**Phase 1 — Build the Verified Internship Sponsor List**

**Goal:** Create a basic, accurate searchable database before touching ML.

1. **Pick Your Data Sources**
   * LinkedIn job postings (filter: “Internship” + location: US).
   * Indeed or Glassdoor internship postings.
   * University career portals (Handshake, Symplicity, etc.).
   * Public lists from international student offices.
   * Crowd-sourced info from current/past students.
2. **Scrape & Extract**
   * Use Python + BeautifulSoup/Selenium (or APIs) to collect:
     + Company name
     + Internship title
     + Job posting text
     + Location
   * Store in a PostgreSQL or MySQL database.
3. **Tag Sponsorship Status (Manual First)**
   * Manually check postings for:
     + **Positive keywords:** “CPT/OPT welcome”, “Will sponsor”
     + **Negative keywords:** “Must be US citizen”, “Work authorization required”
   * Label: Yes Sponsor, No Sponsor, or Unknown.
4. **Build Simple Search UI**
   * React/Next.js frontend + Flask/FastAPI backend.
   * Let users search companies and filter by Yes/No/Unknown.
   * Keep it ugly but functional — focus on accuracy.

**Phase 2 — Add Sponsorship Prediction**

**Goal:** Fill in the “Unknown” companies with a confidence score.

1. **Prepare Training Data**
   * Features:
     + Job description text (NLP bag-of-words or embeddings)
     + Company size, industry, HQ location
     + Internship posting frequency
     + Known CPT/OPT hires from your verified list
   * Labels:
     + Sponsor (1)
     + No Sponsor (0)
2. **Train First Model**
   * Start with **Logistic Regression** or **XGBoost** for tabular + text features.
   * Output probability score (0.0–1.0 → 0–100% likelihood).
3. **Integrate into UI**
   * For Unknown companies:
     + Show “Likely Sponsor (78%)” or “Unlikely Sponsor (15%)”.
   * Add confidence indicators (High / Medium / Low).

**Phase 3 — Feedback & Continuous Improvement**

**Goal:** Make the model smarter over time.

1. **User Feedback Loop**
   * Let users report: “Applied → got sponsorship” or “Applied → no sponsorship”.
   * Store these reports and periodically retrain the model.
2. **Automated Updates**
   * Set your scraper to re-run weekly or monthly.
   * Keep sponsorship info fresh — internships change fast.
3. **Community Building**
   * Encourage international students to contribute data.
   * Reward contributions (leaderboard, recognition badges, etc.).

**What to Build First**

1. **Scraper** for internship postings → store raw data in DB.
2. **Manual tagging** system for verified sponsorship.
3. **Basic search web app** with filters.
4. Only then → **ML model for predictions**.