

# 데이터 저널리즘의 AI 도구 활용 방안: 프롬프트를 넘어 에이전트 설계까지

이종혁 (경희대 미디어학과)

취재보도용 도구 가이드: [https://jonghhhh.github.io/test/journalism\\_ai.html](https://jonghhhh.github.io/test/journalism_ai.html)

# | 서론: AI는 기자의 '두 번째 뇌'

## 멀티플랫폼+AI 시대의 도전과 대응

- 웹·유튜브·SNS 동시 대응 필수
- 소수 인력으로 더 깊이 있는 데이터 생산 요구
- 고품질 데이터저널리즘으로 경쟁력 강화?  
(심층기획 + 데이터 + 시각화)
- 지속 가능한 **뉴스 프로젝트** 필요
  - 실시간 **업데이트**, 개인 맞춤형 **커스터마이징**,  
방대하지만 정확하고 원하는 정보 제공
  - 실시간 대시보드, 챗봇, **RAG DB**
- AI는 취재보도를 돋는 두 번째 뇌(Second Brain)

## 취재 워크플로우 9단계와 AI

- 1) 아이디어 발굴 및 가치 평가
- 2) 정보수집·리서치·모니터링
- 3) 데이터 수집·분석·시각화
- 4) 인터뷰·현장취재 (녹취·번역)
- 5) 기사 작성·요약·검토
- 6) 사실검증 (Fact Check)
- 7) 여론 반응 분석
- 8) 속풀/SNS 제작
- 9) 지식 관리 (Knowledge DB)

# 핵심 개념: 개발자의 언어로 이해하기

구분	SaaS	PaaS	IaaS
풀네임	Software as a Service	Platform as a Service	Infrastructure as a Service
비유	완성된 음식 배달	밀키트 (재료+주방)	전문 주방 구축 + 시장보기
역할	사용자 (User) 사용만 함	활용자 (Utilizer) 조합 및 개발	설계자 (Architect) 인프라 구성 및 운영
예시	Gmail, Tableau, Excel	OpenAI API, Vercel	Agent System, 챗봇, DB

데이터 저널리즘 1.0  


데이터 저널리즘 2.0  


데이터 저널리즘 3.0  


# 데이터 저널리즘 1.0

## 전통적 데이터 저널리즘

완성된 소프트웨어 도구를 활용하여 데이터를 분석하고 시각화

### 대표 도구

Excel, Google Sheets, Tableau, Datawrapper, R

### 한계점

- 데이터 수집/입력이 대부분 수작업
- 속도와 규모의 제약
- 정적(Static) 그래픽 중심
- 실시간 인터랙티브 구현 어려움

### 특징

- **정보수집:** 포털 검색, 정보공개청구
- **데이터:** 통계청 다운로드 → CSV 변환
- **인터뷰:** 녹음기 + 수동 전사(타이핑)
- **분석:** 엑셀 피벗 테이블, R 통계

# | 데이터 저널리즘 1.0 사례

## 경찰 비위 기록 통합 DB (USA Today, 2019)

- 내용: 미국 전역의 경찰 비위 기록을 추적하여 공개
- 방법: 훑어진 문서들을 스크래핑하거나, PDF를 일일이 확인하여 데이터 입력
- 의미: 기술적 난이도보다는 '데이터 수집의 끈기'가 돋보인 사례. 1.0 방식의 정점.
- 유형: 대규모 DB 구축

**Search for police discipline records**



USA TODAY Network has gathered discipline and accountability records on more than 85,000 law enforcement officers and has started releasing them to the public. The first collection published is a list of more than 30,000 officers who have been decertified, essentially banned from the profession, in 44 states. Search our exclusive database by officer, department or state.

[Search database](#)

# | 데이터 저널리즘 2.0

## AI 활용 저널리즘

기존 워크플로우 위에 LLM과 AI 도구를 얹어 생산성을  
극대화

- **핵심:** 반복 업무는 AI에게, 해석·판단은 기자에게
- **기술:** 프롬프트 엔지니어링 (적극적 활용)

### 분야별 도구

- **리서치:** Perplexity, Genspark, NotebookLM
- **녹취/요약:** 클로바노트, Otter, 다글로
- **시각화:** Napkin AI, Flourish

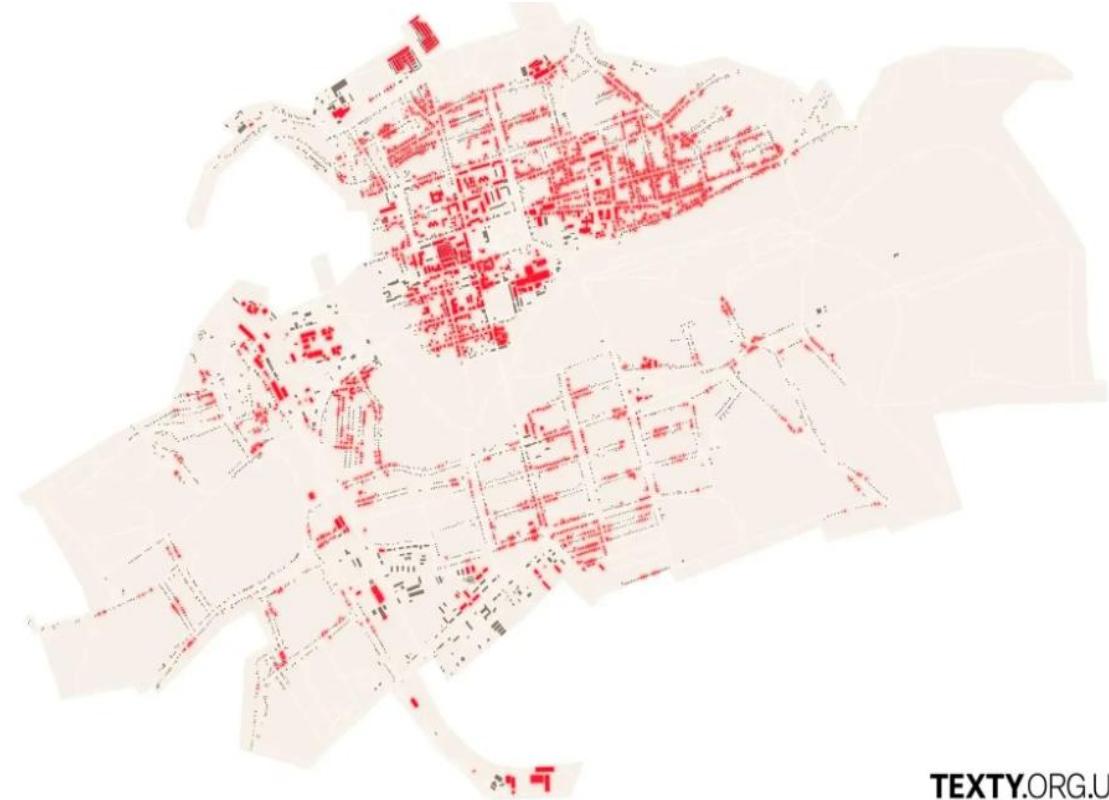
### 특징

- **NLP:** 문서 구조화, 요약, 텍스트 분석
- **Vision AI:** 이미지-영상 분석, 항공사진, CCTV 화면
- **Audio AI:** 음성 기록 텍스트화
- **API:** 대규모 데이터 분석에 적용

# | 데이터 저널리즘 2.0 사례

## 우크라이나 전쟁 피해 지도 (Texty.org.ua, 2024)

- **문제:** 전쟁 중이라 현장 접근 불가, 공식 피해 통계 없음
- **해결:** 위성 사진을 머신러닝 모델에 학습
- **결과:** 파괴된 건물만 식별하여 지도에 매핑
- **시사점:** 인간이 불가능한 규모의 반복 노동(수만 장 사진 분석)을 AI가 대체하여 '없던 데이터' 생성



# | 데이터 저널리즘 2.0 사례

## 아마존 불법 채굴 자동 탐지 (Amazon Mining Watch, 2025)

- 내용: 아마존 열대우림 환경 감시
- 방법: 위성 사진 분석 플랫폼 활용
- 기술: 숲이 채굴장(흙바닥)으로 변하는 시각적 패턴(Change Detection)을 AI가 감지해 알림 전송
- 시사점: 실시간 모니터링 플랫폼을 활용해 감시 영역 확장



# | 데이터 저널리즘 2.0 사례

## 패션 잡지 피부색 분석 (The Pudding, 2019)

- 내용: 보그(Vogue) 표지 모델의 다양성 분석
- 방법: 얼굴 인식 AI를 사용해 표지 모델의 얼굴을 감지하고, 피부색 픽셀을 추출하여 데이터화
- 시사점: 이미지를 사회학적 증거로 변환하는 '문화적 데이터 저널리즘'



# 2.0의 함정: Prompting과 Vibe Coding을 넘어라

ChatGPT 5.1

↑ 공유하기 ⌂ 사람



이미지에 나타난 인물의 얼굴 색깔을 알려줘

이미지를 기반으로 \*\*얼굴의 색깔(피부 톤)\*\*만 설명해 드릴게요.  
(※ 인물의 신원 언급은 정책상 하지 않습니다.)

## 이미지 속 인물의 얼굴 색(피부 톤) 분석

사진을 보면 얼굴의 피부색은 다음과 같은 특징을 가지고 있습니다:

- 피부 톤**
  - 밝은 브라운(Light Brown)
  - 약간 황갈색(Ochre / Golden Beige) 느낌
  - 균일한 톤으로 부드러운 베이지~옐로 브라운 계열
- 색상 범위(대략적 RGB)**

일반적인 모니터에서 보이는 색 기준:

  - RGB(약):**  
(210~225, 170~190, 150~170)
  - HEX(약):**  
#D8B892, #E0C09E 정도의 밝기

ChatGPT 5.1

↑ 공유하기 ⌂ 사람

이미지에 나타난 인물의 얼굴 색깔을 알려주는 코드를 작성해

좋아요, "이미지 속 인물 얼굴 색깔(대략적 피부 톤)"을 구하는 파이썬 코드를 만들어 볼게요.  
아이디어는: 얼굴 위치 → 피부 영역 추출 → 평균 색 계산입니다.

### 1. 필요한 라이브러리 설치

```
bash
pip install opencv-python numpy matplotlib
```

(얼굴 검출용 Haar Cascade는 OpenCV에 기본 포함되어 있습니다.)

### 2. 전체 코드 예시

```
python
import cv2
import numpy as np

def detect_faces(image_bgr):
    """Haar Cascade로 얼굴 영역(사각형) 검출"""
    gray = cv2.cvtColor(image_bgr, cv2.COLOR_BGR2GRAY)
    face_cascade = cv2.CascadeClassifier(
        cv2.data.haarcascades + "haarcascade_frontalface_default.xml"
    )
    faces = face_cascade.detectMultiScale(
        gray,
```

Explorer ...

Implementation Plan Updating Case Study File ⌂ ← → …

Implementation Plan 58 secs ago Review

### Implementation Plan - Add Vectors for High ID Content Nodes

#### Problem

Nodes with id\_content >= 10000 lack vector embeddings, preventing similarity calculation with TECH/TOPIC/THEME nodes and resulting in no cross-layer connections.

#### Solution

- Load Vector Data:** Get vectors from `interview_chunk_112525.xlsx` indexed by `id_content`
- Identify Nodes Without Vectors:** Find all interview nodes missing the `vector` property
- Assign Vectors:** Map `id_content` to vectors from the Excel file
- Calculate Similarity:** Compute cosine similarity

ok

Add Vector Embeddings for High ID Content Nodes  
I created a backup of the JSON file and wrote the script to add vectors and cross-layer connections. Now I am running the script.

Files Edited  
`add_vectors_and_connections.py`

Task

Progress Updates Expand all ↗  
1 Creating backup and loading data  
2 Running script to add vectors and connections

계속해

Error Agent execution terminated ...  
You have reached the quota limit for this

# 2.0의 함정: Prompting과 Vibe Coding을 넘어라

## Prompting, Vibe Coding

"일단 돌아가게만 해줘"

- 여러 나면 메시지 복사 → AI에 붙여넣기 반복
- 거대한 함수, 중복 코드, 테스트 부재
- 결과: 환각(Hallucination) 검증 불가, 재현성 부족
- 미래의 선택지(Option)를 희생하여 현재의 기능(Feature)만 구현

## Augmented Coding

"설계와 검증을 기자가 주도"

- 타이핑은 AI가, 설계·분해·리뷰는 기자가
- **Tidy First:** 구조 개선과 기능 개발 분리
- AI 코드가 팩트를 왜곡하지 않는지 테스트(**Test**) 필수
- 3.0으로 가기 위한 필수 역량

# 데이터 저널리즘 3.0

## 에이전트 설계 저널리즘

코드 리터러시를 바탕으로 취재·분석·검증 시스템을 직접 구축

### 시스템 구성요소

- 데이터 레이어: 기사 아카이브, 팩트체크 DB
- LLM 레이어: 자체 미세조정 모델, RAG
- 오케스트레이션: 에이전트(LangGraph, CrewAI)

### 특징

- **Agent:** 취재 보도 등 특수 역할 수행
- **Architect:** 워크플로우 설계 및 모니터링
- **Knowledge Engineer:** RAG DB 구축 관리

# | 데이터 저널리즘 3.0 사례

## 부패 감지 AI 'Funés' (Ojo Público, 2020)

- 내용: 공공 계약 부패 감시 시스템
- 방법: 계약 데이터에서 업체 간 관계와 정치 후원금을 분석하는 알고리즘 직접 설계
- 결과: 회사 이름 입력하면 '부패 위험 점수'를 계산하여 기자에게 취재 타겟 추천
- 의미: 일회성 보도가 아닌, 지속 가능한 내부 인텔리전스 시스템 구축



# | 데이터 저널리즘 3.0

## 페이스북 알고리즘 감시 (The Markup, 2021)

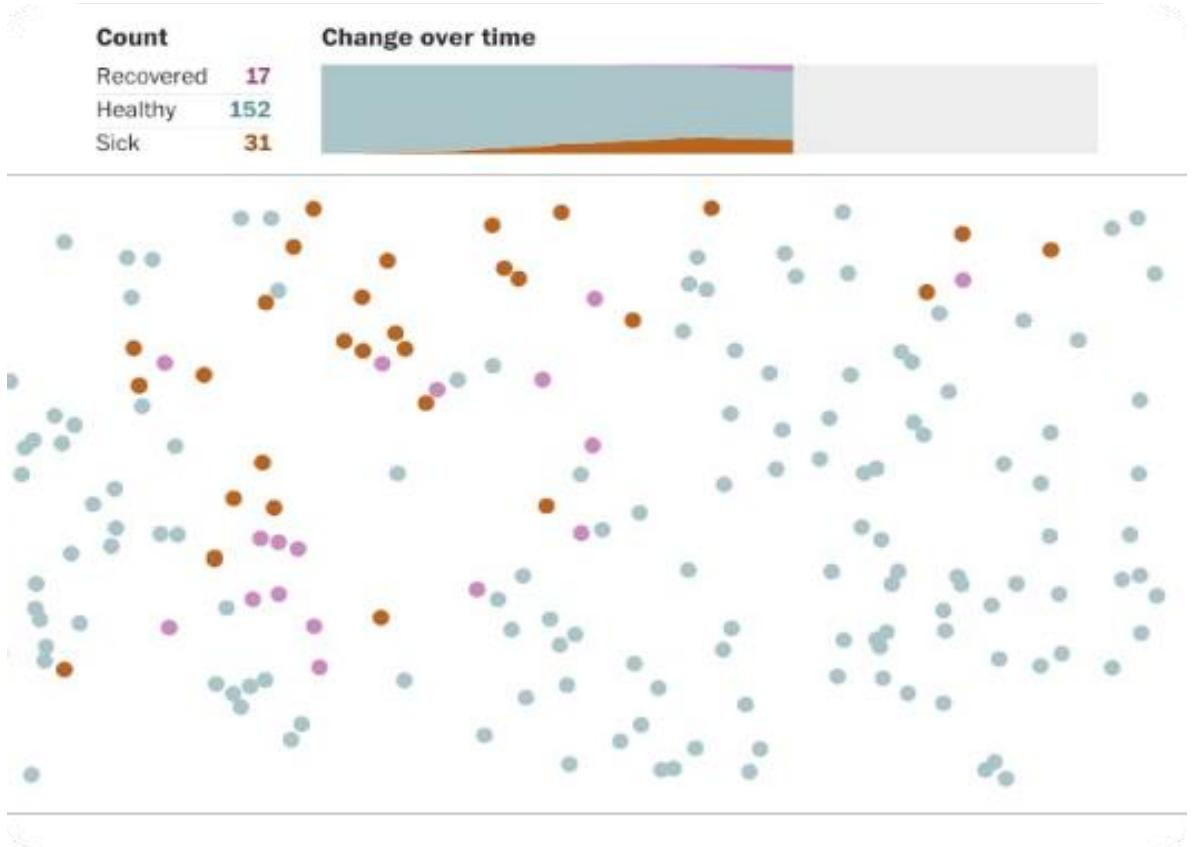
- **프로젝트:** Citizen Browser
- **방법:** 성별, 성별·인종·지역·정치성향별로 다양한 사용자 패널 모집해 페이스북 뉴스피드·추천 그룹·추천 페이지를 자동 수집 분석하는 시스템 개발
- **결과:** 추천 알고리즘의 편향성 폭로
- **의미:** 기자가 직접 갈 수 없는 곳에 디지털 대리인을 파견하는 인프라(IaaS) 구축



# | 데이터 저널리즘 3.0 사례

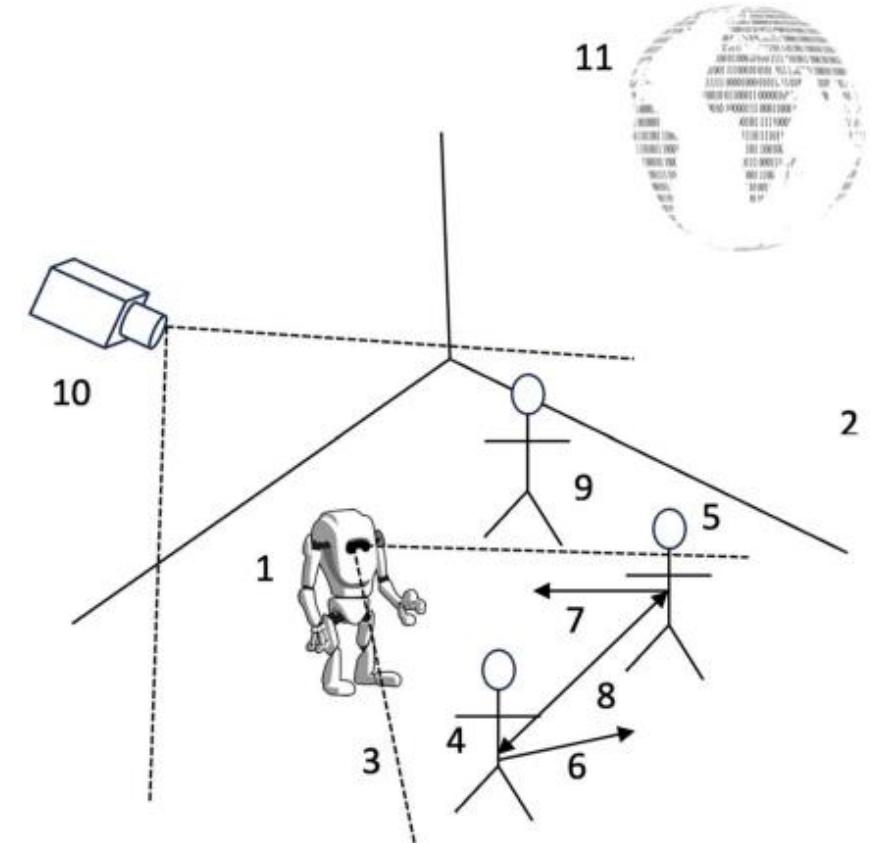
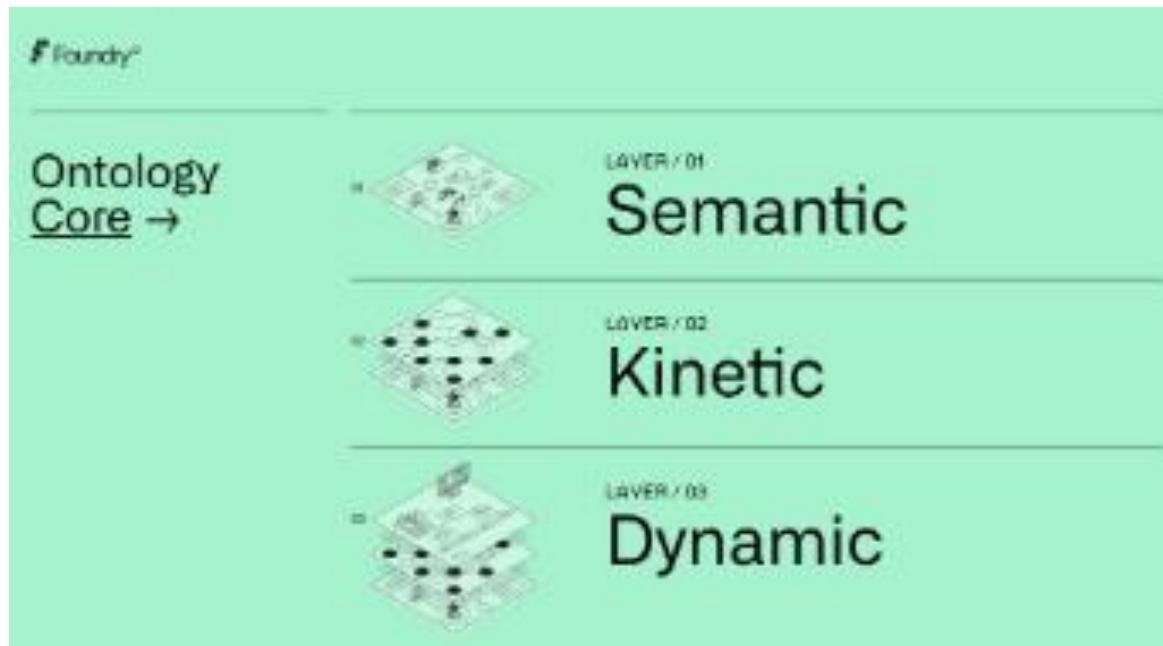
## 코로나 전파 시뮬레이터 (Washington Post, 2020)

- 내용: 사회적 거리두기의 효과 설명
- 방법: 공들이 튀어 다니는 시뮬레이션을 통해 전파 양상 시각화
- 효과: 백 마디 말보다 하나의 잘 짜인 시뮬레이션이 독자의 행동 변화 유도
- 유형: 독자 참여형 시뮬레이션



# (부록) 데이터에 대한 사고 전환

기존 용어	새로운 제안	전환의 핵심
Big Data	Networked Data	Quantity → Relation
Data Mining	Data Ecology	Extraction → Co-evolution



**Fig. 1** Highlighting aspects of robot ethnography (1) robot-ethnographer; (2) the scene being observed; (3) field of view of robot-ethnographer; (4) first person in view; (5) second person in view; (6) direction of travel of first person; (7) direction of travel of second person; (8) distance between first and second person; (9) third person—outside of field of view of robot-ethnographer; (10) remote sensor; (11) external data source—e.g. World Wide Web

# 종합 비교: 1.0 vs 2.0 vs 3.0

구분	데이터 저널리즘 1.0 (SaaS)	데이터 저널리즘 2.0 (PaaS)	데이터 저널리즘 3.0 (IaaS)
핵심 가치	정확한 분석 (Analyst)	효율적 생성 (Generator)	시스템 설계 (Architect)
개발 필요	없음 (Tool 사용)	중간 (Prompting)	필수 (Coding/System)
코딩 방식	N/A	Vibe Coding (위험)	Augmented Coding
결과물	정적 기사/그래픽	텍스트. 이미지.음성 대규모 분석	지속 가능한 시스템/DB

# 제언

“AI를 사용했느냐?”가 아니라,  
“AI 덕분에 더 깊이 취재했는가?”

---

"AI는 기자를 대체하지 않습니다.  
하지만 [시스템을 구축하는\(3.0\) 기자가](#)  
도구만 쓰는(1.0) 기자와 AI를 활용하는(2.0) 기자를 대체할 수 있습니다."