Executive summary
Our research shows that a two-stage pipeline—(1) an **agentic AI web-scraper** that converts publicly available signals into a structured company record and (2) a **deterministic scoring engine** that converts those records into a 0-100 "AI-Opportunity" index—offers a scalable, low-cost substitute for premium data-enrichment APIs. The scraper harvests digital-maturity, operational-complexity, information-flow, market-pressure and budget signals; the engine normalises, weights and sums them to surface which restaurants, schools, law firms or other SMBs are most primed for AI adoption. Because the maths are transparent and fixed, sales teams can filter thousands of prospects, see *why* each scores high or low, and pitch only the use-cases with the best fit.

---

1. Research foundations

1.1 Digital maturity as an adoption predictor

McKinsey's "Digital Quotient" shows that companies scoring in the top quintile for digital capabilities pull far ahead in AI value capture(McKinsey & Company). Website quality metrics such as Google's PageSpeed Insights offer objective proxies for that maturity(Google for Developers), while CRM or e-commerce tags detected by BuiltWith indicate deeper stack sophistication(api.builtwith.com).

1.2 Deterministic lead-fit models

HubSpot's manual lead-scoring playbooks demonstrate how point-based rules can rank prospects without ML(HubSpot Blog). Sean Ellis's 40 % PMF survey rule provides a precedent for single-threshold indicators(Learning Loop).

1.3 Operational complexity and automation ROI

McKinsey studies link higher employee and location counts to greater automation pay-offs(McKinsey & Company). Document-intensive processes amplify the case for NLP and RAG tools(ScienceDirect).

1.4 External pressure and budget capacity

Competitive intensity accelerates tech adoption among SMEs(Startups.co.uk), while Microsoft's 2024 SMB survey confirms budget remains the primary barrier once awareness is high(Source).

1.5 API cost constraints

Crunchbase's data-licensing tiers start in the mid-five figures annually, making them unattractive for broad prospect sweeps at seed stage(Crunchbase). The agentic scraper strategy eliminates that fixed cost.

---

2. System architecture

2.1 Stage 1 – Agentic AI scraper

- **Agents**: autonomous Reason-and-Act loops using headless browsers.
- **Tasks**:
    1. Crawl target domain; capture HTML, DNS and publicly linked assets.
    2. Detect tech stack via pattern rules (BuiltWith schemas).
    3. Parse revenue cues (pricing pages, "about" copy, careers listings).
    4. Query public APIs with free tiers (Google Places for competitor density, SEC/EDGAR where available).
- **Output**: JSON row conforming to `company_schema v1.0`.

2.2 Stage 2 – Deterministic scorer

For each company row, compute five sub-scores (0–1):

Symbol	Formula (inputs normalised)	Weight
D Digital Maturity	0.4·SiteSpeed + 0.3·CRM_Flag + 0.3·E-com_Flag	0.25
O Operational Complexity	z(Employees) + z(Locations) + z(Services)	0.20
I Information-Flow	log10(DailyDocs+1)/4	0.20
M Market Pressure	z(CompDensity)+z(IndustryGrowth)–z(Rivalry)	0.20
B Budget Signal	log10(Revenue)/7	0.15

Global score:

AI-Opportunity=100 (0.25D+0.20O+0.20I+0.20M+0.15B)\text{AI-Opportunity}=100\,(0.25D+0.20O+0.20I+0.20M+0.15B)

All maths are linear or logarithmic—fully repeatable and auditable.

---

3. Data flow & governance

1. **Nightly ingest:** scraper writes raw snapshots → object store.
2. **Feature extraction:** deterministic parsers populate metric table.
3. **Scoring job:** engine calculates sub-scores and final index.
4. **API / UI:** sales app queries by score threshold, shows “why” panel.
5. **Version control:** weight file (`weights.yaml`) locked after calibration sprint.

---

4. Validation & calibration strategy

- **Back-test cohort:** 200 SMBs where AI pilots succeeded or failed; examine correlation between score and real ROI.
- **Sensitivity analysis:** ±10 % weight perturbations; monitor rank stability.
- **Human sanity check:** sales leads review top-20 / bottom-20 list for face validity.

---

5. Implementation roadmap

Phase	Deliverable	Timeline
P-0	Finalise `company_schema` & weight file	Week 1
P-1	Build minimal agent (crawl + site-speed + tech stack)	Weeks 2-3

P-2	Add revenue heuristics & competitor density module	Week 4
P-3	Scorer micro-service + REST endpoint	Week 5
P-4	Back-test & freeze weights	Weeks 6-7
P-5	Integrate with outreach UI, push to prod	Week 8

---

6. Risks & mitigations

Risk	Likelihood	Impact	Mitigation
Scraper blocked by bots	Med	Med	Rotate IPs, respect robots.txt
Heuristic revenue errors	High	Med	Add confidence flag; prioritise manual validation on low-confidence leads
Metric drift over time	Med	High	Quarterly weight review locked by change-control

---

7. Conclusion

This two-stage architecture replaces costly enrichment feeds with an in-house, explainable pipeline that can rank any local business on AI readiness. By combining deterministic maths with flexible agentic data capture, we keep variable cost near zero while retaining full transparency—ideal for systematic outbound campaigns across heterogeneous industries.

References

1. McKinsey Digital & AI leader spread analysis(McKinsey & Company) – establishes digital-maturity impact.
2. Crunchbase Data-Licensing pricing(Crunchbase) – demonstrates API cost barrier.
3. Sean Ellis PMF 40 % rule(Learning Loop) – precedent for deterministic thresholds.

4. HBR on TAM/SAM/SOM sizing challenges(Harvard Business Review) – defines market-size variables.

5. Study on competitive pressure driving tech adoption(ScienceDirect).

6. Microsoft 2024 SMB AI-adoption survey(Source) – budget as primary hurdle.

7. Google PageSpeed Insights specification(Google for Developers) – objective site-quality metric.

8. McKinsey “unsung AI ideas” (leadership gaps)(McKinsey & Company) – supports digital-quotient logic.

9. HubSpot lead-scoring instructions(HubSpot Blog) – deterministic point model.

10. BuiltWith Domain & Lists API docs(api.builtwith.com) – free tech-stack detection.

11. UK SME tech-adoption pressure stats(Startups.co.uk) – validates market-pressure metric.