DATA SCIENCE PORTFOLIO

# AI驅動的跨領域情報自動化系統

**單次執行過濾 86% 無效雜訊，精準鎖定高價值商業情報**

整合 **Google Gemini** 語意理解與堅韌爬蟲架構的
Python 全自動化解決方案

---

簡報人：**王譽鈞**

# 痛點與挑戰：資訊雜訊與資料清理的困境

### 資訊超載

每日面對數萬條非結構化數據，人工篩選耗時且易錯，難以捕捉轉瞬即逝的 Alpha 情報。

### 時間陷阱

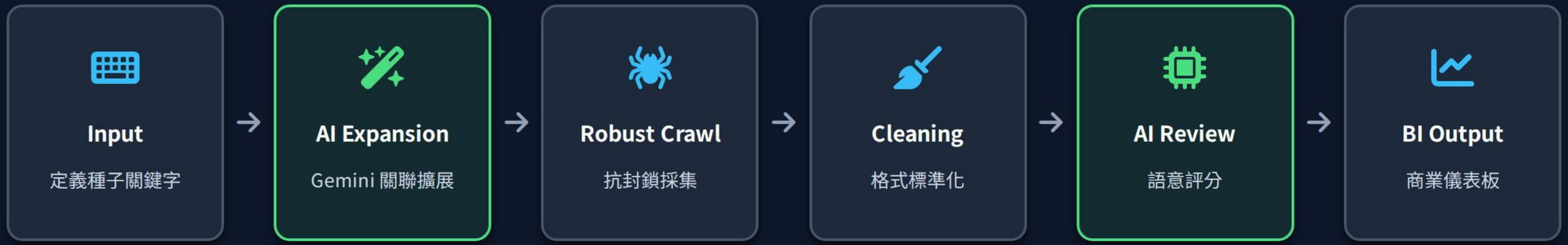
來源網站大量使用 "2天前" 相對時間，導致歷史回測與長期趨勢分析出現嚴重偏差。

### 語意雜訊

關鍵字匹配缺乏語境（如 "Apple" 指水果還是科技？），傳統規則式爬蟲無法有效區分。

# 系統邏輯架構：端到端自動化流水線

Process Flow



# 工程技術亮點：對抗髒資料的三大防線

Python Engineering

## Yahoo 時間校正

針對「2天前」模糊時間，實作 JSON-LD 結構化數據解析。

- 修正 166 則日期錯誤
- 剔除 104 則偽裝舊聞

## Google RSS 代理

開發 RSS 代理模組，突破 ETtoday 等網站的站內搜尋頻率限制。

- 穩定獲取即時索引
- 0% IP Ban Rate

## 多執行緒平行架構

實作 4 Threads 平行架構，在有限運算資源下達成最佳 I/O 平衡。

- 1,300+ 則 / 10 min
- 效能提升 500%

# AI 賦能 I：趨勢獵人 (Trend Hunter)

Generative AI

從「靜態關鍵字」自動轉化為「動態商業場景」，擴展情報視野。

| INPUT ENTITY | | AI IDENTIFIED CONTEXT |
| --- | --- | --- |
| 好市多 (Costco) | → | 2025 黑色星期五 (促銷季佈局) |
| 00940 (ETF) | → | 成分股 輪動效應 (投資痛點) |
| 輝達 (NVIDIA) | → | GB200 供貨狀況 (供應鏈追蹤) |

# AI 賦能 II：嚴格守門員 (Strict Gatekeeper)

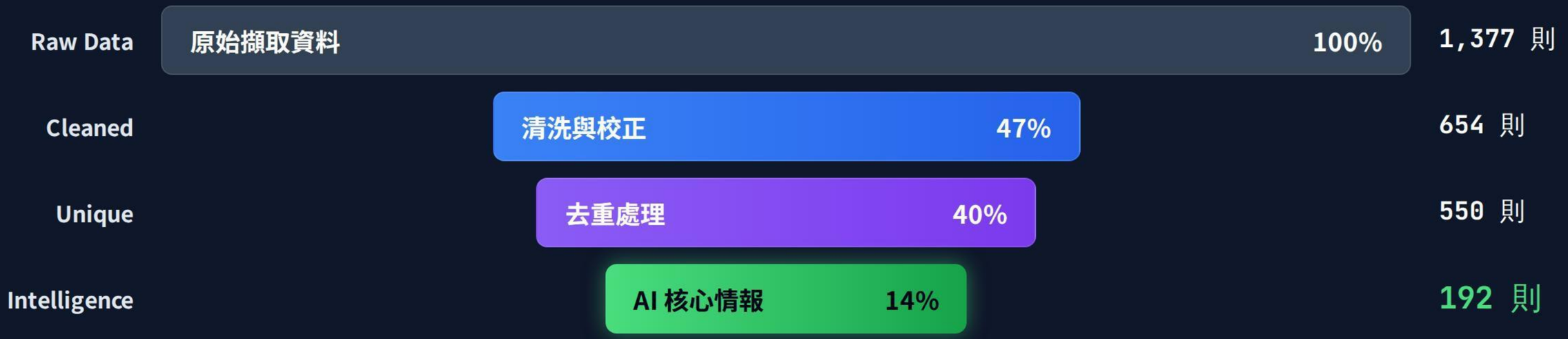
Semantic Filter

AI 成功攔截的真實雜訊案例，精準區分「字面符合」與「語意相關」。

| 搜尋關鍵字 | 爬取到的標題 | AI 決策 | 商業洞察 / 攔截原因 |
| --- | --- | --- | --- |
| OpenAI | 「馬斯克談 **Grok5** 的未來展望...」 | REJECT | 競品動態 (非本體) |
| 黃金價格 | 「陽明海運獲頒環保**金級獎**肯定」 | REJECT | 形容詞誤判 |
| 鴻海 | 「全球**海運**聯盟即將進行重組」 | REJECT | 領域錯置 (航運業) |
| 台積電 | 「台積電 CoWoS 產能擴充計畫提早」 | ACCEPT | 高價值情報 |

# 商業影響力：情報密度漏斗

Data Funnel




總雜訊過濾率達 86%，審閱時間 ~~4hr~~ → 15min

**測試場景：**
科技趨勢(輝達/OpenAI/鴻海) x 品牌(星巴克/好市多/特斯拉) x 金融(BTC/黃金/00940)

**執行數據：**
2025/11/26 (回推3日)

# 作品展示與未來展望

## 1. 自動化報表 Deliverable

目前系統已能自動生成結構化 Excel 報表，包含來源、校正後日期、AI 摘要與信心評分。使用者可直接透過篩選器查看高價值情報。

## 2. Next Step: RAG 整合

未來將導入檢索增強生成 (RAG) 技術，建立企業專屬知識庫，允許使用者進行自然語言問答：

> User: "上週特斯拉的主要競爭對手動態為何？"
> Bot: "根據情報庫，BYD 發布了新的刀片電池技術..."