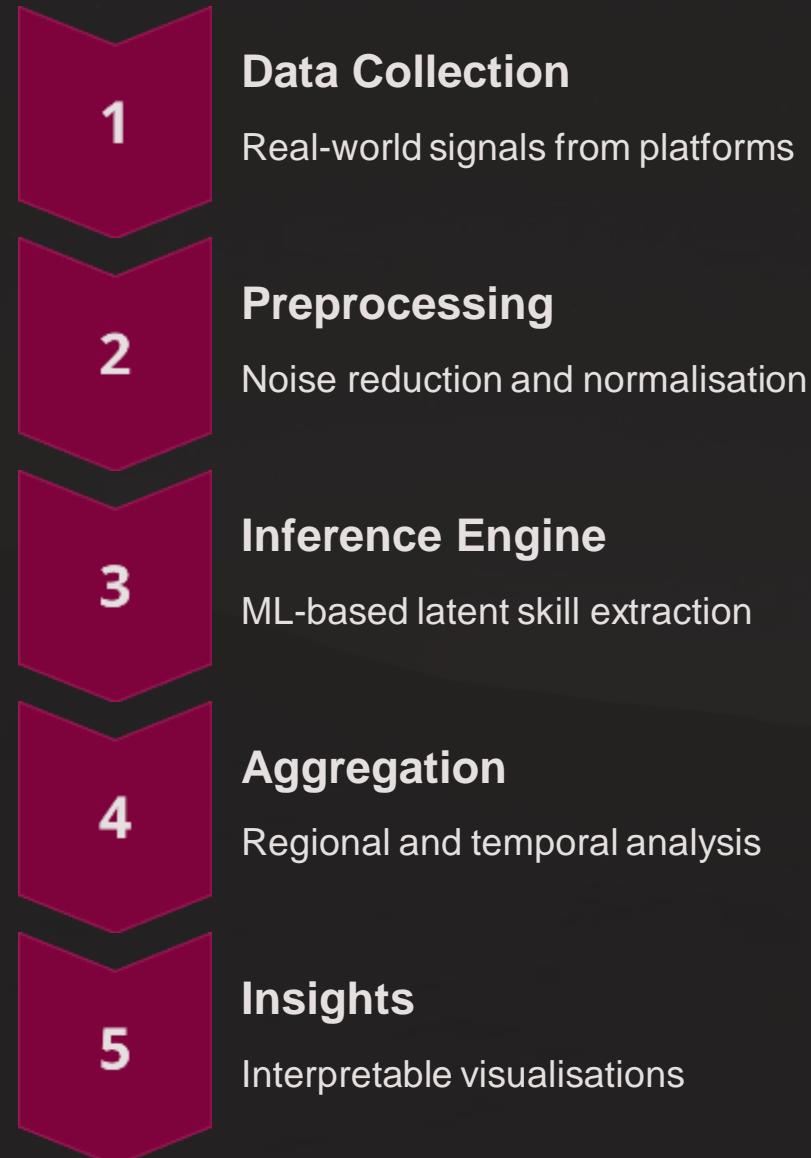
Skill demand evolves faster than institutions can track or curricula can adapt.

## Strategic Risk

Late detection of skill gaps creates national economic and security vulnerabilities.

*You cannot design policy, education, or workforce strategy without visibility into real skills.*

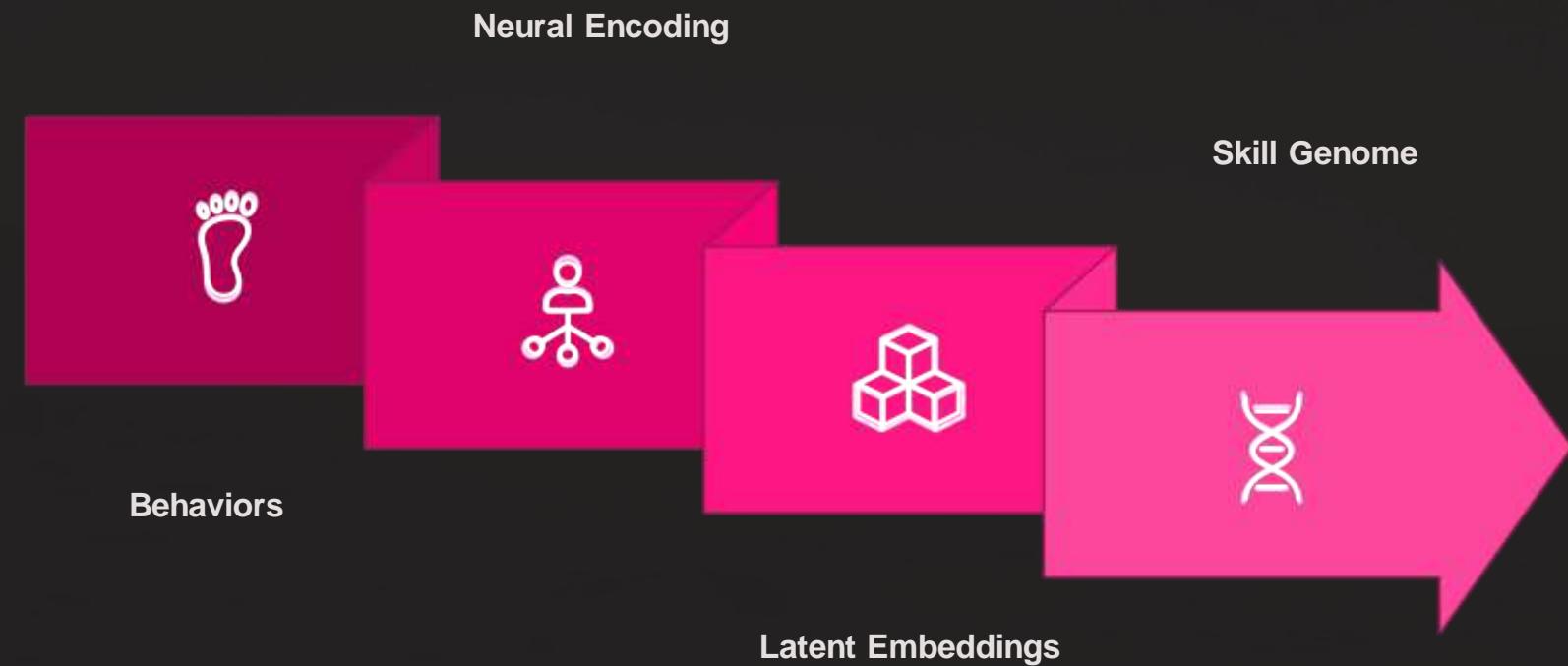
# System Architecture



The focus is on **inference and insight**, not raw data reporting. Each stage transforms noisy signals into actionable intelligence.

# The Core Engine

## Latent Skill Inference



Skills are not directly observable. The system uses a neural encoding model to learn **latent skill embeddings** from behavioural patterns whilst reducing noise and anomalous signals.

These embeddings form the **Skill Genome layer**, representing underlying capability rather than surface-level activity. This approach captures tacit knowledge, informal expertise, and emerging competencies that traditional metrics miss entirely.

# From Inference to Impact

01

## Regional Aggregation

Embeddings are aggregated region-wise to reveal talent clusters and geographical concentrations of specific skill sets.

02

## Temporal Analysis

Time-based analysis highlights emerging and declining skills, tracking how the skill landscape evolves quarter by quarter.

03

## Gap Detection

Distribution comparisons surface potential skill gaps, mismatches, and structural risks before they become critical.

*Inference enables visibility; aggregation enables policy relevance.*

# Technical Feasibility

## Real Data Foundation

Uses publicly available GitHub data as a primary signal source, demonstrating immediate viability.

## Validated Prototype

Prototype validated on limited-scale datasets with measurable accuracy in skill inference.

## Scalable Architecture

System designed to scale seamlessly with additional real-world sources as they become available.

## Deployment Considerations

Data access is a deployment challenge, not a technical limitation. The architecture supports integration with multiple signal sources including:

- Open-source contribution platforms
- Freelance and gig economy activity
- Online learning completion data
- Professional community engagement

# Interactive Demonstration

The dashboard displays the distribution of skill archetypes across three main sections:

- Cluster Statistics:**
  - Technical Specialist: 33 people (11.0%)
  - Continual Learner: 80 people (26.7%)
  - Market Professional: 15 people (5.2%)
  - Creative Innovator: 24 people (8.4%)
  - Orchestrator: 22 people (7.7%)
  - Hybrid Generalist: 15 people (5.2%)
- Skill Genome Map:** A scatter plot showing the relationship between Skill Dimension 1 (x-axis) and Skill Dimension 2 (y-axis). Data points are color-coded by archetype and marked with stars. The plot shows a clear separation between archetypes, with Technical Specialist and Hybrid Generalist clustered at the top left, Continual Learner in the center, Market Professional and Creative Innovator in the middle right, Orchestrator in the bottom right, and Hybrid Generalist at the bottom left.
- Geographic Distribution:** A scatter plot showing the distribution of individuals across a geographic area. The x-axis represents Longitude and the y-axis represents Latitude. Points are colored by archetype, showing a concentration of individuals in North America (USA and Canada) and Europe.

Regional Skill Distribution

### Regional Profiles

- West
- Central
- East
- South
- North

### Regional Skill Distribution Heatmap

	West	Central	East Region	South	North
Technical Specialist	0.18	0.12	0.06	0.07	0.12
Continuous Learner	0.23	0.25	0.27	0.37	0.25
Market Professional	0.04	0.06	0.06	0.07	0.01
Creative Innovator	0.07	0.08	0.08	0.04	0.12
Entrepreneur	0.08	0.04	0.08	0.07	0.09
Hybrid Generalist	0.04	0.06	0.05	0.00	0.09
Data Expert	0.30	0.27	0.35	0.35	0.28
Builder	0.08	0.08	0.05	0.04	0.06

Legend: Impression (0.00 to 0.35)

The chart displays the skill gap analysis across different roles, comparing current distribution (blue bars) against required distribution (orange bars). The y-axis lists the roles, and the x-axis shows the gap in percentage points from -0.15 to 0.05. A dashed vertical line at 0.00 indicates the surplus threshold. A solid vertical line at approximately -0.02 indicates the shortage threshold.

Role	Current Distribution (%)	Required Distribution (%)	Gap (%)
Builder	~0.02	~0.05	~-0.03
Data Expert	~-0.15	~-0.05	~-0.10
Hybrid Generalist	~0.05	~0.08	~-0.03
Entrepreneur	~0.05	~0.02	~0.03
Creative Innovator	~0.05	~0.01	~0.04
Market Professional	~0.05	~0.08	~-0.03
Continuous Learner	~-0.15	~-0.05	~-0.10
Technical Specialist	~-0.02	~-0.01	~-0.01

#	Customer	Age	Adolescent	Following	Date	Language	Source	Region	Urban	Lat	Long	Comments	Prod	Created	First	OptIn	Active	Projects	Rating	Average	Posts	Views	Score	Comments	Ranking	Last View
0	vishaldeep	49	1889	79	114	T	India	West	1	20.039	03.38	1250	22	39	1250	61.4	49	3,708	740000	26	31349	455T	1	2024000	8	
1	moush	411	321	0	382	A	North, Bangladesh	Central	1	23.3918	75.7813	2880	44	28	350	130	36	4,282	310000	31	20800	462P	1	2800000	9	
2	anuradha1998	38	897	0	1294	F	Bangladesh, India	East	1	24.0303	09.5434	975	29	44	1100	130	39	0	17998	31	80799	8818.0	1	8979008	0	
3	rahmeng	61	581	22	394	T	Bangladesh	Central	1	24.5909	02.4468	915	24	34	350	100	30	4,180	21998	30	38898	1804.5	1	3889000	1	
4	manasi4	78	15249	18	4624	A	New Delhi, India	Central	1	26.0208	81.2053	2140	39	38	1250	130	39	2	941618	32	211000	4100.3	1	2112000	9	
5	Vishal-V	219	2229	0	9353	O	India	South	1	10.1889	82.1098	3070	59	58	1250	130	49	3	1511823	46	209993	3026.3	1	2099900	0	
6	PraveenSS	315	3229	127	2307	O	India	East	1	28.8652	85.0553	2725	46	58	1250	130	37	0	1388200	39	133170	8606.0	1	13317000	0	
7	T_ymd4kkesw	68	2062	T	243	O	Gujarat, India	East	1	23.2364	01.3773	1620	27	38	1250	94.3	34	3,980	368888	32	80333	4210.7	1	3953000	0	
8	guptam	61	15673	1	6488	O	India	East	1	25.7026	86.0713	3110	22	38	1250	130	40	1	4209220	27	846403	41312.2	1	8464000	0	
9	anushka199609	58	8584	1	85978	O	New Delhi, India	Central	1	23.2703	79.3439	3270	22	38	1250	130	37	0	2250000	38	2985700	3206.0	1	19857000	0	

# The Path Forward

SkillGenome X demonstrates how fragmented, noisy real-world data can be transformed into a coherent **Skill Genome** using machine learning. By shifting from credential-based measurement to inference-based understanding, the system enables earlier detection of strengths, gaps, and risks in a nation's skill landscape.

This approach provides policymakers, educators, and workforce strategists with unprecedented visibility into actual capabilities, not just documented qualifications. It reveals hidden talent, tracks emerging competencies, and identifies structural vulnerabilities before they manifest as national challenges.

## Core Insight

If skills are latent, inference is the only path forward.

Traditional measurement will always lag behind reality. Only inference can keep pace with change.