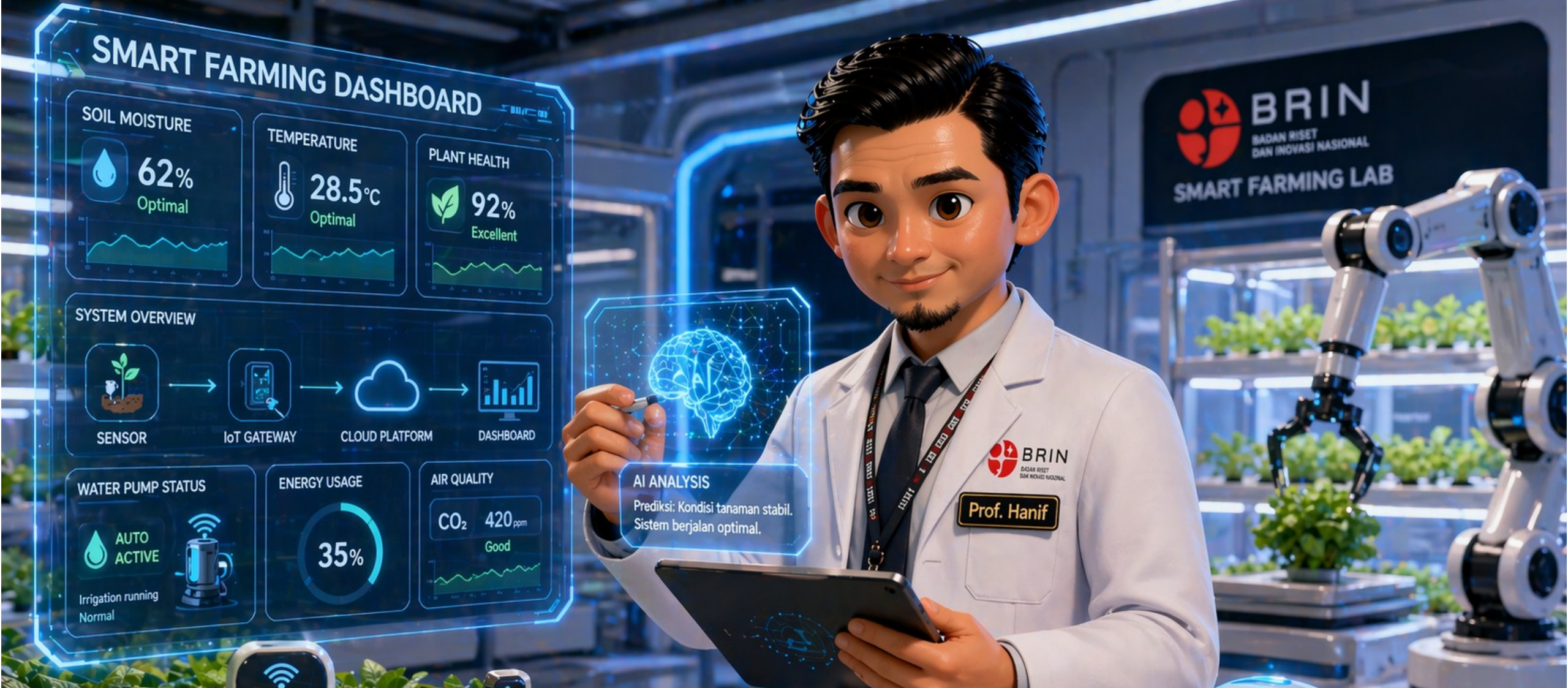




BRIN
BADAN RISET
DAN INOVASI NASIONAL

Pelatihan Teknis Substantif JF ADI Batch 4

PENGUMPULAN DAN PERSIAPAN DATA

Prof. Dr. Hanif Fakhurroja, S.Si., MT. ▪ Badan Riset dan Inovasi Nasional
Rabu, 29 Juli 2026

Prof. Dr. Hanif Fakhurroja, S.Si., MT.,

MCE, MCAAIF, CBDA, CRP, ITS AI, ITS DA, ITS PY, GAIF

Education:

- **Research Professor in Information System, Kepres No 31/M Tahun 2025**
- **Doctor** of Electrical Engineering and Informatics, **ITB**, 2015-2021
- **Magister** of Informatic Technology, **ITB**, 2008-2010
- **Bachelor** of Science in Computational Physics, **UNPAD**, 1999-2003
- Microsoft Certified Educator from Certiport (MCE)
- Microsoft Certified Azure Artificial Intelligence Fundamentals from Certiport (MCAAIF)
- Certified in Big Data Analyst with TÜV Rheinland Certified Qualification (CBDA)
- Certified in Remote Pilote (CRP) from Ministry of Transportation
- Certified Information Technology Specialist in Artificial Intelligence from Certiport (ITS AI)
- Certified Information Technology Specialist in Data Analytics from Certiport (ITS DA)
- Certified Information Technology Specialist in Python Programing from Certiport (ITS Py)
- Certified Generative AI Foundations from Certiport (GAIF)

Work Experiences:

- 2006 – 2021: **LIPI: Researcher**
- 2021 – Now: **BRIN: Research Professor/Peneliti Ahli Utama IV/c**
- 2017 – Now: **Telkom University: Professional Lecturer**
- 2018 – Now: **Trainer Big Data Analyst with TÜV Rheinland Certified Qualification**
- 2023 – Now: **Master Trainer Artificial Intelligence at B One Corporation**
- 2025 – Now: **AI Trainer at Pondasi, Hcelerate, and Krya**

Research Area:

Internet of Things, Big Data and Data Analytics, Machine Learning, Man-Machine Interaction, Intelligent Instrumentation Technology.



Hanif Fakhurroja



Hanif Fakhurroja



hanifoza@gmail.com



Hanif.fakhurroja



0818647004



PETIKAN
KEPUTUSAN PRESIDEN REPUBLIK INDONESIA
NOMOR 31/M TAHUN 2025
TENTANG
PENGANGKATAN DALAM DAN PEMBERHENTIAN DARI
JABATAN FUNGSIONAL AHLI UTAMA
DENGAN RAHMAT TUHAN YANG MAHA ESA
PRESIDEN REPUBLIK INDONESIA,

Menimbang : dst.
Mengingat : dst.
MEMUTUSKAN:
Menetapkan: dst.
KESATU : Mengangkat dalam jabatan fungsional ahli utama terhitung mulai tanggal pelantikan, masing-masing:
1. s.d. 64. dst.
65. Dr. Hanif Fakhurroja, S.Si., M.T., NIP 198006012006041004, Pembina Utama Muda (IV/c), sebagai Peneliti Ahli Utama pada Badan Riset dan Inovasi Nasional;
66. s.d. 139. dst.
dan kepada yang bersangkutan diberikan tunjangan jabatan fungsional sesuai dengan ketentuan peraturan perundang-undangan.
KEDUA : dst.
KETIGA : Keputusan Presiden ini mulai berlaku pada tanggal ditetapkan.
SALINAN dst.
PETIKAN Keputusan Presiden ini disampaikan kepada yang bersangkutan untuk digunakan sebagaimana mestinya.

Ditetapkan di Jakarta
pada tanggal 23 September 2025
PRESIDEN REPUBLIK INDONESIA,

ttd.

PRABOWO SUBIANTO

UNIT SEKRETARIAT
KEMENTERIAN SEKRETARIAT NEGARA RI
Unit Bidang Administrasi Aparatur,



SK No 232846 A



Scopus

This author profile is generated by Scopus

Fakhurroja, Hanif

Telkom University, Bandung, Indonesia • Scopus ID: 55869180100 • 0000-0002-2483-251X

Show all information

685

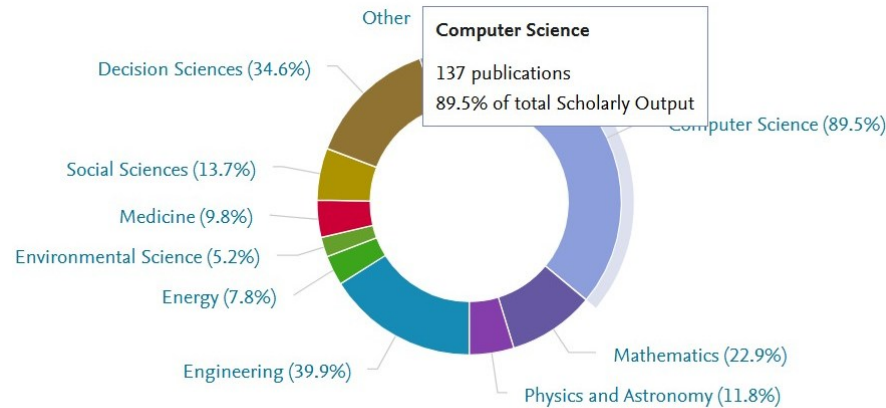
Citations by 559 documents

154

Documents

12

h-index



Quartiles

- Q1 (top 25%)
- Q2 (26% - 50%)
- Q3 (51% - 75%)
- Q4 (76% - 100%)

Publications



7
12
9
26

Publication share (%)

13.0
22.2
16.7
48.1



Hanif Fakhurroja

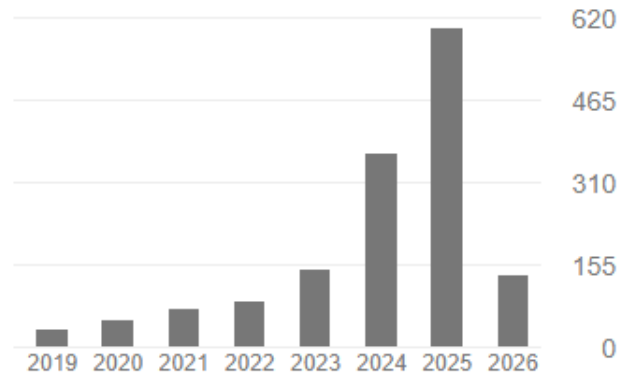
Telkom University and National Research and Innovation Agency Republic of Indonesia (BRIN)
Verified email at telkomuniversity.ac.id
Internet of Things Big Data and Data Analytics Machine Learning Information System IT Governance

Journals' Rankings					Report a Bug · Options · Vote for Next Features	
Q1	Q2	Q3	Q4	NA		
4% 4	4% 4	4% 4	2% 2	86% 74		
— based on 88 most cited papers in the last 10 years @						
					EXCITATION [™]	
☐ TITLE					CITED BY	YEAR
☐ Counterattacking cyber threats: A framework for the future of cybersecurity					389 [*]	2023
MF Saffira, M Lubis, H Fakhurroja						
Sustainability 15 (18), 13369						
☐ Design of real-time weather monitoring system based on mobile application using automatic weather station					87	2017
A Munandar, H Fakhurroja, M Rizqyawan, RP Pratama, JW Wibowo, ...						
2017 2nd International Conference on Automation, Cognitive Science, Optics ...						
☐ Pengelolaan Sistem Informasi Organisasi					57	2024
M Lubis, FE Susilawati, S Sudirman, I AP, A Kartini, F Adam, A Nurdin, ...						
Yayasan Kita Menulis						

Cited by

VIEW ALL

	All	Since 2021
Citations	1597	1427
h-index	20	17
i10-index	36	29



Authored 166 international publications and 14 books; holds 8 patents, 6 industrial design rights, and 10 copyrights.

Outline

- Akusisi data
- Pelabelan Data
- Prapemrosesan Data
- Data Outlier/Deteksi Outlier
Perekayasaan Fitur



Apa itu Akusisi Data?



BRIN
BADAN RISET
DAN INOVASI NASIONAL

BerAKHLAK
Berorientasi Pelayanan Akutabadi Kompeten
Harmonis Loyal Adaptif Kolaboratif

#bangga
melayani
bangsa

Mitos vs. Fakta: Audit Akuisisi Data

Tindakan

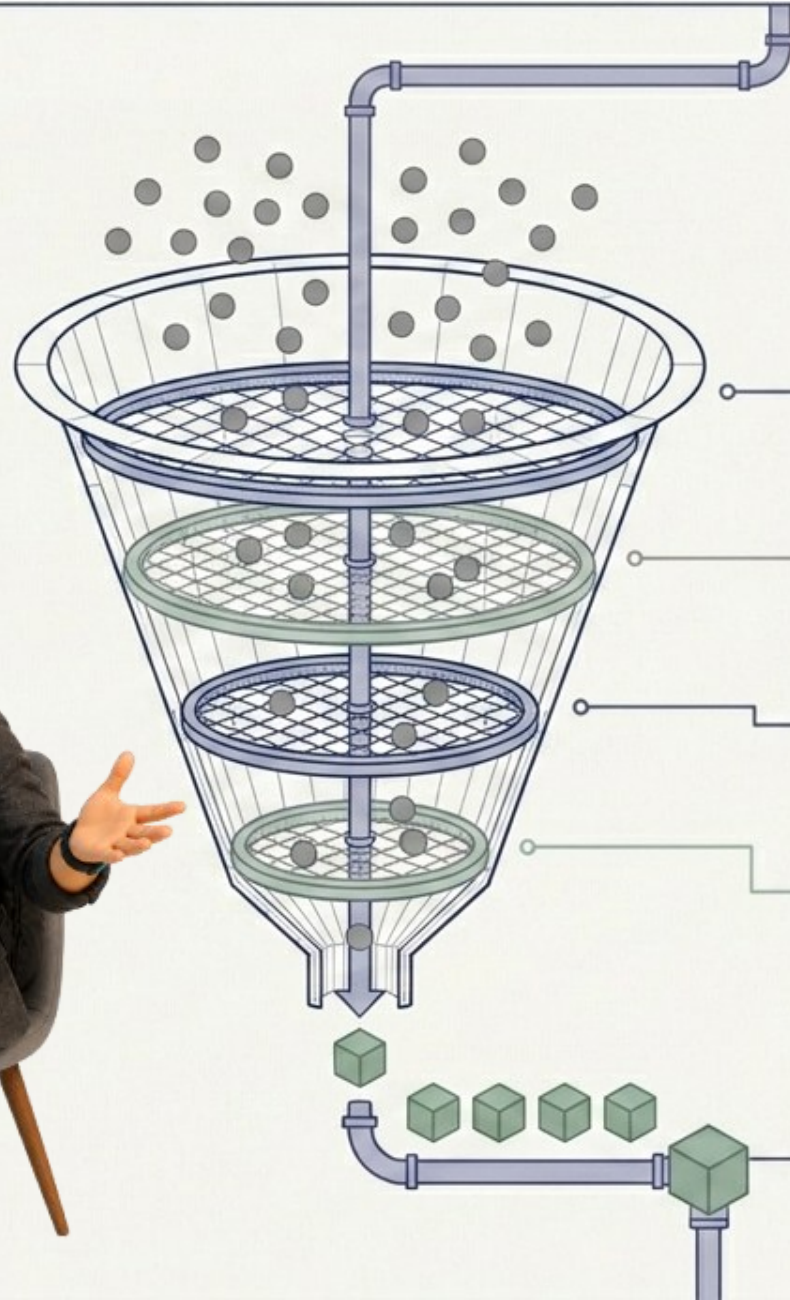
- Mengunduh dataset dari repositori internet.
- Membeli dataset dari penyedia pihak ketiga.
- Menerima sharing data dari kolega.
- Mengumpulkan data internet tanpa metodologi.
- Menyimpan data acak yang menarik dari web.

Pelanggaran

- ✗ Bukan hasil pengamatan fenomena alam/sosial/teknis.
- ✗ Tidak ada intervensi pengamatan langsung/tidak langsung.
- ✗ Bukan observasi orisinal.
- ✗ Tidak dirancang untuk menjawab pertanyaan spesifik.
- ✗ Proses tidak sistematis.

Akuisisi data ADI bukan sekadar 'mendapatkan data', melainkan memproduksi data melalui desain pengamatan yang terstruktur.

Empat Pilar Akuisisi Data



Pengamatan Fenomena:

Mengukur kejadian alam, sosial, atau teknis.

Intervensi Alat Bantu:

Menggunakan instrumen spesifik untuk menjamin data kuantitatif.

Keterbacaan Komputer (*Machine-Readable*):

Hasil akhir dapat diproses oleh sistem komputasi.

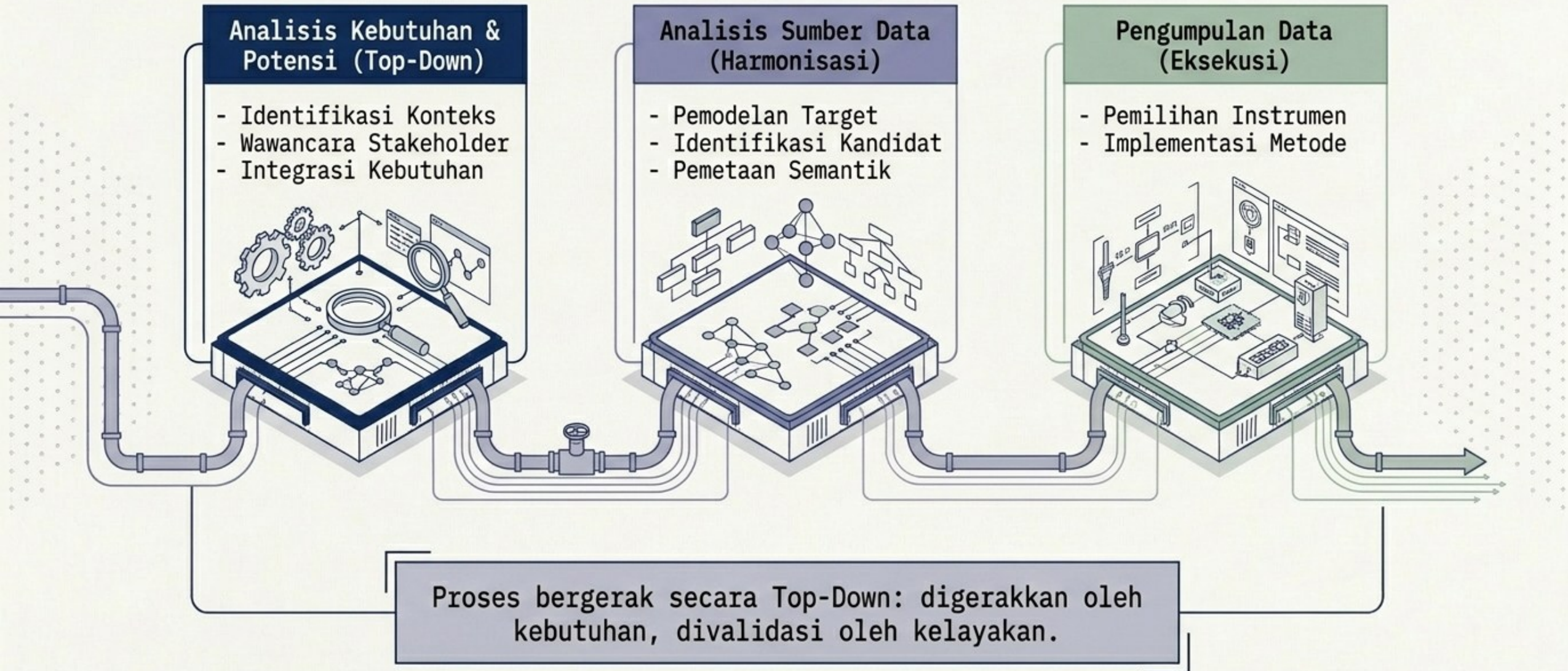
Berbasis Pertanyaan (*Question-Driven*):

Dikumpulkan sistematis untuk menjawab hipotesis.

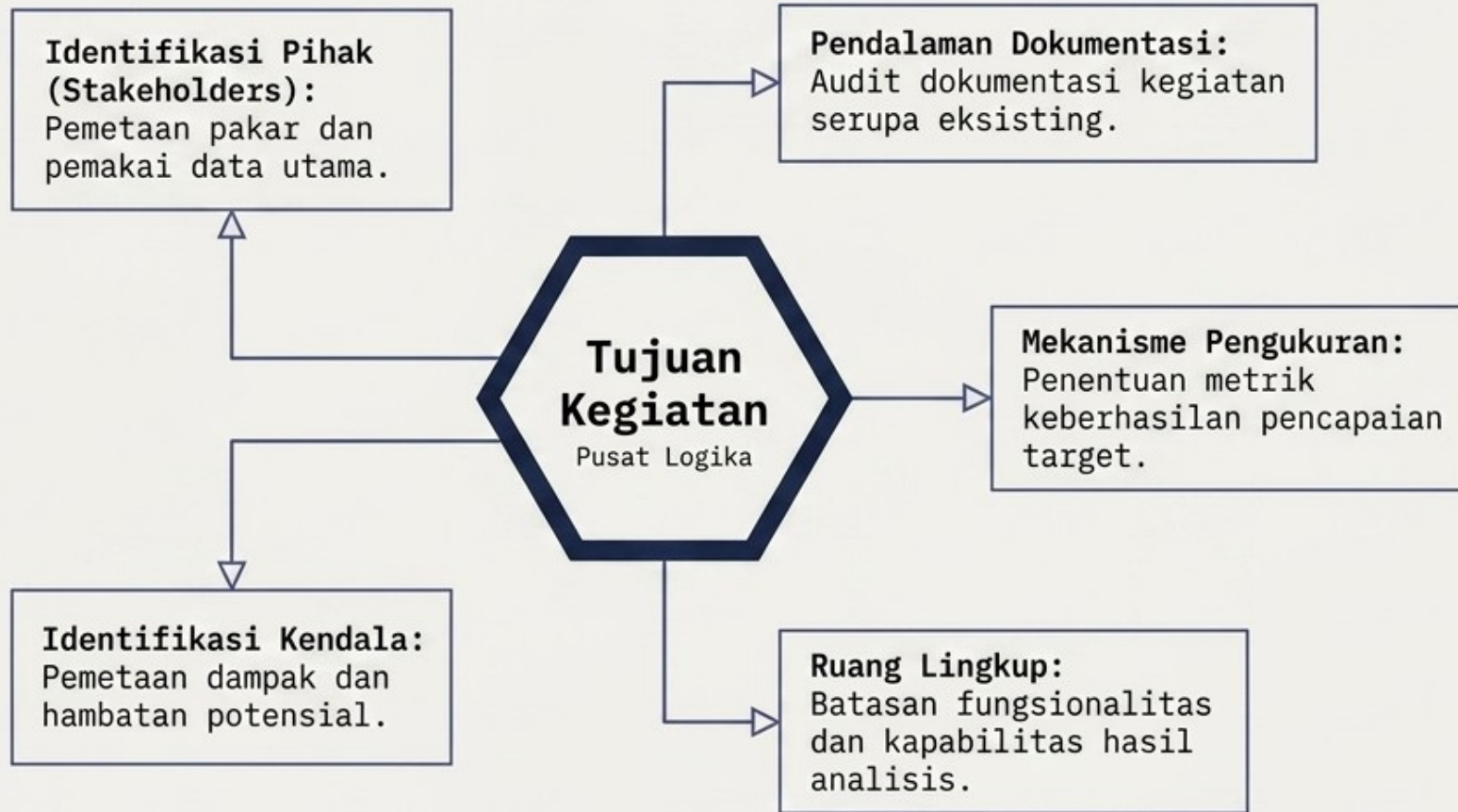
Akuisisi Data Ilmiah yang Valid



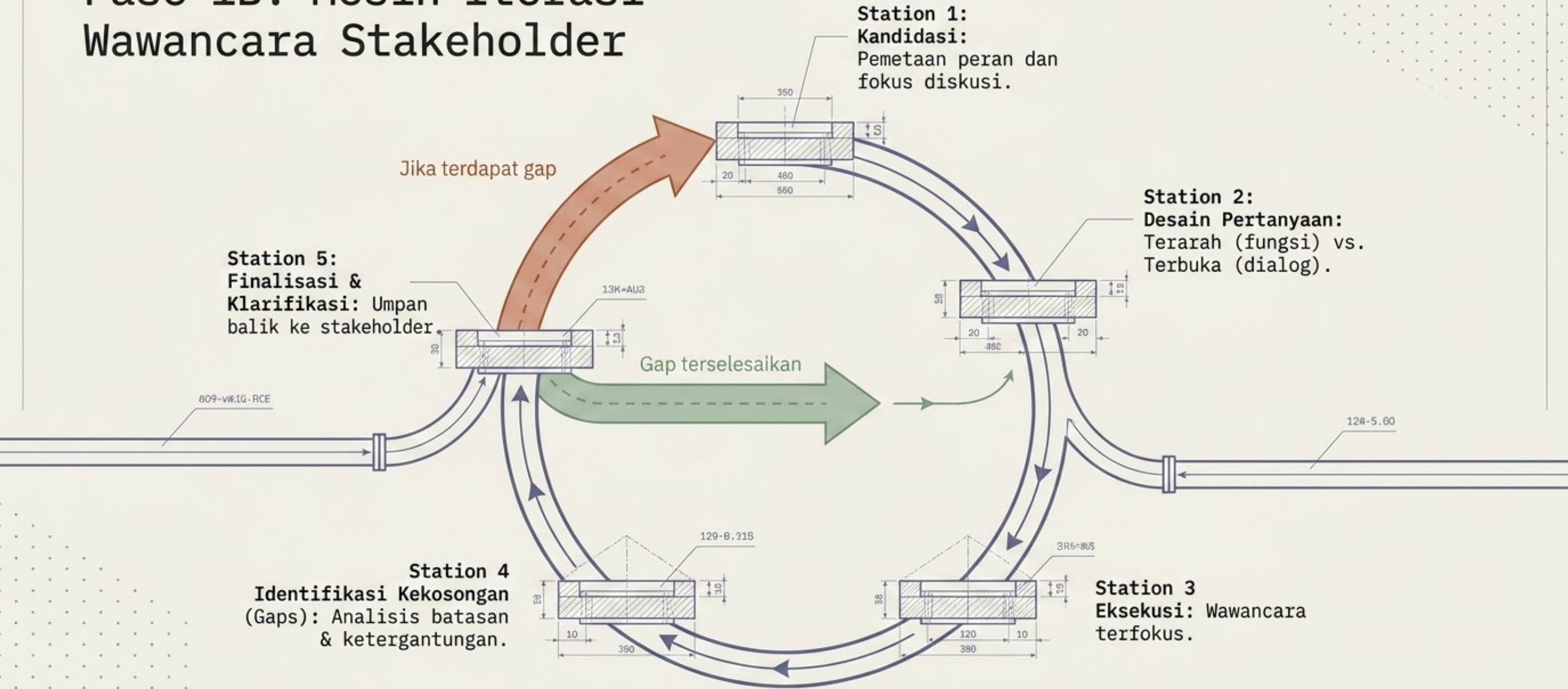
Arsitektur Makro Akuisisi Data



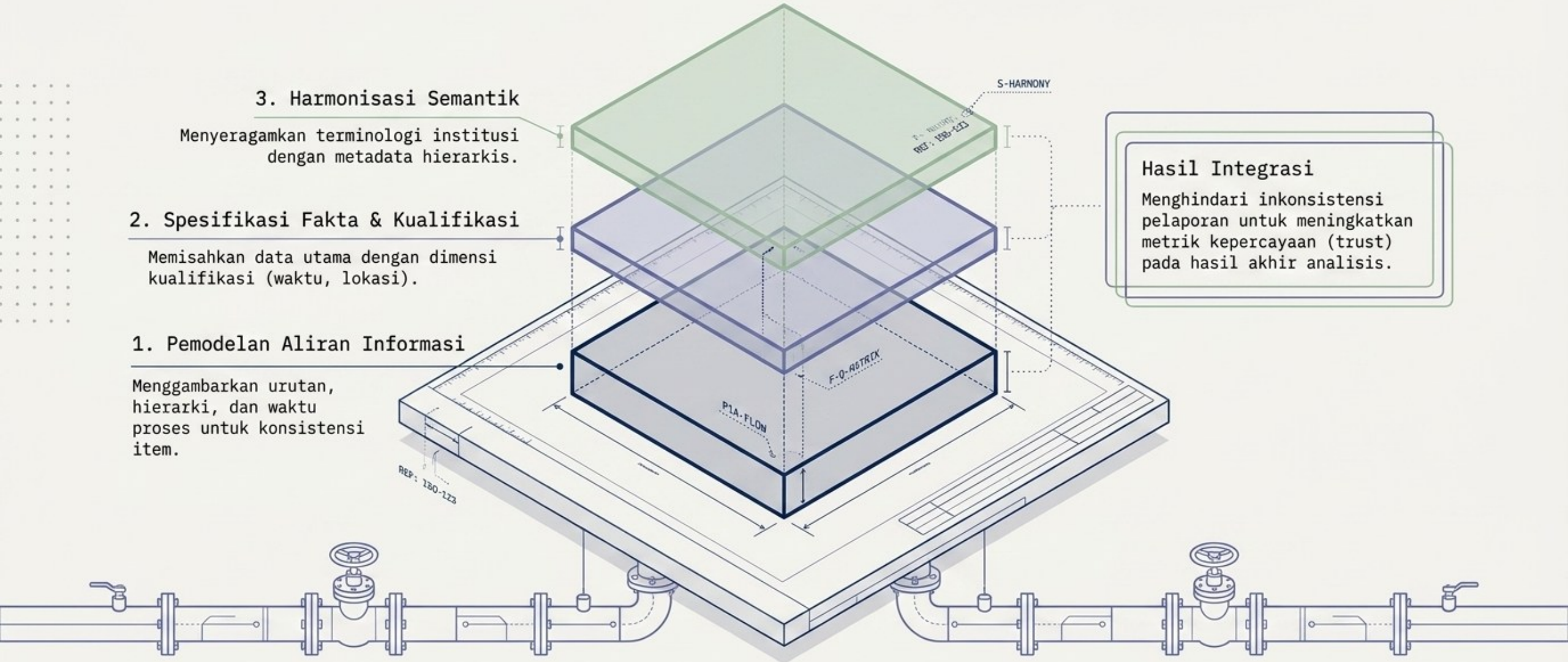
Fase 1A: Cetak Biru Konteks Kebutuhan



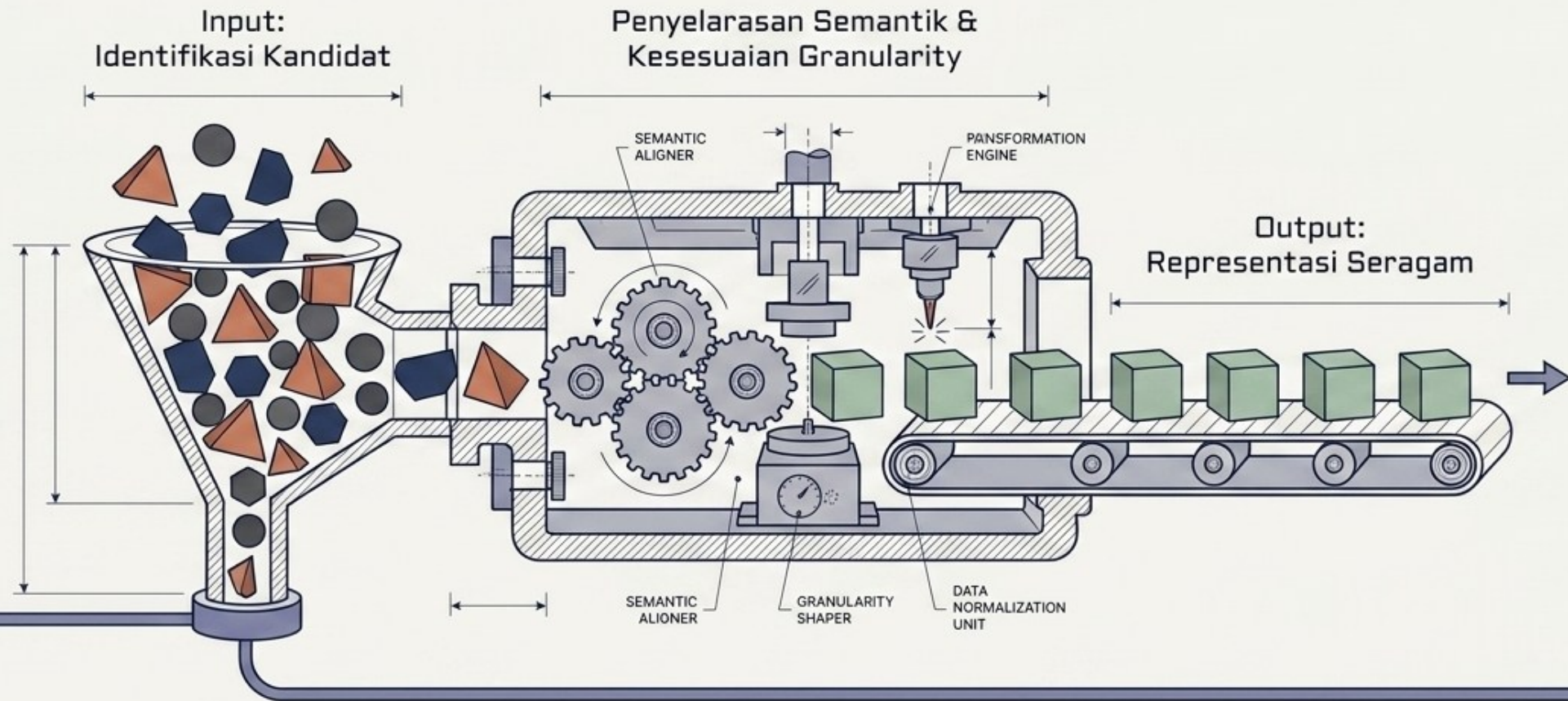
Fase 1B: Mesin Iterasi Wawancara Stakeholder



Fase 1C: Integrasi Kebutuhan Data



Fase 2: Mesin Harmonisasi Sumber & Target

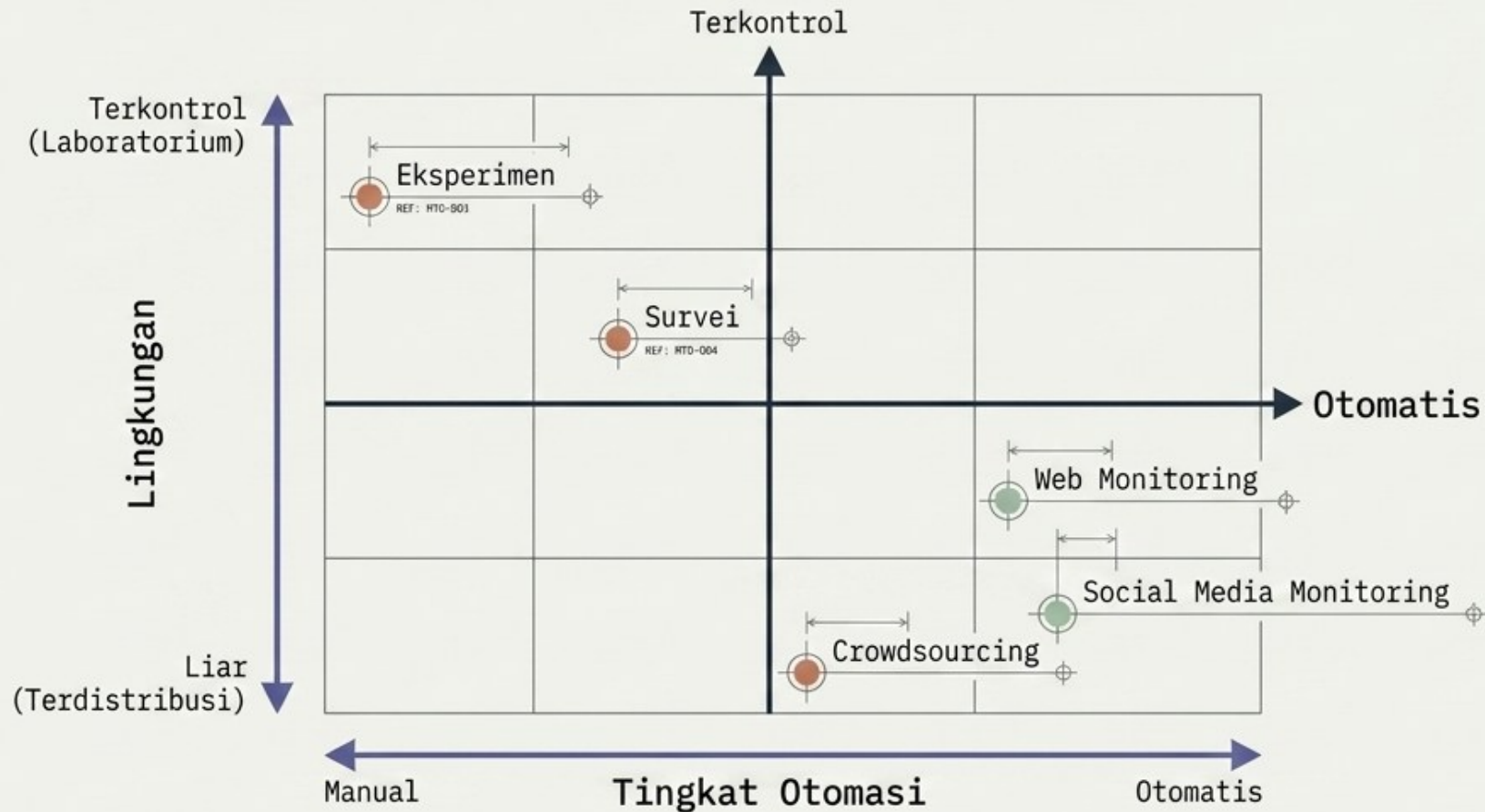


Evaluasi elemen & fakta komputasi dari berbagai sumber.

Menyatukan definisi data dan menyamakan resolusi tingkat detail.

Satu bentuk standar untuk analisis lanjutan.

Fase 3: Matriks Strategi Pengumpulan Data



Pemilihan metode bergantung pada karakteristik sumber data dan resolusi yang dibutuhkan oleh Target Mapping.

Article

Ontology-Driven Emotion Multi-Class Classification and Influence Analysis of User Opinions on Online Travel Agency

Putri Utami Rukmana ¹, Muhamman Lubis ^{1,*}, Hanif Fakhrrurroja ^{2,*}, Asriana ¹ and Alif Noorachmad Muttaqin ¹

¹ Master of Information System Study Program, School of Industrial Engineering, Telkom University, Main Campus (Bandung Campus), Jl. Telekomunikasi no. 1, Bandung 40257, West Java, Indonesia; putriutami@student.telkomuniversity.ac.id (P.U.R.); asriana@student.telkomuniversity.ac.id (A.); weaboo@student.telkomuniversity.ac.id (A.N.M.)

² Research Center for Smart Mechatronics, National Research and Innovation Agency, Jl. Sangkuriang, Dago, Bandung 40135, West Java, Indonesia

* Correspondence: muhammanlubis@telkomuniversity.ac.id (M.L.); haniff@telkomuniversity.ac.id (H.F.)

Abstract

The rise in social media has transformed Online Travel Agencies (OTAs) into platforms where users actively share their experiences and opinions. However, conventional opinion mining approaches often fail to capture nuanced emotional expressions or connect them to user influence. To address this gap, this study introduces an ontology-driven opinion mining framework that integrates multi-class emotion classification, aspect-based analysis, and influence modeling using Indonesian-language discussions from the social media platform X. The framework combines an OTA-specific ontology that formally represents service aspects such as booking support, financial, platform experience, and event with fine-tuned IndoBERT for emotion recognition and sentiment polarity detection, and Social Network Analysis (SNA) enhanced by entropy weighting and TOPSIS to quantify and rank user influence. The results show that the fine-tuned IndoBERT performs strongly with respect to identification and sentiment polarity detection, with moderate results for multi-class emotion classification. Emotion labels enrich the ontology by linking user opinions to their affective context, enabling the deeper interpretation of customer experiences and service-related issues. The influence analysis further reveals that structural network properties, particularly betweenness, closeness, and eigenvector centrality, serve as the primary determinants of user influence, while engagement indicators act as discriminative amplifiers that highlight users whose content attains high visibility. Overall, the proposed framework offers a comprehensive and interpretable approach to understanding public perception in Indonesian-language OTA discussions. It advances opinion mining for low-resource languages by bridging semantic ontology modeling, emotional understanding, and influence analysis, while providing practical insights for OTAs to enhance service responsiveness, manage emotional engagement, and strengthen digital communication strategies.

Keywords: ontology-based sentiment analysis; multi-label emotion classification; IndoBERT; social network analysis; entropy weighting; TOPSIS; online travel agency; social media opinion mining

1. Introduction

The rapid growth of user-generated content on social media has fundamentally changed how companies monitor and respond to public opinion. Platforms



Academic Editor: Gianluigi Ferrari

Received: 13 October 2025

Revised: 20 November 2025

Accepted: 20 November 2025

Published: 17 December 2025

Citation: Rukmana, P.U.; Lubis, M.; Fakhrrurroja, H.; Asriana; Muttaqin, A.N. Ontology-Driven Emotion Multi-Class Classification and Influence Analysis of User Opinions on Online Travel Agency. *Future Internet* **2025**, *17*, 582. <https://doi.org/10.3390/fi17120582>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Article

Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm

Ahmad Nahid Ma'aly ¹, Dita Pramesti ^{1,*}, Ariadani Dwi Fathurahman ¹ and Hanif Fakhrrurroja ²

¹ School of Industrial Engineering, Telkom University, Bandung 40257, West Java, Indonesia; nahidm@student.telkomuniversity.ac.id (A.N.M.); ariadanidf@student.telkomuniversity.ac.id (A.D.F.); haniff@telkomuniversity.ac.id (H.F.)

² Research Center for Smart Mechatronics, National Research and Innovation Agency, Bandung 40135, West Java, Indonesia

* Correspondence: ditapramesti@telkomuniversity.ac.id

Abstract: Presidential elections are an important political event that often trigger intense debate. With more than 139 million users, YouTube serves as a significant platform for understanding public opinion through sentiment analysis. This study aimed to implement deep learning techniques for a multi-label sentiment analysis of comments on YouTube videos related to the 2024 Indonesian presidential election. Offering a fresh perspective compared to previous research that primarily employed traditional classification methods, this study classifies comments into eight emotional labels: anger, anticipation, disgust, joy, fear, sadness, surprise, and trust. By focusing on the emotional spectrum, this study provides a more nuanced understanding of public sentiment towards presidential candidates. The CRISP-DM method is applied, encompassing stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment, ensuring a systematic and comprehensive approach. This study employs a dataset comprising 32,000 comments, obtained via YouTube Data API, from the KPU and Najwa Shihab channels. The analysis is specifically centered on comments related to presidential candidate debates. Three deep learning models—Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and a hybrid model combining CNN and Bi-LSTM—are assessed using confusion matrix, Area Under the Curve (AUC), and Hamming loss metrics. The evaluation results demonstrate that the Bi-LSTM model achieved the highest accuracy with an AUC value of 0.91 and a Hamming loss of 0.08, indicating an excellent ability to classify sentiment with high precision and a low error rate. This innovative approach to multi-label sentiment analysis in the context of the 2024 Indonesian presidential election expands the insights into public sentiment towards candidates, offering valuable implications for political campaign strategies. Additionally, this research contributes to the fields of natural language processing and data mining by addressing the challenges associated with multi-label sentiment analysis.

Keywords: deep learning; multi-label sentiment analysis; Indonesian 2024 presidential elections

1. Introduction

Presidential elections represent the pinnacle of the political process, serving as the decisive mechanism through which leaders, responsible for shaping state policies in the forthcoming term, are selected. The enthusiasm of the community in the moment of the presidential election has always been in the spotlight and triggers intense debates such as debates related to political ideology, social and economic issues, dissatisfaction with the previous government, identity and religious issues, social media, and candidates. In 2024, Indonesia held its presidential election, selecting a president and vice president. More than two hundred million voters participated in the election on 14 February 2024 to elect a President. Initially, in the social media era, people seek information through posts



Citation: Ma'aly, A.N.; Pramesti, D.; Fathurahman, A.D.; Fakhrrurroja, H. Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm. *Information* **2024**, *15*, 705. <https://doi.org/10.3390/info15110705>

Academic Editors: Ioannis Tsamperidis, Eleni Vrochidou and George A. Papakostas

Received: 29 September 2024

Revised: 26 October 2024

Accepted: 29 October 2024

Published: 5 November 2024



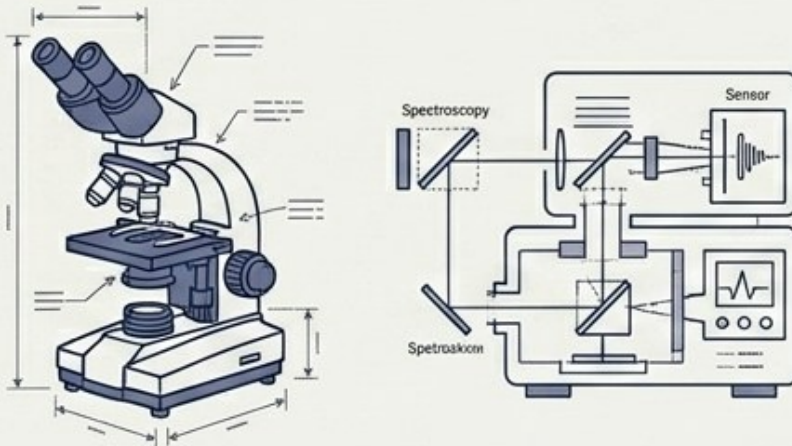
Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Contoh penelitian yang datanya diambil dari Social Media



Instrumen Lingkungan Terkontrol

Eksperimen (Observasi Fisik)



- Lingkungan: Laboratorium / Terkontrol.
- Alat Sederhana: Mikroskop, Stetoskop, Atmometer.
- Alat Sederhana: Mikroskop, Stetoskop, Atmometer.
- Alat Kompleks: Spektroskopi, Genome Sequencer.
- Output: Numerik, gambar, audio, video.

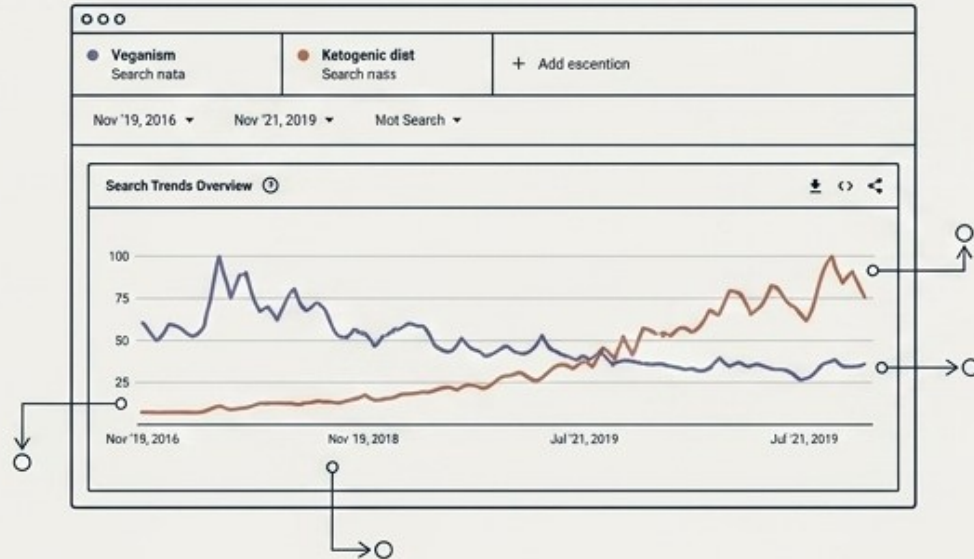
Survei (Observasi Sosial/Opini)



- Metode: Kuesioner elektronik, Focus Group.
- Platform Digital: Google Forms, Survey Monkey, Lime Survey.
- Keunggulan: Instalasi server mandiri (Lime Survey) untuk keamanan data institusi.

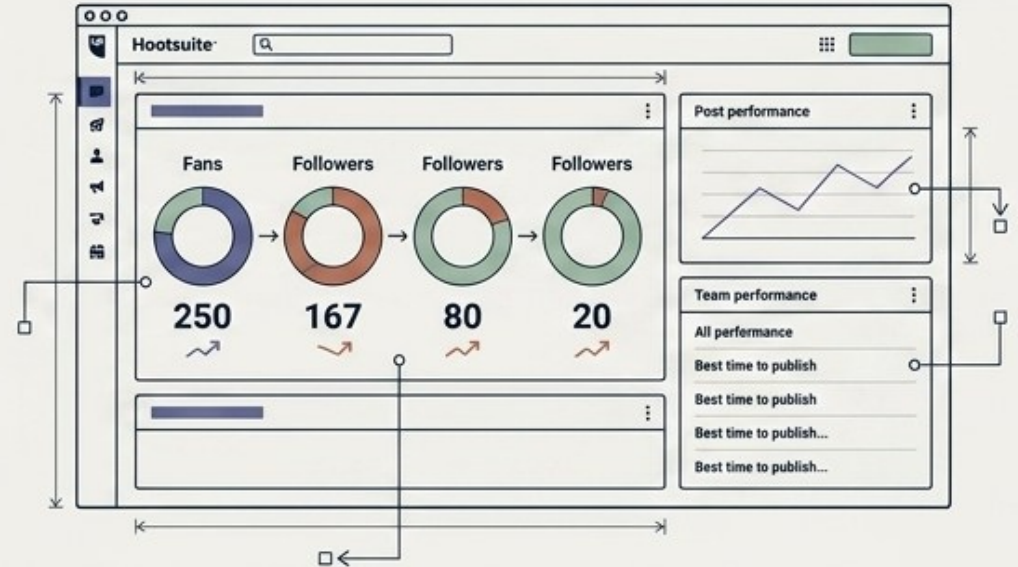
Ekstraksi Data di Ekosistem Digital

Web Monitoring (Penelusuran Sistematis)



- **Tantangan:** Volume raksasa (Big Data).
- **Instrumen:** Google Alerts (notifikasi keyword) & Google Trends (analisis popularitas spasial-temporal).

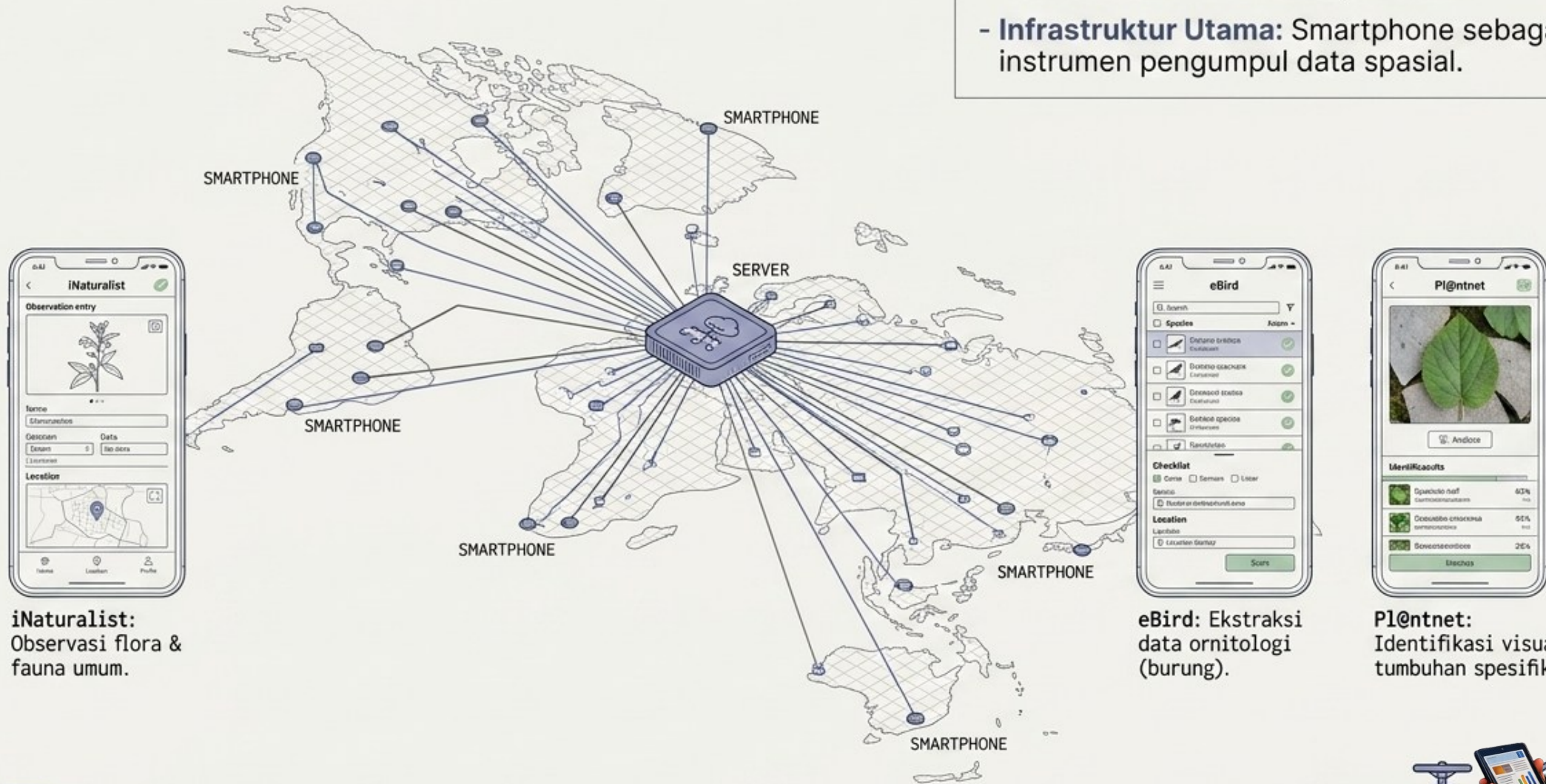
Social Media Monitoring (Observasi Tidak Langsung)



- **Target Primer:** Teksual, Gambar, Sentimen (Like/Comment).
- **Target Sekunder (Metadata):** Geolokasi, cap waktu (timestamp), data perangkat.
- **Instrumen:** Dashboard analitik pihak ketiga terotomasi.

Citizen Science: Sensor Terdistribusi Masif

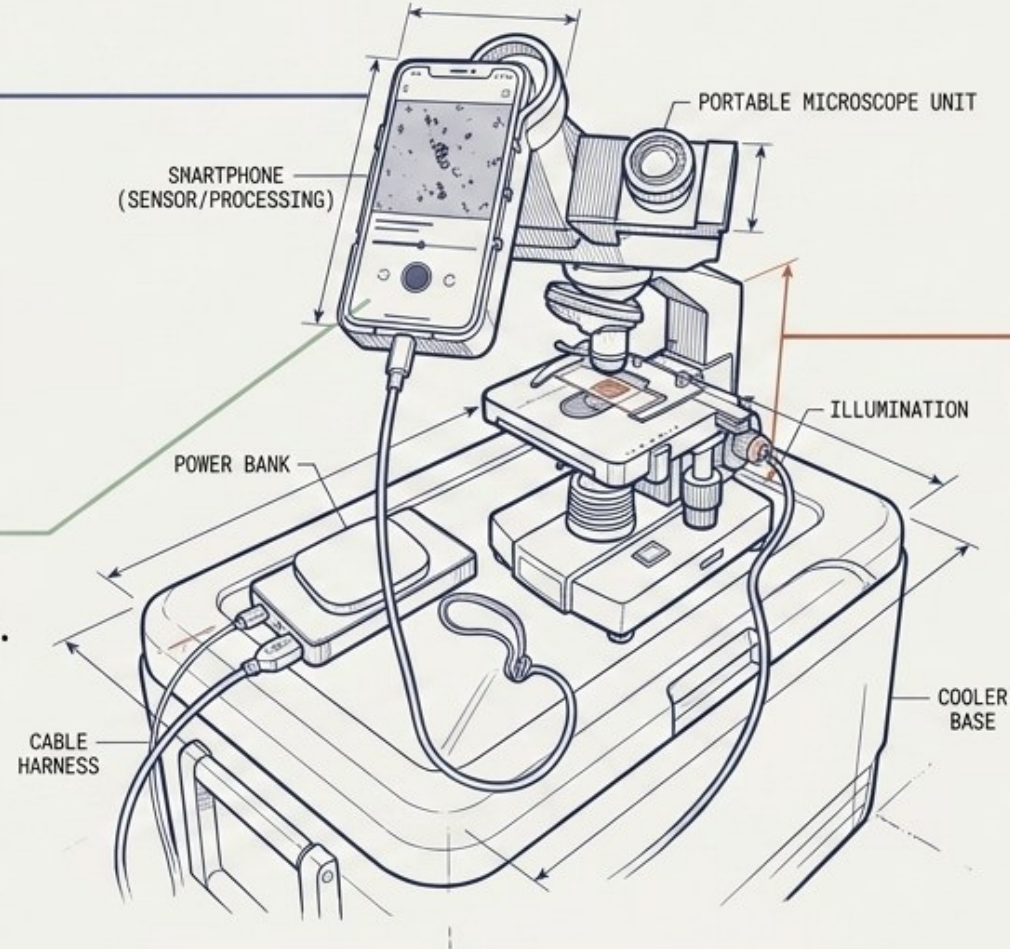
- **Definisi:** Memanfaatkan partisipasi masyarakat secara sukarela untuk eksplorasi sains.
- **Infrastruktur Utama:** Smartphone sebagai instrumen pengumpul data spasial.



Arsitektur Hybrid: Integrasi Hardware & Software

Problem: Kapabilitas sensor smartphone standar tidak mencukupi untuk analisis mikroskopis.

Solusi Software: Aplikasi smartphone untuk merekam, memproses, dan mengirim data.



Solusi Hardware: Mikroskop portabel untuk kuantifikasi alga pada sampel air.

Takeaway: Analisis Data Ilmiah harus mampu mendesain rig akuisisi gabungan ketika instrumen tunggal menemui batasan teknis.



Daftar Periksa (Checklist) Eksekusi Akuisisi



Validasi Definisi: Apakah metode ini mengobservasi fenomena atau hanya menyalin data?



Audit Kebutuhan: Apakah aliran informasi dan metrik stakeholder sudah terpetakan?



Kesesuaian Semantik: Apakah granularity dan struktur data target sudah selaras dengan sumber?

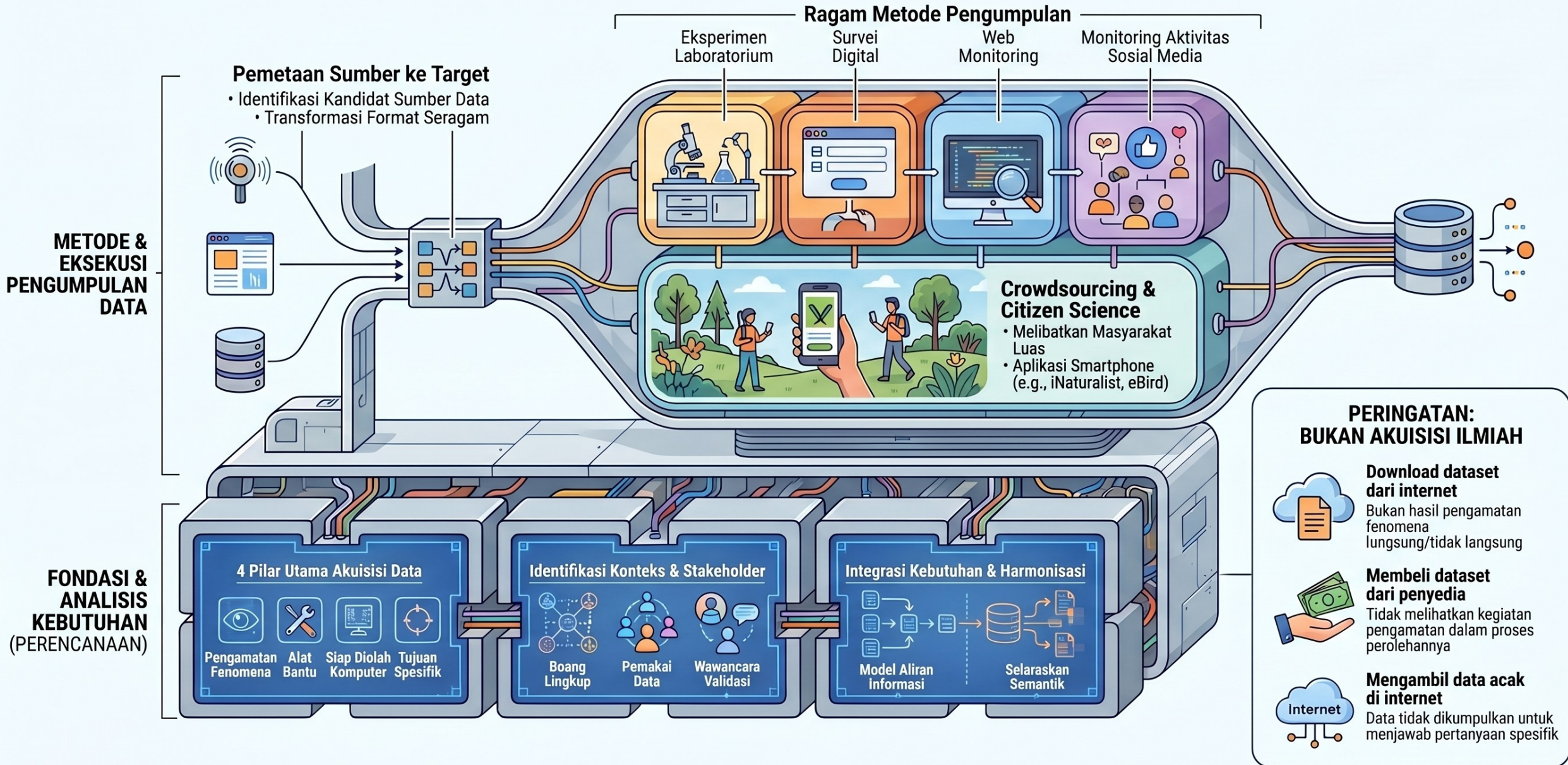


Kalibrasi Instrumen: Apakah instrumen (Eksperimen/Survei/Web/Citizen Science) tepat untuk lingkungan?



Kualitas analisis esok hari ditentukan oleh presisi akuisisi hari ini.
Eksekusi pipa data Anda.

Arsitektur Akuisisi Data Ilmiah: Dari Fondasi hingga Koleksi



Apa itu Pelabelan Data?

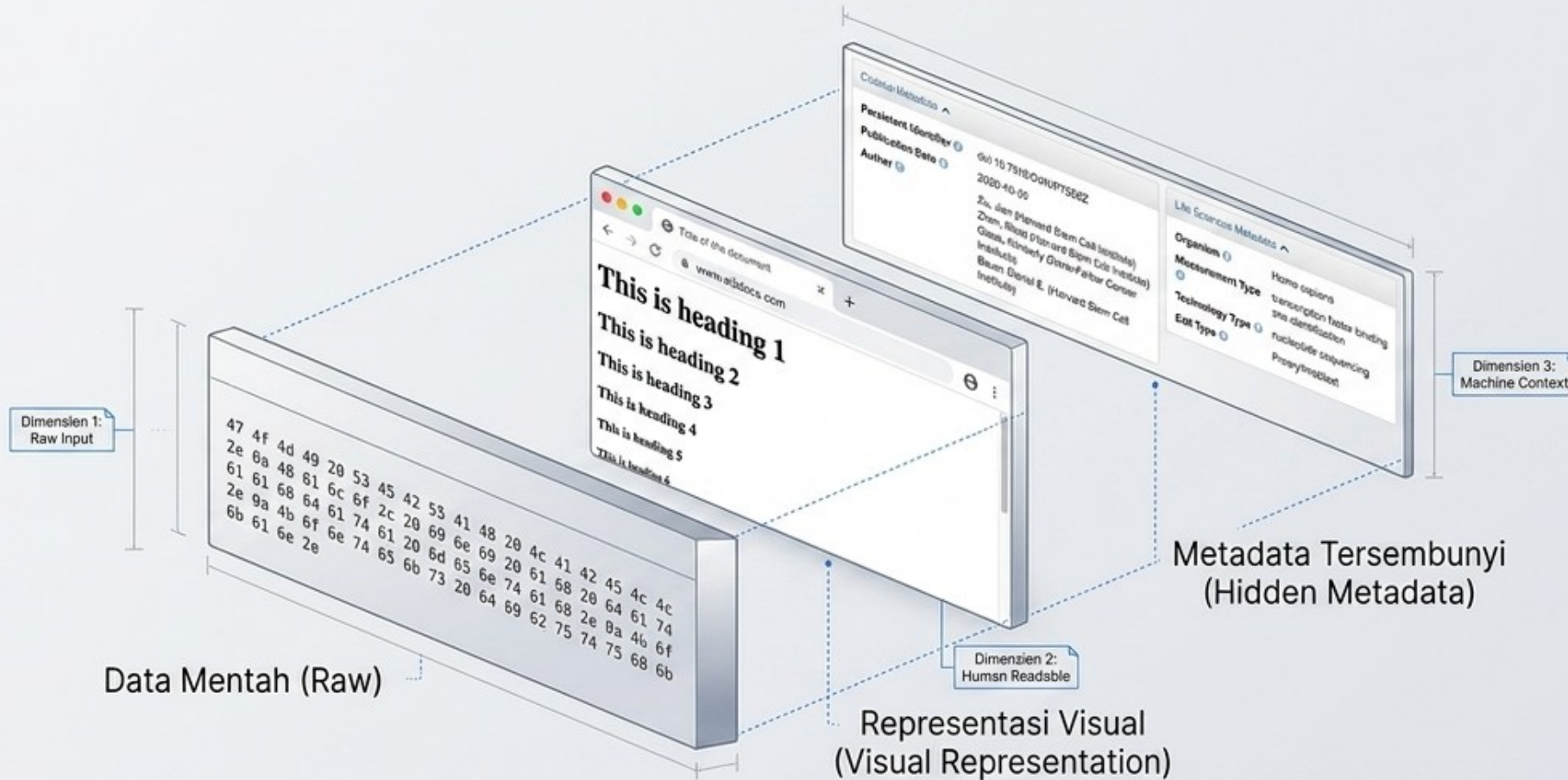


BRIN
BADAN RISET
DAN INOVASI NASIONAL

BerAKHLAK
Berorientasi Pelayanan Akuntabel Kompeten
Harmonis Loyal Adaptif Kolaboratif

#bangga
melayani
bangsa

Anatomi Sebuah Label: Memberikan Konteks pada Dimensi Data

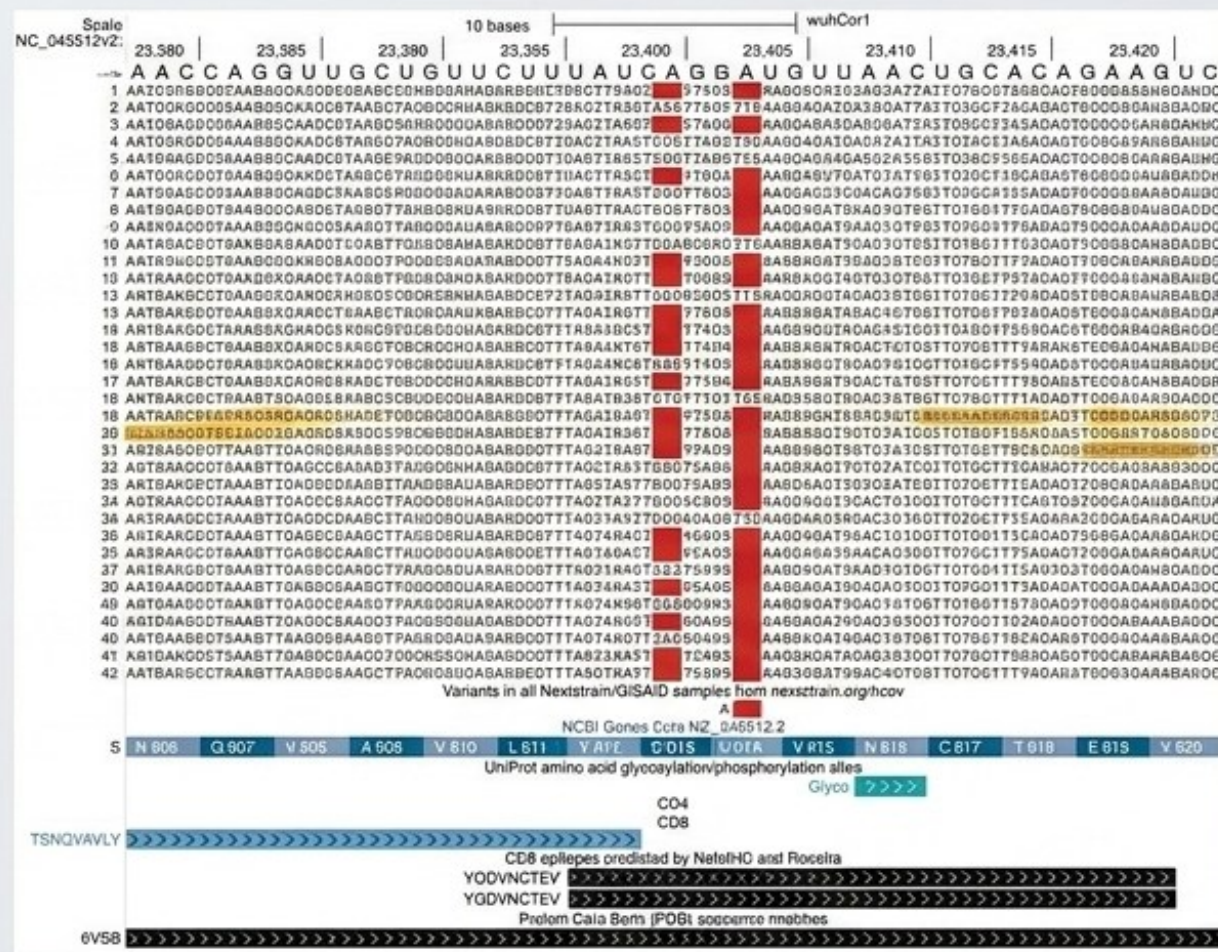


Pelabelan data (*data annotation*) adalah proses menyematkan konteks agar informasi mentah dapat dicerna dengan seragam, logis, dan tanpa ambiguitas oleh manusia maupun mesin.



A blue line drawing of a human head in profile, facing right. Inside the head, the brain is depicted with wavy lines. A horizontal arrow points from the center of the brain to a small, white, tag-like shape on the right. The tag has a small 'D' on it.

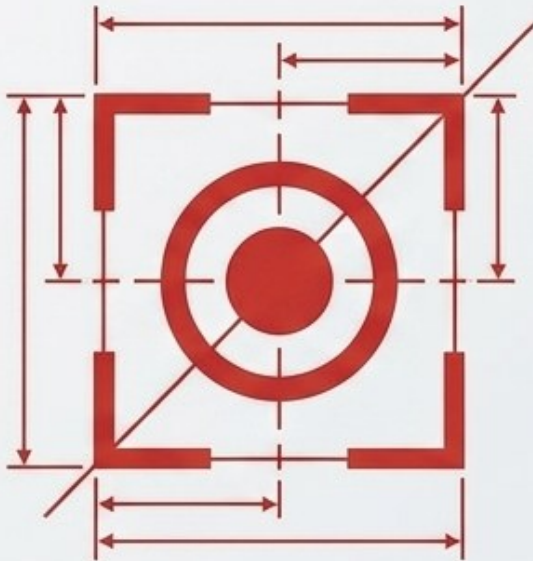
Supervised Learning



Tantangan Eksekusi: Skala data raksasa (misal: urutan genome) mustahil dianalisis manual tanpa algoritma yang dilatih dengan pelabelan presisi tinggi.

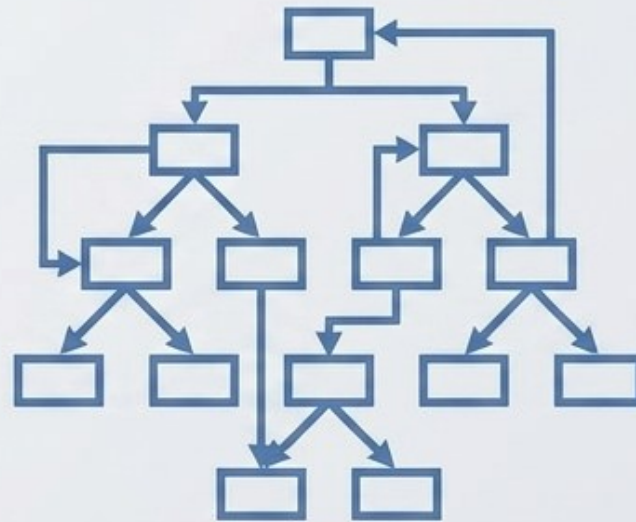
Arsitektur 3 Pilar Objektif Pelabelan Data

Pilar 1: Visual (Identifikasi Objek)



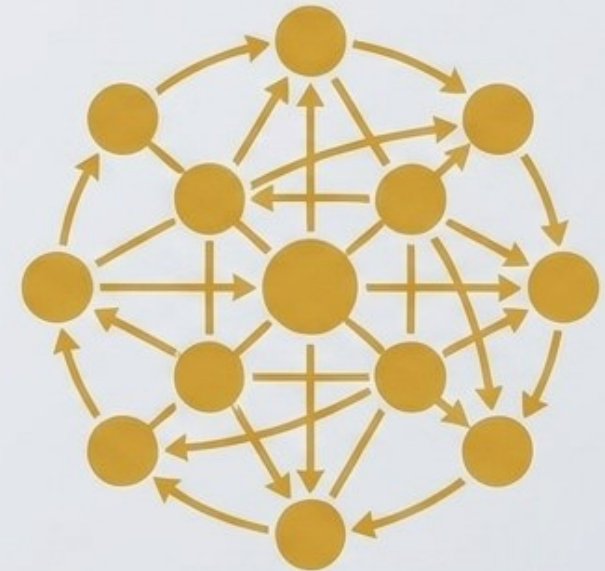
Menentukan koordinat dan luasan entitas fisik dalam ruang visual (Region-of-Interest).

Pilar 2: Sintaksis (Relasi & Struktur)

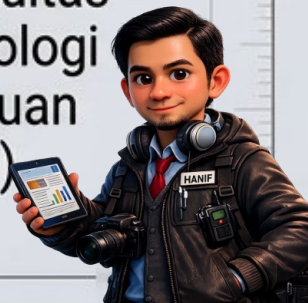


Memetakan peran gramatikal dan hubungan antar elemen (Part-of-Speech & Sentimen).

Pilar 3: Semantik (Makna & Pengetahuan)



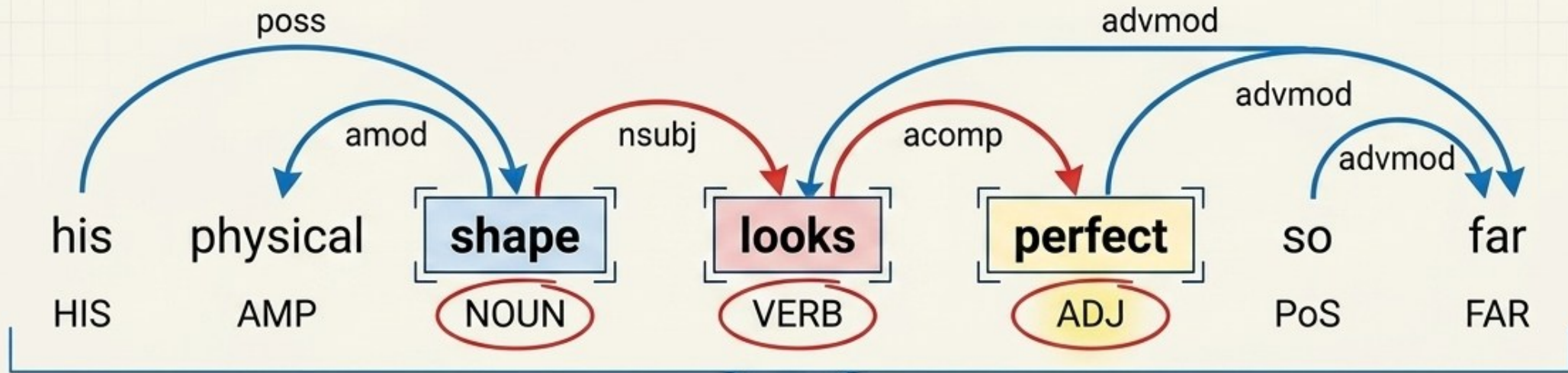
Mengeliminasi ambiguitas melalui integrasi Ontologi dan Basis Pengetahuan (Knowledge Base)



The image consists of two side-by-side screenshots. The left screenshot shows a microscopic view of a plant specimen, likely a leaf, with several small, labeled regions of interest. The labels are in Indonesian, such as 'Konteks Mikro (Spesimen Ilmiah)'. The right screenshot shows a macro view of a house, with several large, labeled regions of interest. The labels are in Indonesian, such as 'Konteks Makro (Ruang Terbuka)'.

Mesin tidak melihat gambar; mereka melihat piksel. Label mengubah blok piksel menjadi Region-of-Interest (ROI) yang memiliki batas, koordinat, dan kelas identitas (Multi-label).

Pilar 2: Dekonstruksi Sintaksis (Struktur & Relasi)



➔ Sentimen Ekstraksi: Positif ✓

Dalam Pemrosesan Bahasa Alami (NLP), **pelabelan Part-of-Speech (PoS)** membangun **pohon** kalimat. Pohon ini mendeteksi bukan hanya apa yang dikatakan, tetapi bagaimana **elemen tersebut berinteraksi**—misalnya, menentukan arah sentimen secara algoritmik.

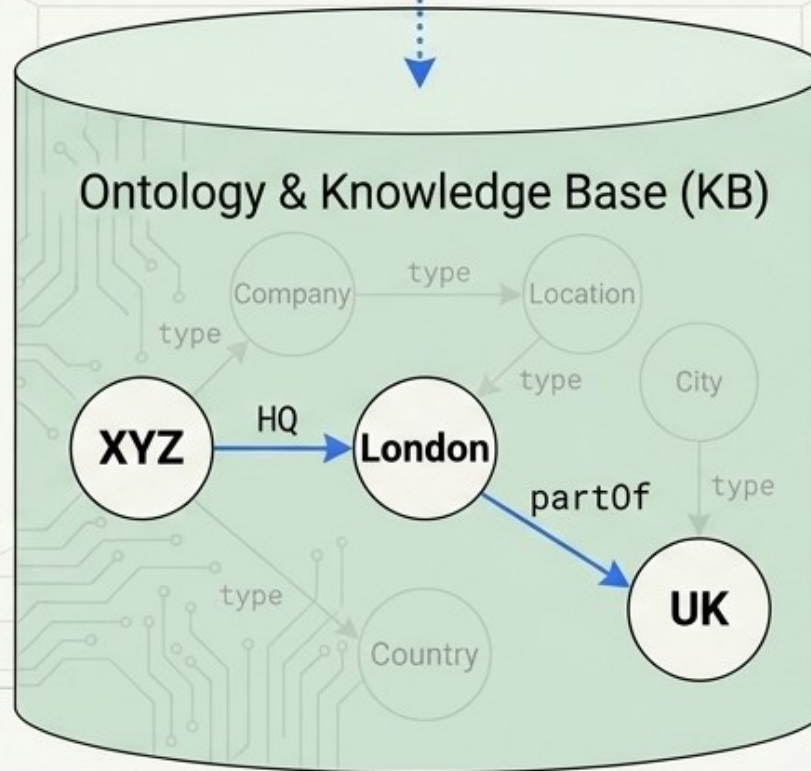


Pilar 3: Resolusi Ambiguitas Semantik (Makna)

XYZ announced profits in Q3, planning to build a \$120M plant in Bulgaria...

Masalah:

Mesin melihat string karakter yang sama (contoh: 'Jaguar' hewan vs. 'Jaguar' mobil).



Solusi:

Pelabelan Semantik menghubungkan entitas lokal ke hierarki Ontologi global. Pencarian presisi mengabaikan 'XYZ' sebagai nama orang, dan memahaminya secara absolut sebagai entitas perusahaan di London.

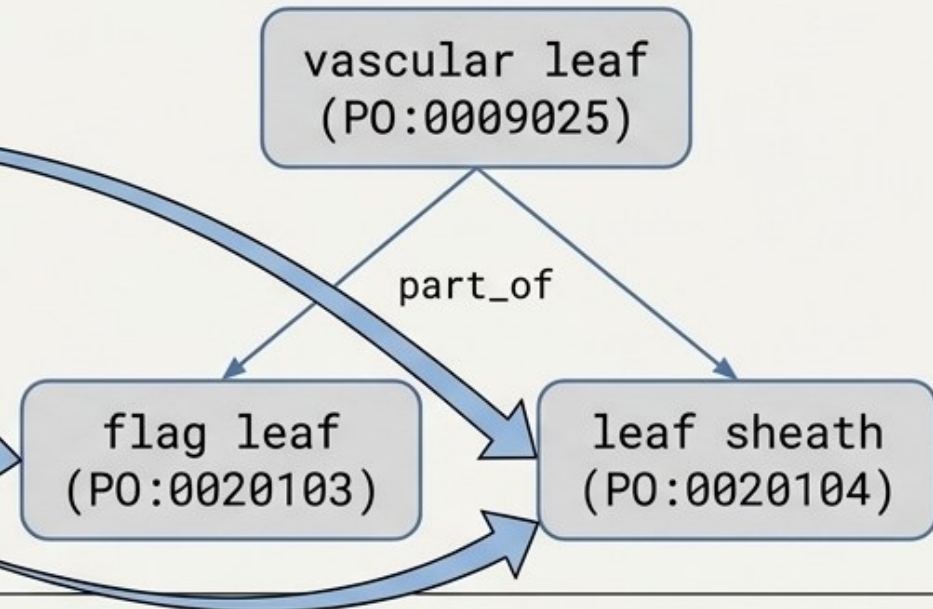


Praktik Semantik: Standardisasi Terminologi Ilmiah

Masalah: Terminologi Deskriptif Bebas

...penelitian menunjukkan bahwa pada tanaman padi (*Oryza sativa*), **leaf sheath** mengelilingi batang dan **flag leaf** adalah daun terakhir yang muncul sebelum malai. Struktur ini sangat penting untuk perlindungan dan fotosintesis...

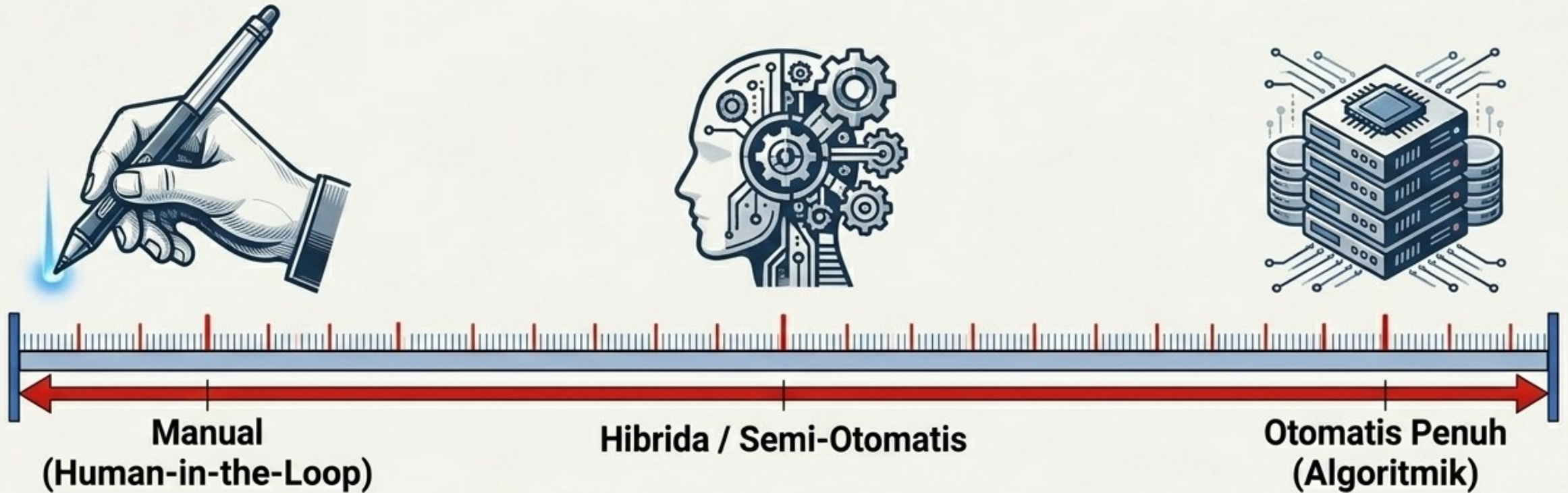
Solusi: Pelabelan Ontologi Standar (Plant Ontology/PO)



Pelabelan Ontologi menggantikan **terminologi deskriptif bebas** dengan kode referensi standar internasional. Ini menghilangkan bias subjektif peneliti dan memungkinkan deduksi struktural **otomatis** (contoh: *flag leaf* secara fungsional dikenali mesin sebagai bagian dari *vascular leaf*). Pelabelan Ontologi di secara fungsional dikenali mesin sebagai bagian dari *vascular leaf*).



Spektrum Eksekusi: Dari Intuisi Pakar ke Logika Skala Besar



Setelah tujuan pelabelan ditetapkan (Objek, Relasi, Semantik), tantangan selanjutnya adalah bagaimana label disematkan. Pemilihan metode selalu menjadi **kompromi** antara **tingkat presisi pakar manusia** dan **kebutuhan kecepatan komputasi mesin** pada skala raksasa.

Matriks Eksekusi: Manual vs. Otomatis

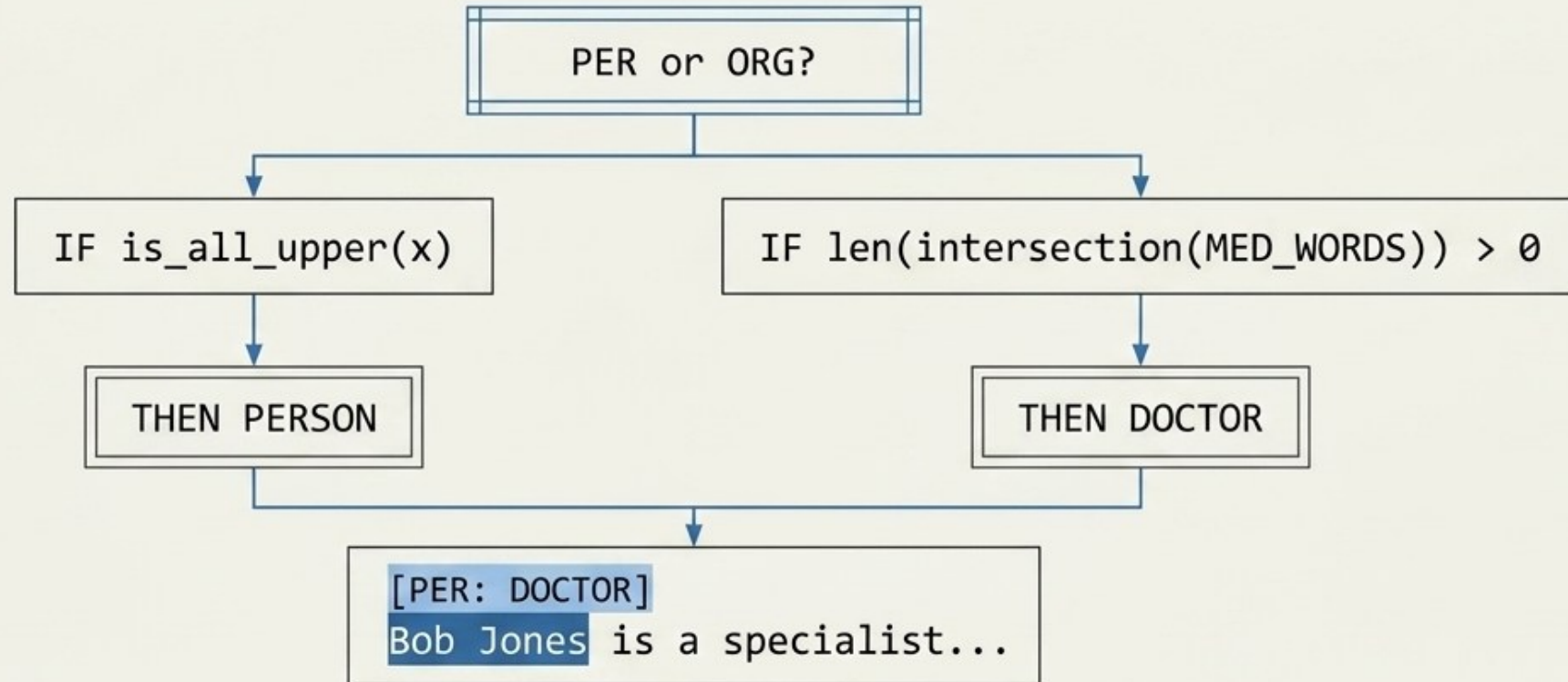
Pendekatan Manual	Pendekatan Otomatis
Sangat lambat (Item-per-item)	Sangat cepat (Skala eksponensial)
Langsung menandai batas objek dan metadata	Tidak langsung (pakar bertugas menyusun heuristik)
Kualitas bergantung ketat pada konsistensi pakar individu	Kualitas bergantung pada arsitektur fungsi pelabelan
Dataset berukuran kecil/baru, korelasi highly-nuanced	Dataset masif, volume tinggi, domain berstruktur ketat

Anatomi Pelabelan Manual: Sinkronisasi Multisensori



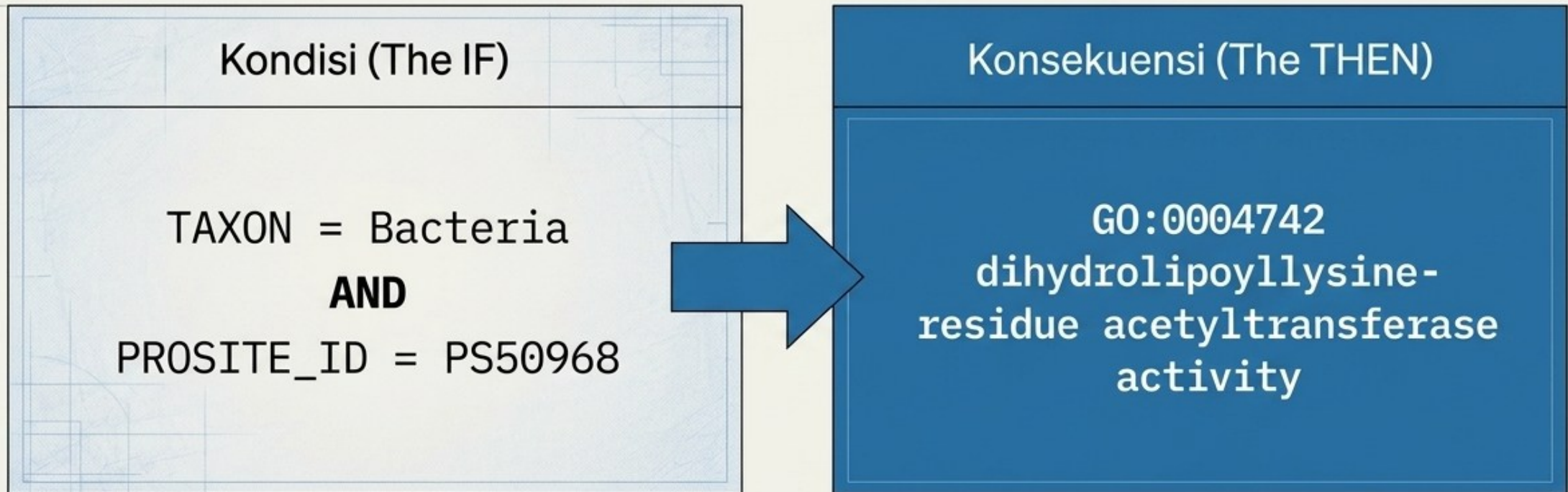
Ekspert membedah data melalui antarmuka grafis. Dari poligon batas spasial untuk gambar, kluster titik (*point-count*) intensitas koral, hingga segmentasi temporal berbasis rentang waktu untuk instrumen audio dan video.

Otomatisasi I: Heuristik & Weak Supervision



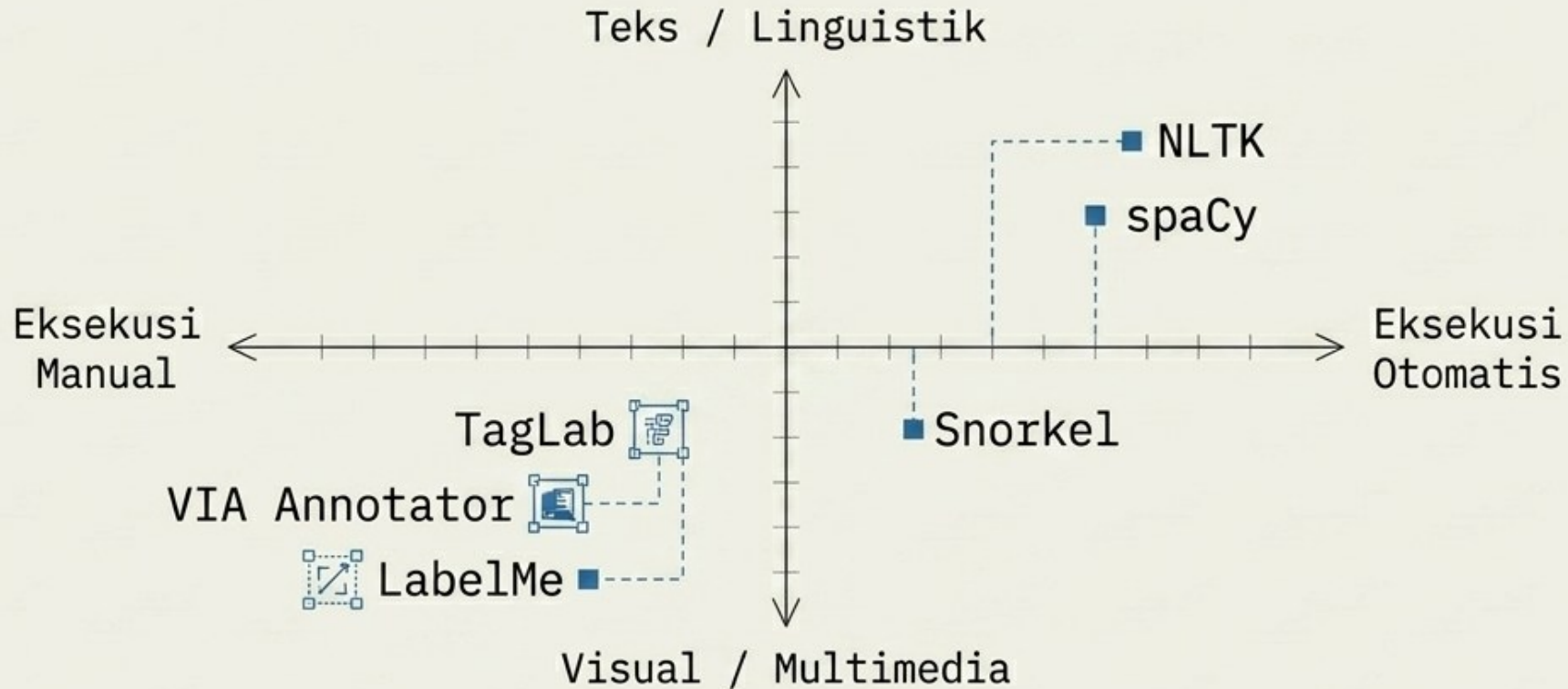
Daripada melabeli jutaan kata secara manual, pakar menulis Fungsi Pelabelan berbasis aturan pola, kamus, atau heuristik. Mesin mengintegrasikan aturan sederhana ini untuk menebak label akhir secara otomatis dengan probabilistik.

Otomatisasi II: Penelusuran Deduktif (*Rule-Based Systems*)



Rules Engine bekerja seperti corong logika absolut. Pakar mendefinisikan prasyarat metadata ketat. Jika data mentah secara absolut memenuhi semua kondisi (IF), mesin secara otomatis menerapkan konsekuensi akhir (THEN) berupa anotasi ontologi terstandar.

Peta Ekosistem Instrumen Pelabelan



Tidak ada satu alat penawar untuk semua jenis data. Pemilihan platform analitik sangat bergantung pada letak matriks ini: format sensorik data yang dihadapi dan kesiapan sumber daya untuk penulisan logika otomatis.

Evaluasi & Tindak Lanjut: Kerangka Implementasi Instansi



1. Identifikasi Inventaris

Audit repositori.
Teks, Gambar, Audio,
atau Video?



2. Pemetaan Instrumen

Cocokkan format
data mentah dengan
Peta Ekosistem alat.



3. Uji Coba Terbatas

Lakukan pelabelan
alpha-test pada 100
sampel item pertama.



4. Audit Waktu & ROI

Hitung waktu rata-
rata per item vs.
total man-hours untuk
memvalidasi efisiensi.

Rencana Aksi Anda: Gunakan **Cetak Biru** ini untuk menginisiasi standarisasi pelabelan data di instansi masing-masing. Data mentah repositori Anda tidak memiliki nilai intrinsik operasional hingga ia dilabeli secara terstruktur.

The diagram illustrates two AI services used for data labeling. On the left, a vertical stack of three colored squares (orange, blue, and green) represents different data types or stages. The orange square is labeled 'VIA / LabelMe' and 'Annotasi Gambar dan Video'. The blue square is labeled 'Snorkel' and 'Pelabelan Otomatis (Weak Supervision)'. The green square is labeled 'Model' and 'Model Training'. Arrows indicate a flow from the orange square to the blue square, and from the blue square to the green square. The background features a stylized circuit board pattern with lines and dots.

VIA / LabelMe
Annotasi Gambar dan Video

Snorkel
Pelabelan Otomatis (Weak Supervision)

METODE DAN IMPLEMENTASI

MANUAL VS OTOMATIS

MANUAL

Mengandalkan ekspert
Teknik seperti
Weak Supervision
atau Rule-Based

ALUR KERJA SUPERVISED LEARNING

Data yang telah dilabeli digunakan sebagai data latih bagi mesin untuk mempelajari karakteristik data.

DATA DILABELLI

MESIN PEMBELAJARAN

TRAINED MODEL

ESENSI DAN RUANG LINGKUP PELABELAN

MEMBERIKAN KONTEKS PADA DATA MENTAH

Menambahkan informasi latar belakang agar data dipahami secara seragam oleh manusia dan mesin.

PERANGKAT LUNAK PENDUKUNG (CONTOH)

PEMANFAATAN ONTOLOGI UNTUK PRESISI

Menggunakan referensi standar (Knowledge Base) untuk mengtiilangkan ambiguitas makna pada entitas data.

**KNOWLEDGE BASE
(REFERENSI STANDAR)**

TARGET DATA

VIA / LabelMe

Annotasi Gambar dan Video

Snorkel

Pelabelan Otomatis
(Weak Supervision)

NLTK / spaCy

TIGA FOKUS UTAMA IDENTIFIKASI

**OBJEK
(VISUAL)**

RELASI/STRU
(BANASA)

SEMANTIK (MAKNA MELALUI ONTOLOGI)

Buah

Merah

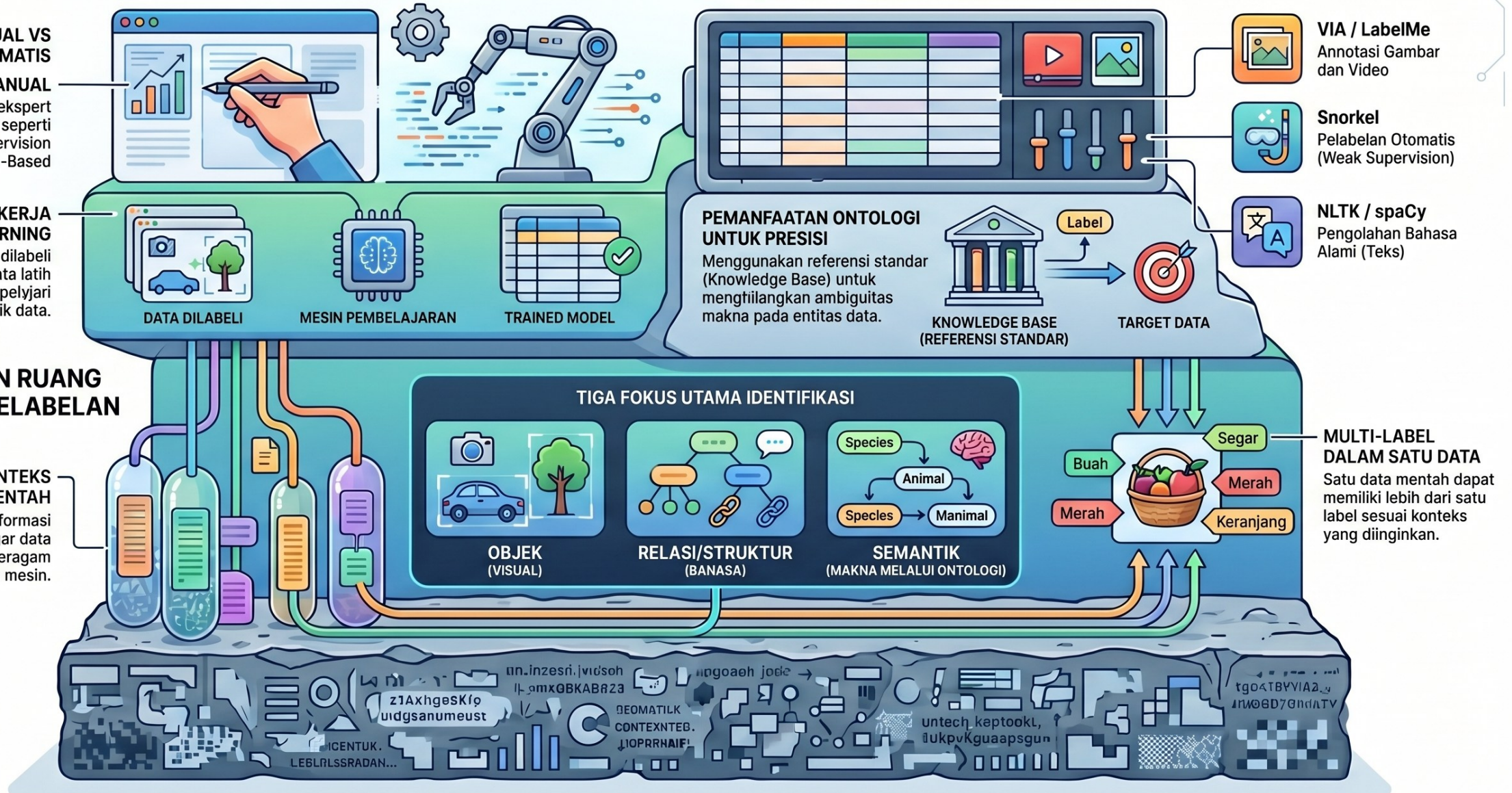
Segal

Merah

Kelanjutan

- MULTI-LABEL
DALAM SATU DATA

Satu data mentah dapat memiliki lebih dari satu label sesuai konteks yang diinginkan.



Apa itu Prapemrosesan Data?



BRIN
BADAN RISET
DAN INOVASI NASIONAL

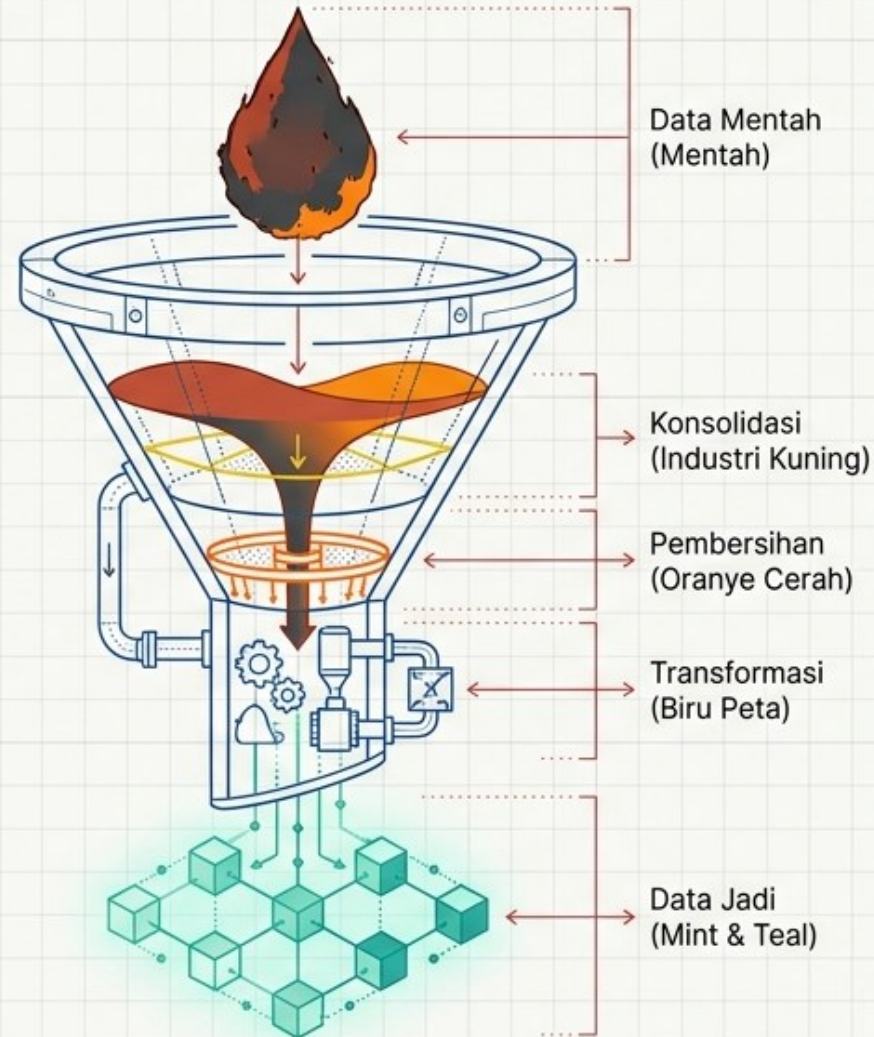
BerAKHLAK
Berorientasi Pelayanan Akuntabel Kompeten
Harmonis Loyal Adaptif Kolaboratif

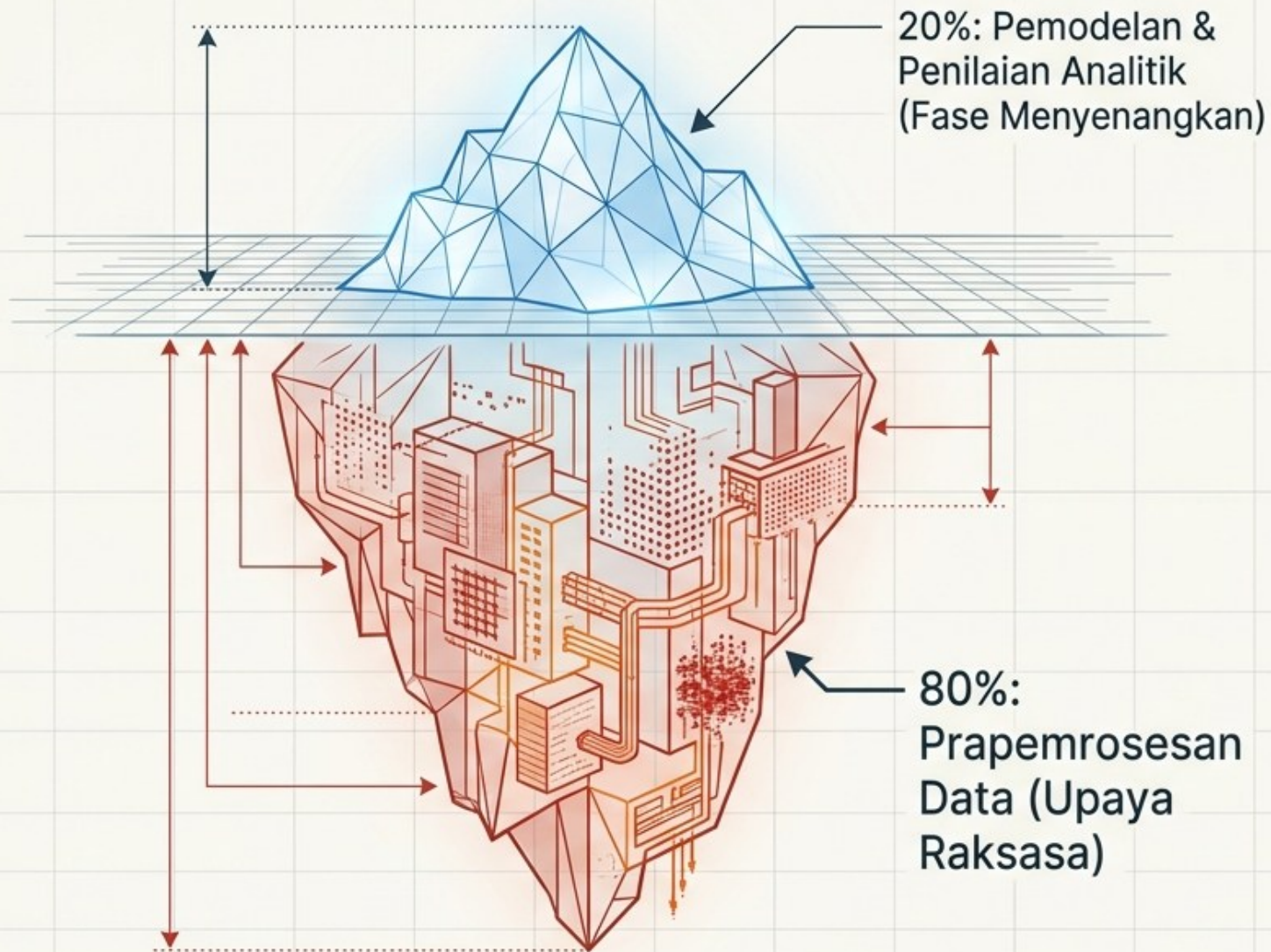
#bangga
melayani
bangsa

Cetak Biru Kilang Data

Merekayasa Data Mentah Menjadi Aset Analitik Bernilai Tinggi

Panduan Sistematis Prapemrosesan Data





Realita Penambangan Data: 80% Berada di Bawah Permukaan

Profesional analitik menghabiskan sebagian besar waktu dan tenaga mereka bukan untuk melatih algoritma, melainkan untuk mengubah data dunia nyata yang kotor menjadi format yang dapat dicerna oleh mesin.

Mengapa Data Mentah Tidak Pernah Siap Pakai

Data dunia nyata umumnya kotor, terlalu rumit, dan mengandung anomali yang mendistorsi keakuratan model prediktif.

Anatomi Data Buruk

Ketidaklengkapan (Missing Value): Data tidak tercatat, sengaja dihapus, atau alat malfungsi.

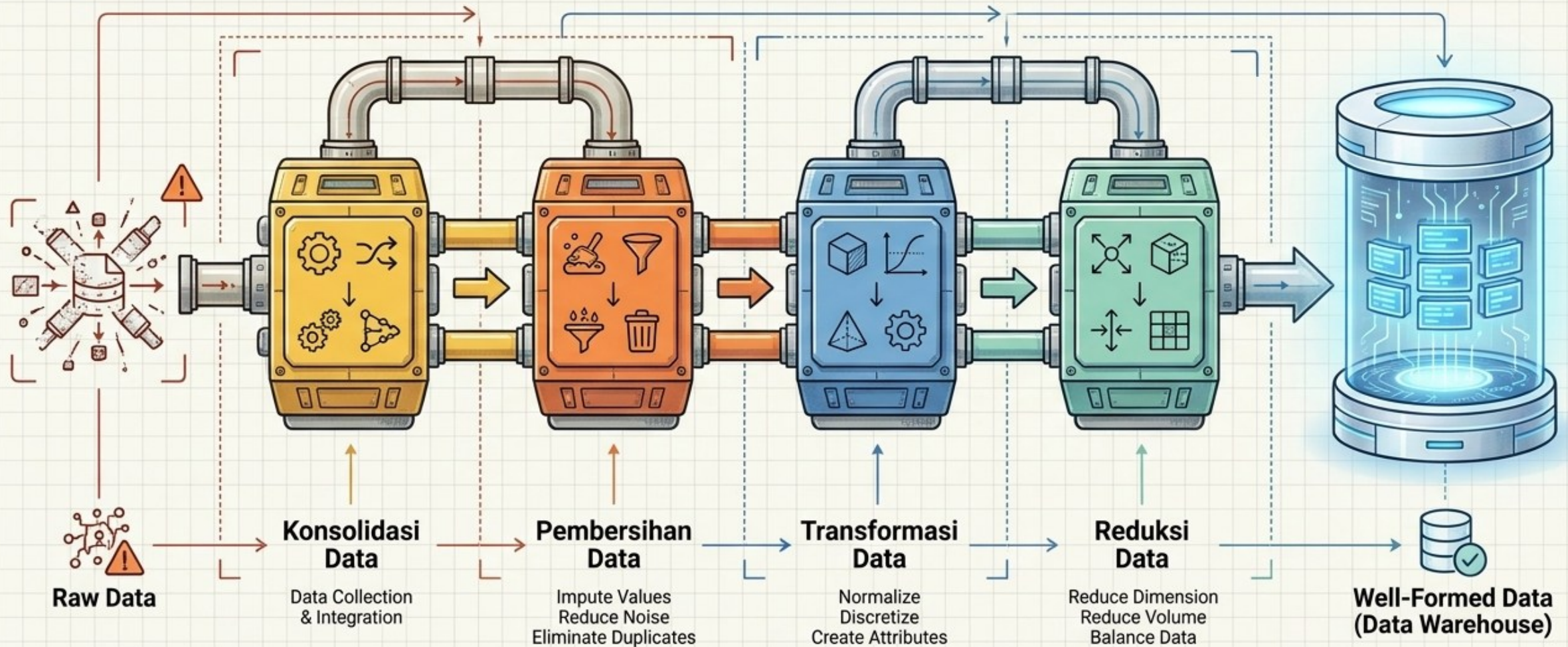
1

Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03		5.95%
04	700	6.40%
05	800	5.40%
06	900	5.70%
07	750	5.90%
08	N/A - ERROR	7.00%
09	650	6.50%
10	825	5.70%

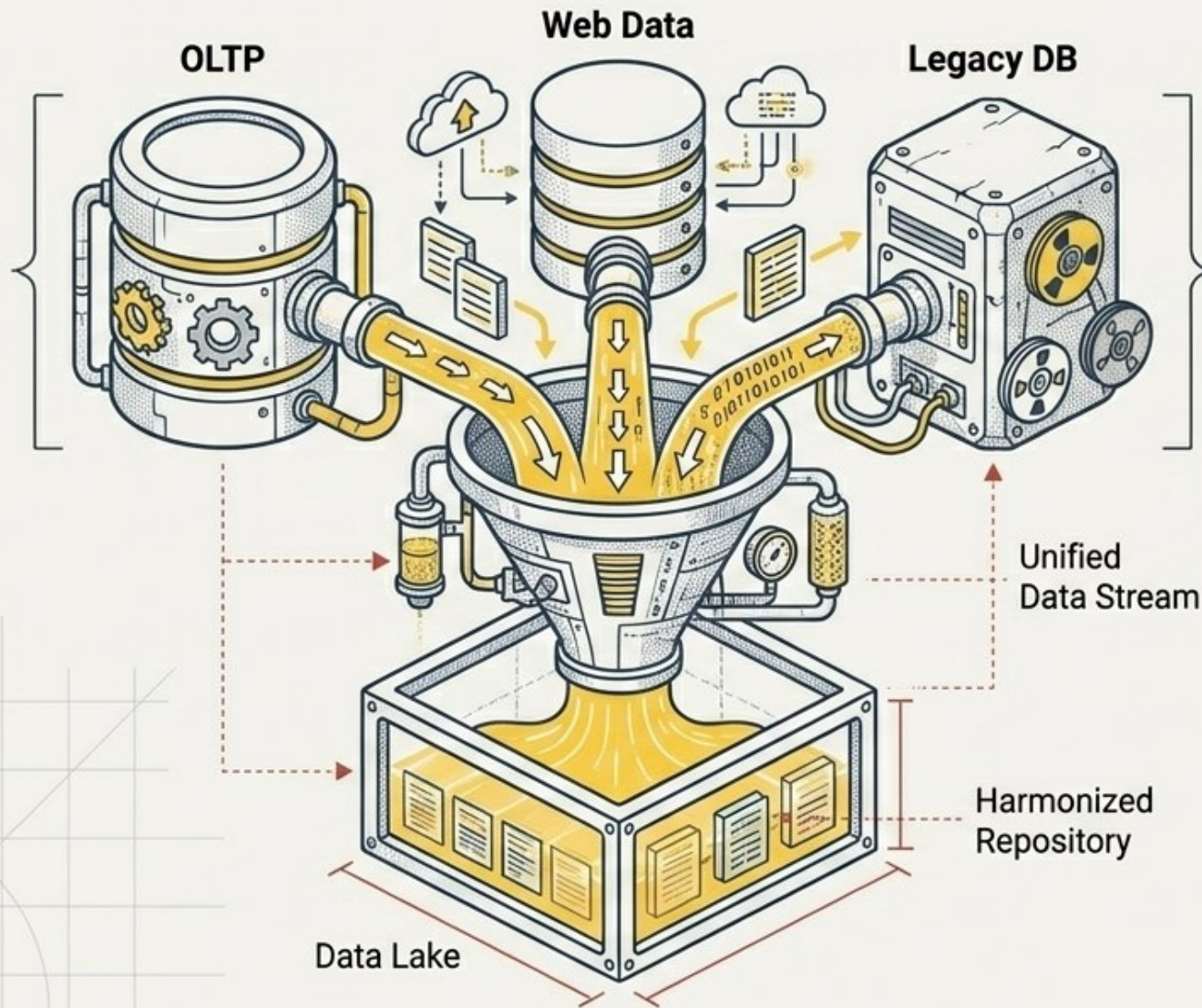
2 **Gangguan (Noise/Outlier):** Nilai anomali di luar batas teoretis atau kesalahan entri.

3 **Ketidakkonsistenan:** Perubahan sistem pengkodean antar periode operasional.

Arsitektur Prapemrosesan: 4 Fase Kilang Data



Fase 1: Konsolidasi Data (Pengumpulan & Penyatuan)



1. Akses & Kumpulkan

Menarik data menggunakan kueri SQL, layanan web, atau agen perangkat lunak.

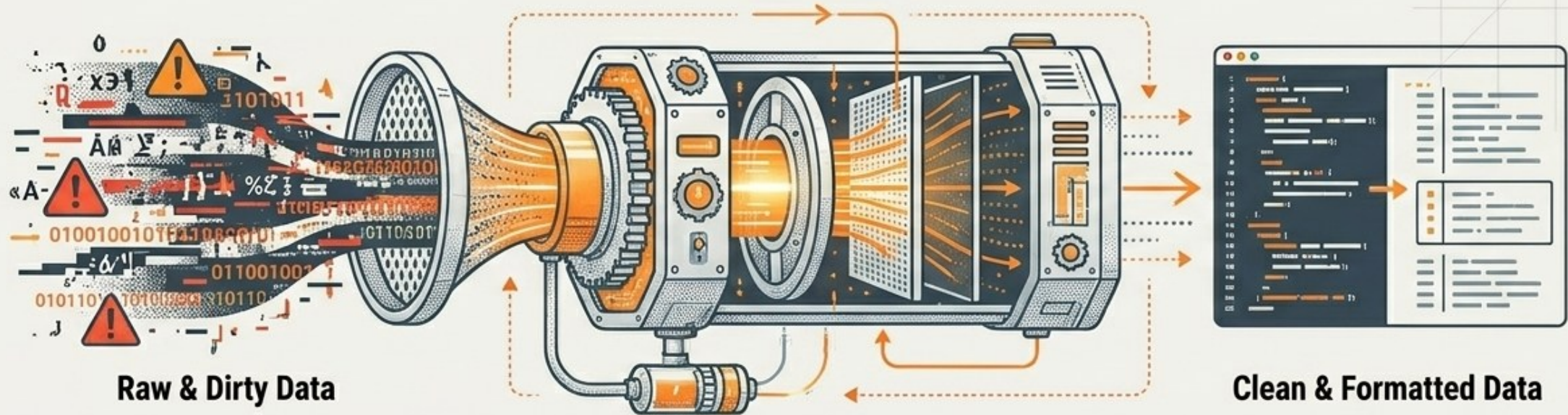
2. Pilih & Filter

Menyaring informasi yang tidak relevan berdasarkan keahlian domain pakar dan uji statistik.

3. Integrasikan

Menyatukan data dari berbagai sumber menggunakan pemetaan berbasis ontologi.

Fase 2: Pembersihan Data (Menyaring Kotoran)



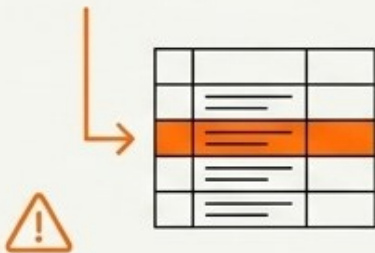
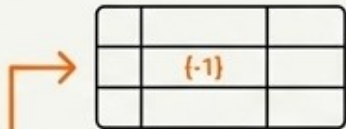
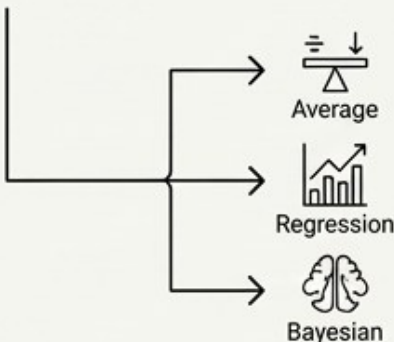
Tugas Utama (Key Tasks)

-  **Imputasi:** Memperhitungkan dan mengisi nilai yang hilang.
-  **Reduksi Noise:** Mengidentifikasi dan memuluskan anomali.
-  **Eliminasi:** Menghapus duplikat dan catatan yang salah.

Wawasan Analitik

Tidak semua nilai hilang adalah error. Contoh: Kolom pendapatan rumah tangga sering kali dibiarkan kosong secara sengaja oleh individu berpenghasilan teratas. Konteks domain sangat penting.

Matriks Mitigasi: Menangani Nilai yang Hilang (Missing Values)

1. Eliminasi	2. Inspeksi	3. Identifikasi	4. Pengganti (Imputasi)
Membuang seluruh baris.  Risiko: Kehilangan informasi substansial jika distribusi nilai hilang tidak beraturan.	Pakar domain mengevaluasi manual.  Karakteristik: Sangat akurat, namun subjektif dan lambat untuk Big Data.	Menggunakan nilai konvensional (mis. {-1}).  Fungsi: Menandai kekosongan tanpa menghapus rekaman.	Substitusi otomatis menggunakan Rata-rata, Regresi, atau Teorema Bayesian. 

Sensitivitas Algoritma

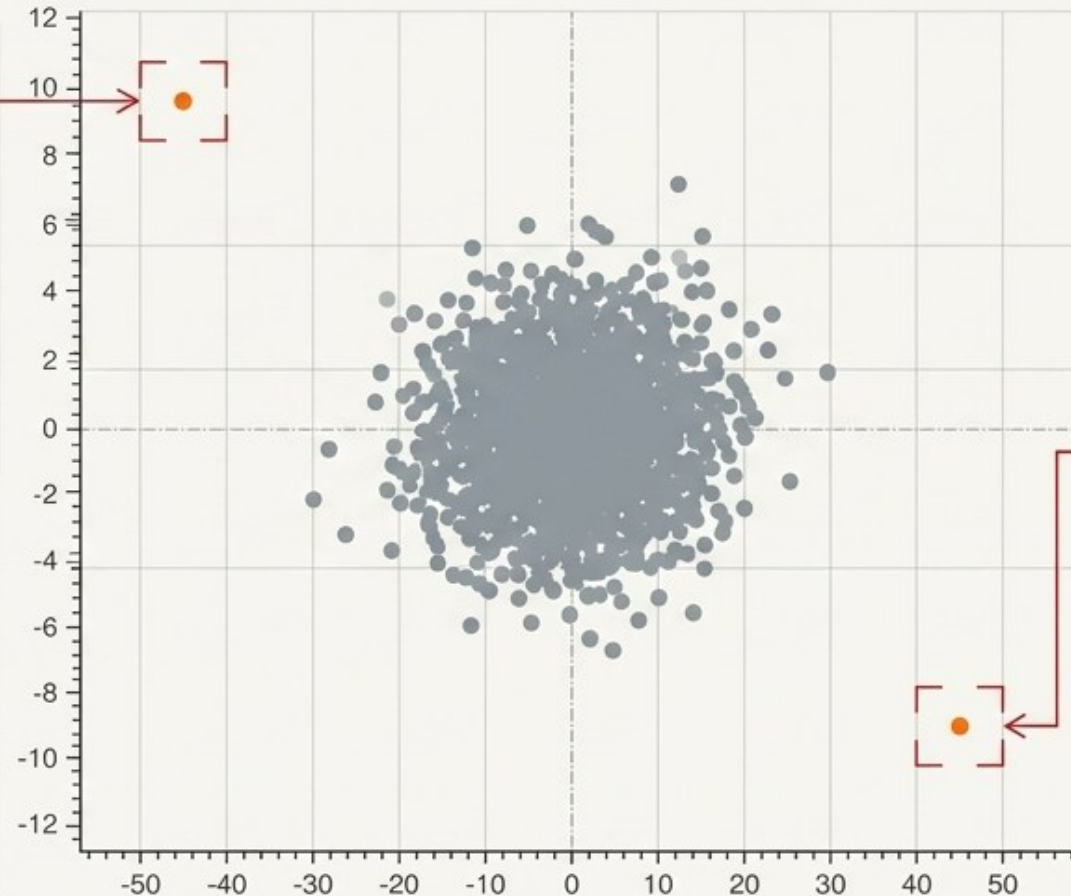
Algoritma **k-NN** sangat tangguh terhadap nilai yang hilang, sedangkan model **Jaringan Saraf Tiruan (Neural Networks)** akan gagal tanpa imputasi yang tepat.

k-NN

Deteksi Outlier: Valid vs. Anomali

Outlier adalah anomali ekstrem yang mencondongkan (skew) pola model representatif.

Sah (Valid): Kejadian nyata yang ekstrem (Contoh: Pendapatan miliaran rupiah).



Keliru (Error): Kesalahan input murni (Contoh: Tinggi manusia tercatat 1,73 sentimeter).

Action

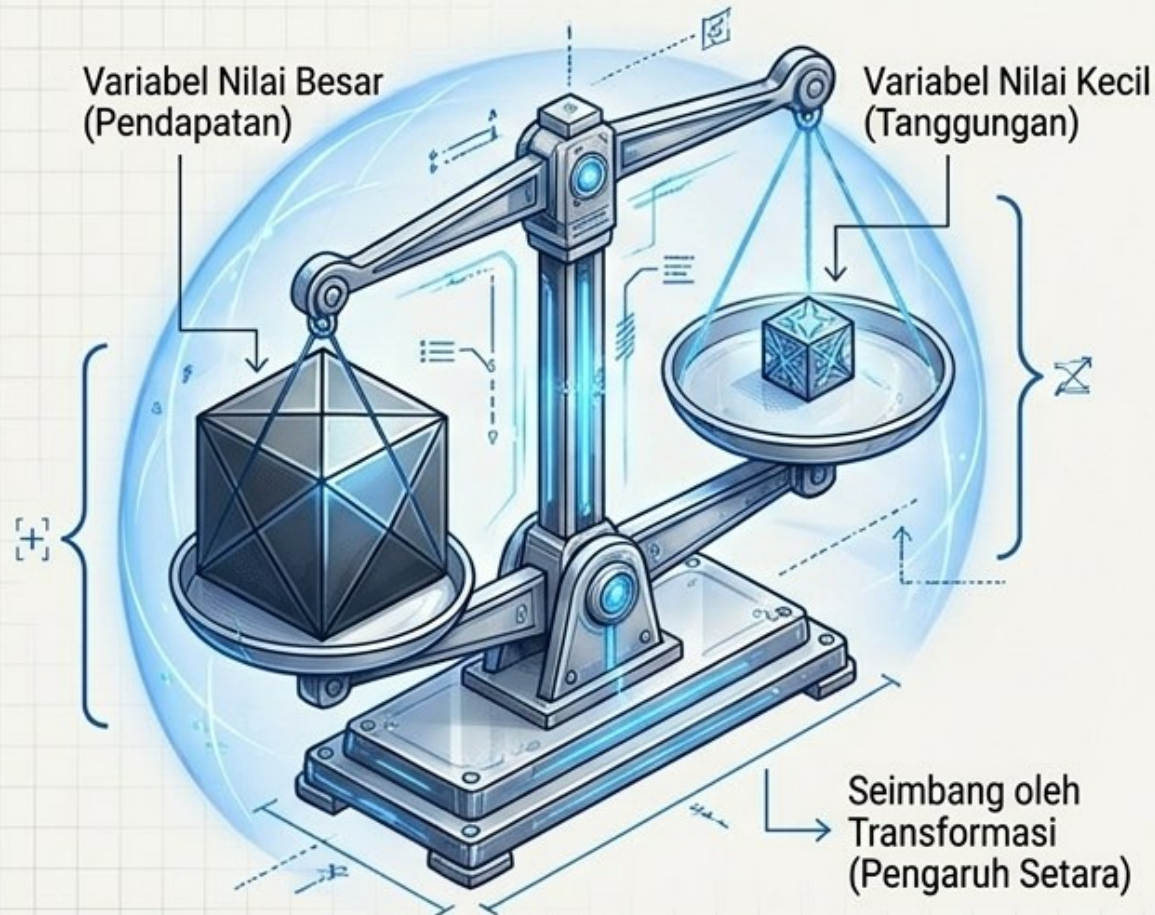


Identifikasi menggunakan rata-rata/standar deviasi, lalu hapus atau ratakan (binning).



Fase 3: Transformasi Data (Membentuk Ulang Sinyal)

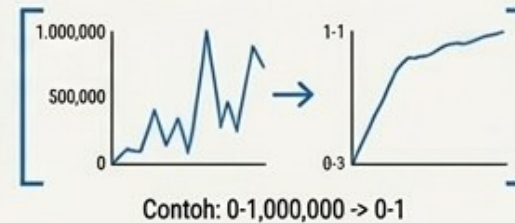
Untuk mencegah satu variabel numerik besar (mis. Pendapatan) mendominasi variabel bernilai kecil namun penting (mis. Jumlah tanggungan).



3 Strategi Transformasi

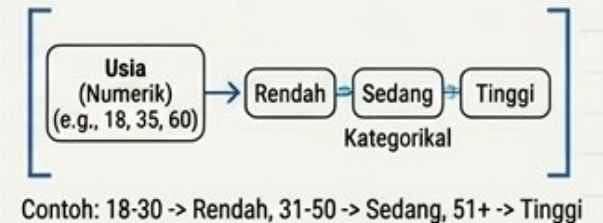
1. Normalisasi

Menstandarkan rentang nilai.



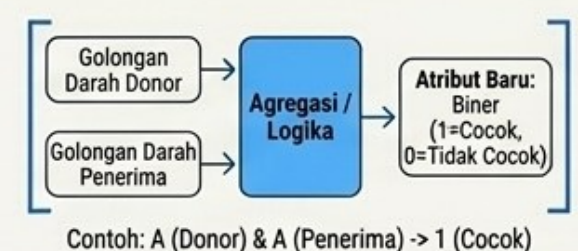
2. Diskritisasi

Mengubah numerik menjadi kategorikal (mis. Usia -> Rendah, Sedang, Tinggi).



3. Agregasi / Atribut Baru

Membuat variabel turunan (mis. Menggabungkan golongan darah donor dan penerima menjadi satu atribut biner: 1=Cocok, 0=Tidak Cocok).



Anatomi Normalisasi Data

Skala Desimal

$$x'_{ij} = \frac{x_{ij}}{10^h},$$



Efek: Menggeser titik desimal ke kiri.

- Output: Rentang nilai berada di antara $[-1, 1]$.

Min-Maks

$$x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}} (x'_{\max,j} - x'_{\min,j}) + x'_{\min,j}$$

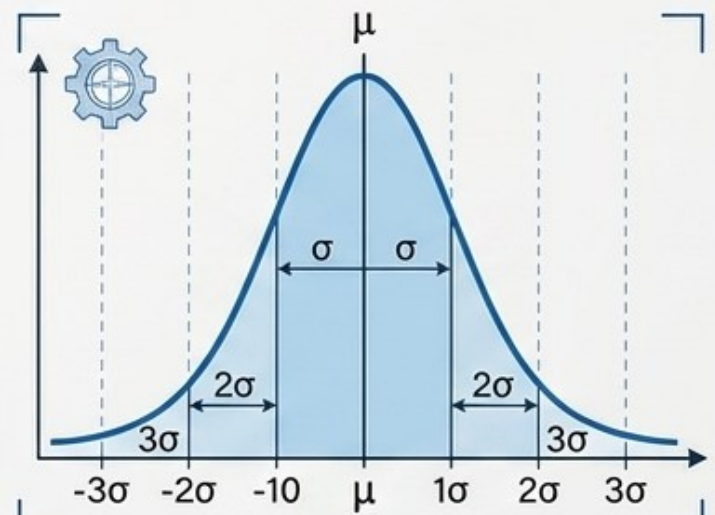


Efek: Menekan skala ke batas presisi yang diinginkan secara seragam.

- Output: Rentang nilai spesifik, umumnya $[0, 1]$.

Z-Index

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j},$$

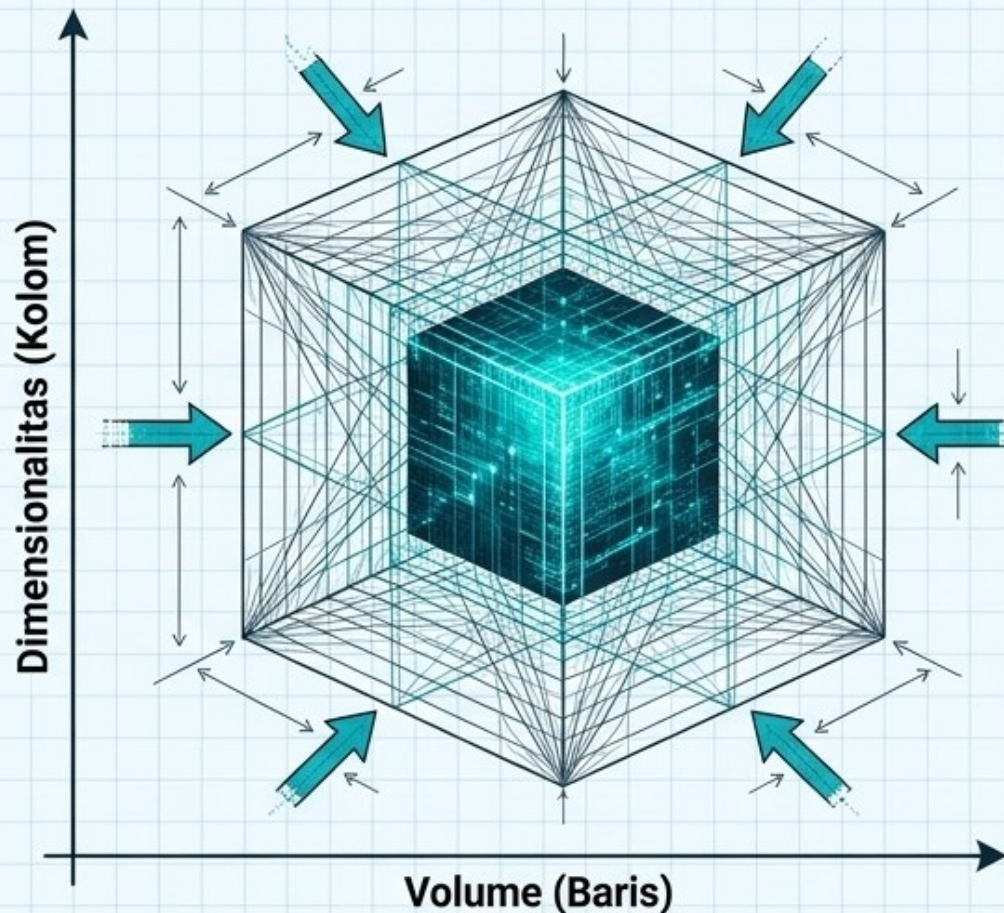


Efek: Menggunakan rata-rata (mean) dan deviasi standar.

- Output: Distribusi kurva normal, nilai hampir pasti berada di rentang $(-3, 3)$.

Fase 4: Reduksi Data (Memadatkan Nilai)

Ilmuwan data menyukai Big Data, tetapi terlalu banyak dimensi dan rekaman memperlambat analitik (misal: data microarray genom atau pemrosesan gambar).



3 Sumbu Solusi

1. Reduksi Dimensi (Kolom)

Memilih variabel paling relevan (Principal Component Analysis, Chi-square).



2. Reduksi Volume (Baris)

Mengurangi jumlah kasus menggunakan Random/Stratified Sampling.

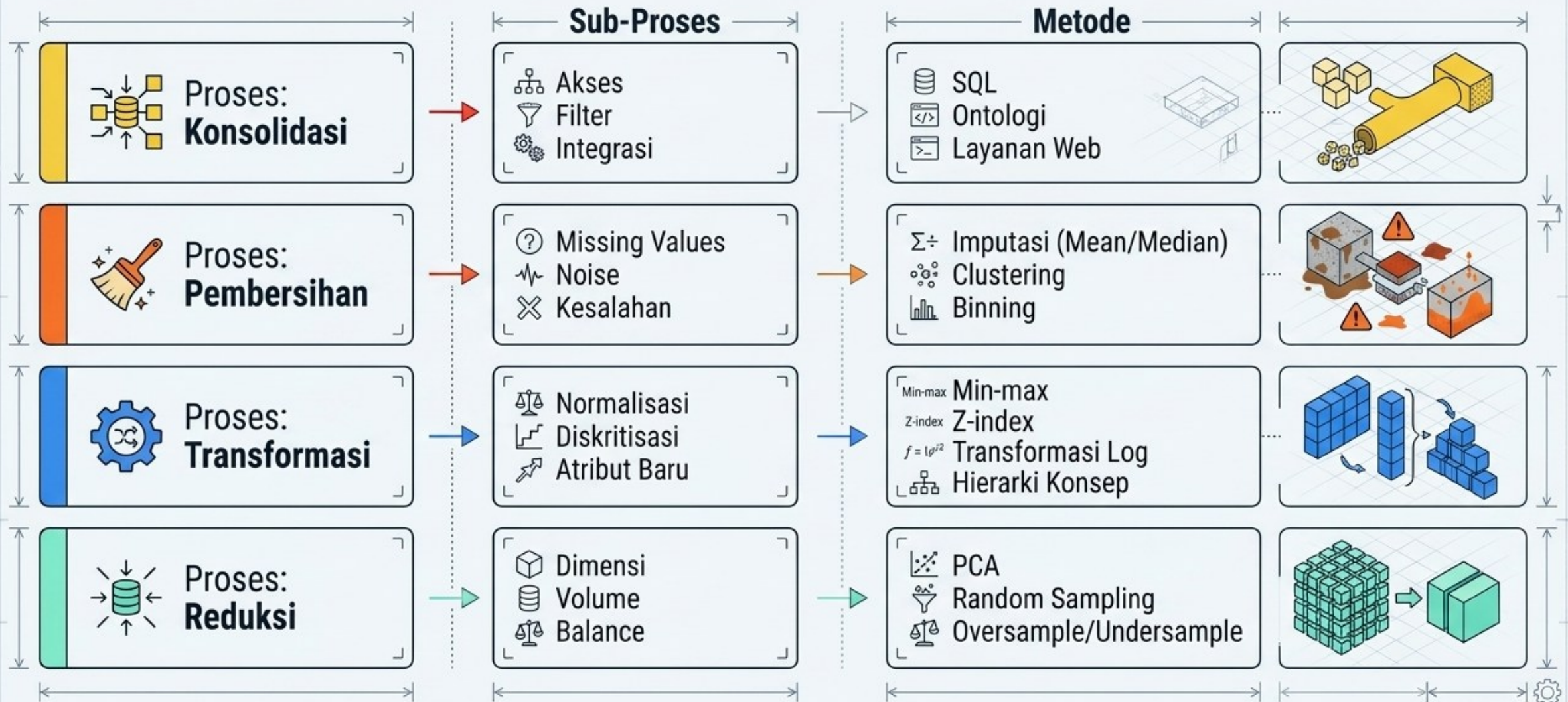


3. Balancing Data

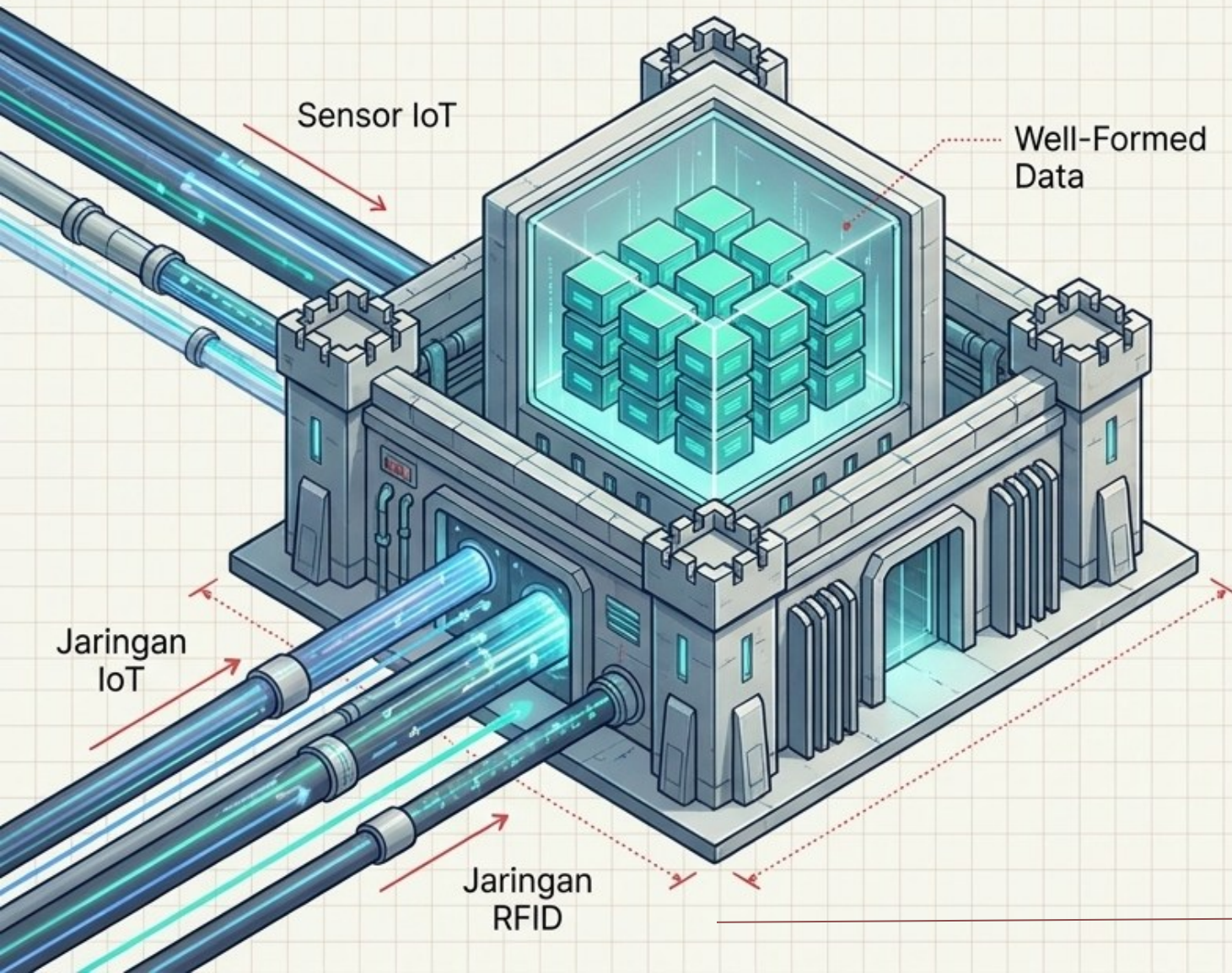
Melakukan Oversampling (kelas minoritas) atau Undersampling (kelas mayoritas) pada data yang miring (skewed).



Lembar Contekan (Cheat Sheet): Ekosistem Prapemrosesan Data



Data Warehouse: Repositori Hasil Pemurnian



Kualitas data adalah bahan bakar utama untuk Business Intelligence.

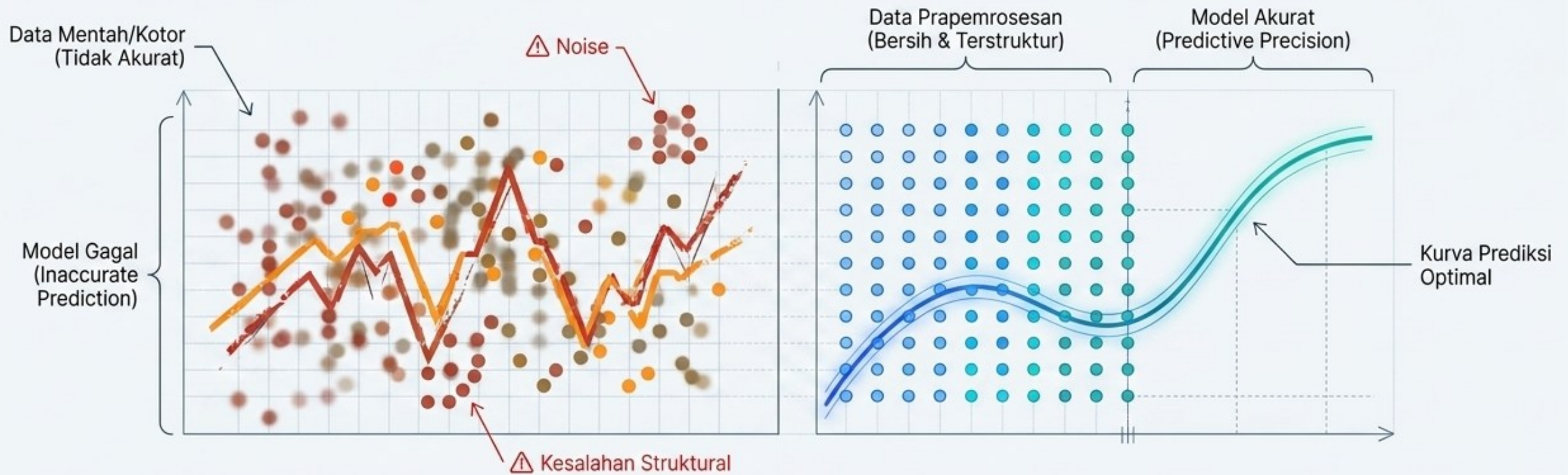
The Big Data Shift:

Otomatisasi pengumpulan data (Sensor, RFID, Internet) meningkatkan volume data secara radikal, membuat prapemrosesan yang kuat menjadi kewajiban mutlak, bukan sekadar opsi.

Data bersih di Data Warehouse menjamin akurasi wawasan perusahaan.



Fondasi Algoritma Prediktif



Takeaway: Model machine learning dan analitik prediktif terhebat di dunia tidak dapat mengkompensasi kualitas data yang buruk.

Closing thought: Kualitas wawasan yang dihasilkan oleh teknologi keputusan sepenuhnya ditentukan oleh upaya rekayasa yang Anda investasikan dalam fase Prapemrosesan Data. Kehebatan algoritma dimulai dari data yang bersih.

Anatomi Prapemrosesan Data: Memurnikan Data Mentah untuk Analitik

Reduksi Dimensi & Data

Mengurangi jumlah atribut ke ukuran yang dapat dikelola tanpa kehilangan informasi substansiai.

Transformasi & Normalisasi

Mengubah skala data (misalnya 0 ke 1) untuk mengurangi potensi bias variabel.

Arsitektur Pemurnian Data

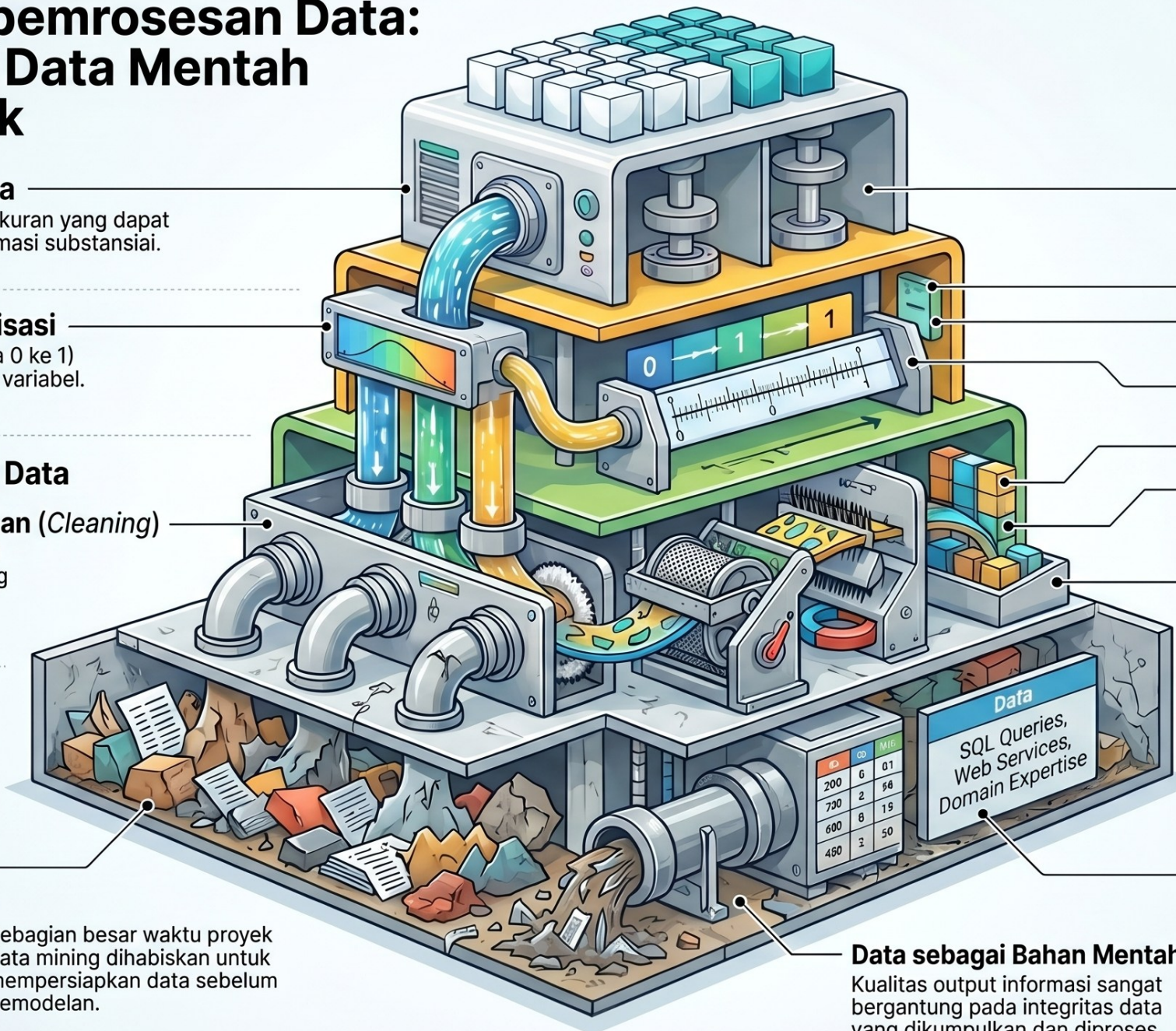
Konsolidasi & Pembersihan (Cleaning)

Mengintegrasikan sumber data dan menangani nilai yang hilang atau outlier.

Fondasi & Urgensi Prapemrosesan



Sebagian besar waktu proyek data mining dihabiskan untuk mempersiapkan data sebelum pemodelan.



Min-Max Scaling
Adjustor hydraunticaliv presses & laser cutters.

Min-Max Scaling
Z-Score Index
Calibrator precision range.

Normalisasi Desimal

Imputasi (Mean/Median)

Clustering
Separator grouping similar elements and outliers.

Binning
Sorter organized elements and buckets.

Masalah Utama Data Mentah

-  Data asli biasanya tidak lengkap
-  mengandung noise (outlier)
-  dan tidak konsisten.

Data sebagai Bahan Mentah
Kualitas output informasi sangat bergantung pada integritas data yang dikumpulkan dan diproses.

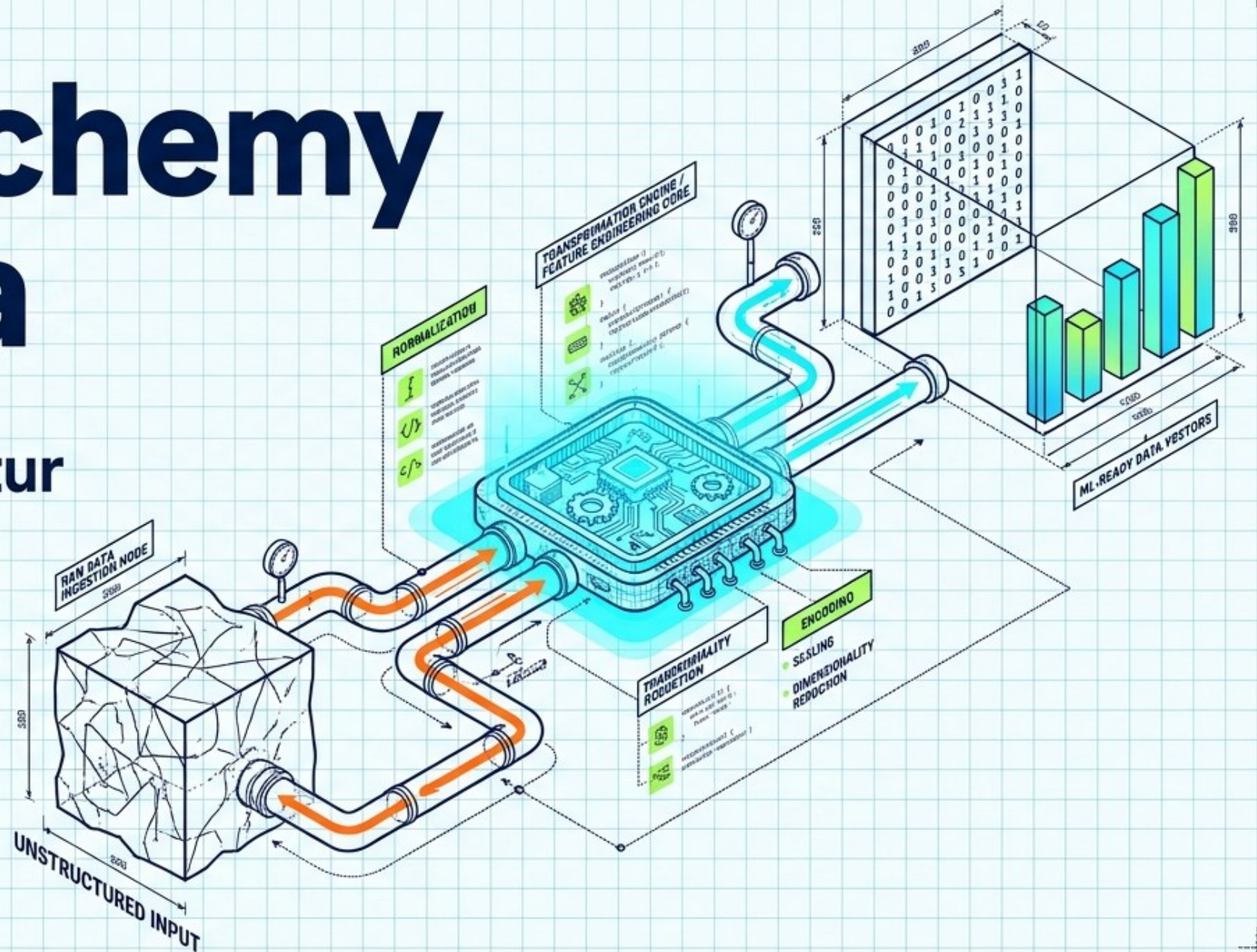


Apa itu Data Outlier? Bagaimana Deteksi Outlier Perekrutannya Fitur?

The Alchemy of Data

Panduan Visual Perekayasaan Fitur

Mengubah data mentah menjadi bahan bakar teroptimasi untuk Machine Learning.



Prapemrosesan Membersihkan; Perekayasaan Mengekstrak Makna

Perekayasaan fitur adalah transformasi komprehensif. Bukan sekadar memperbaiki format, melainkan menciptakan representasi yang langsung dapat dikonsumsi oleh mesin.

Prapemrosesan (Cleaning)

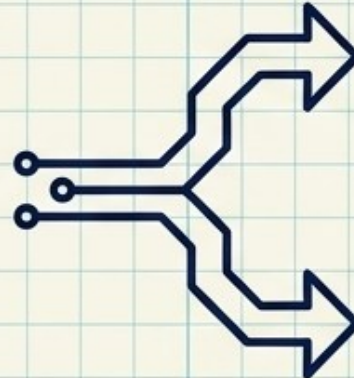
Dec 16 9am: server shutting
down & restarting



Dec 16 9am: server shutting
down ✓ restarting

Perekayasaan Fitur (Organizing)

Dec 16 9am: server shutting
down & restarting



Date	Text
Dec 16 9am	Server shutting down & restarting
Dec 16 10am	Server resuming all operations

Anatomi Data dan Batasan Operasionalnya

Identifikasi tingkat data menentukan transformasi matematis apa yang diizinkan.

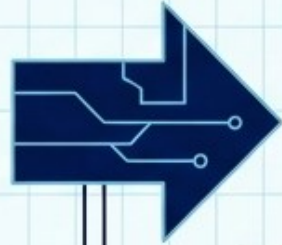
Diagnostic Matrix

Tingkatan Data	Penjumlahan / Pengurangan	Nilai Tengah (Median)	Perkalian / Pembagian	Nol Mutlak
Nominal (Kualitatif) Nama, Golongan Darah	✗	✗	✗	✗
Ordinal (Kualitatif) Ranking Kelas	✗	✓	✗	✗
Interval (Kuantitatif) Suhu (Celcius)	✓	✓	✗	✗
Rasio (Kuantitatif) Nilai Uang	✓	✓	✓	✓

Mengubah Kategori Menjadi Vektor Numerik Biner

One-Hot Encoding

Color
Red
Green
Blue

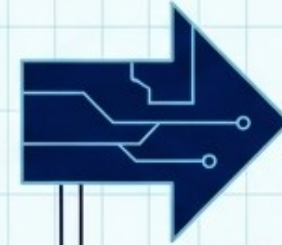


d1	d2	d3
1	0	0
0	1	0
0	0	1

Menghasilkan N fitur.
Boros memori untuk kategori berjumlah besar.

Dummy Encoding

Color
Red
Green
Blue



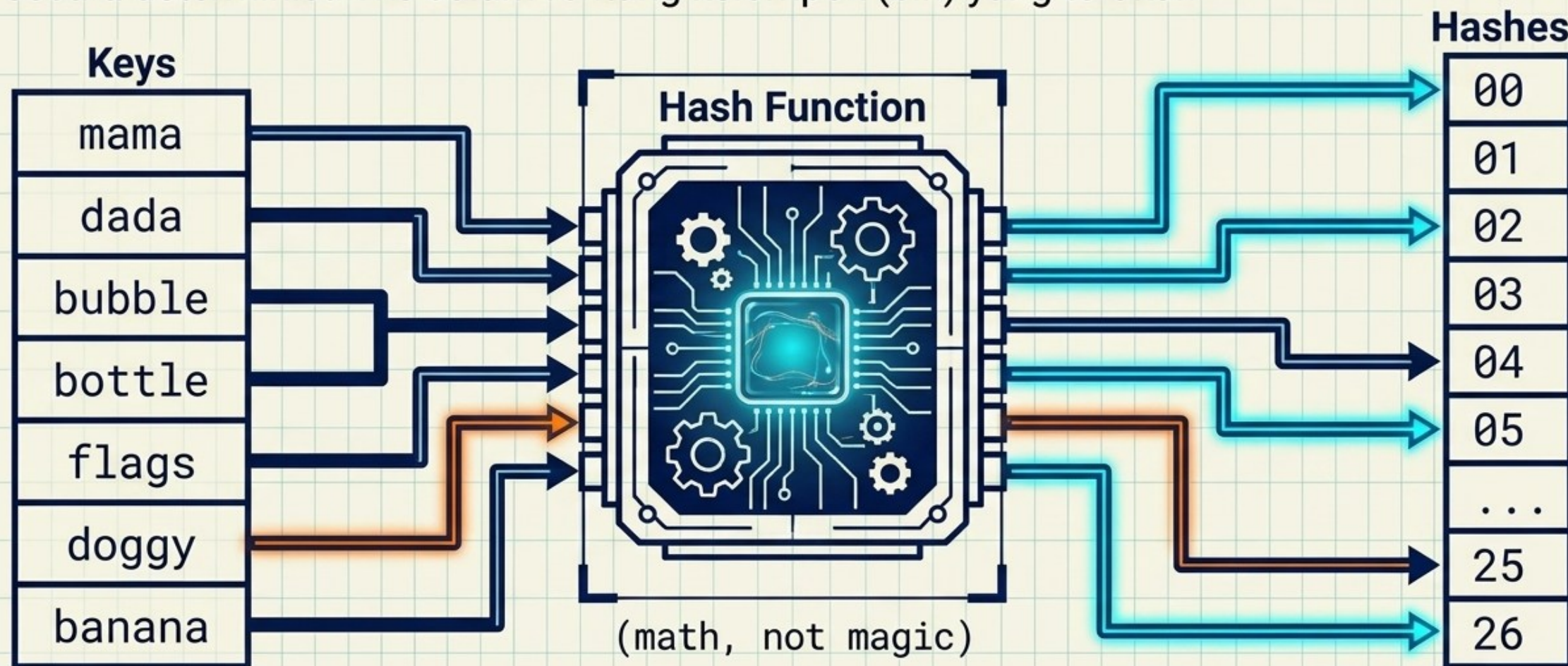
d1	d2
1	0
0	1
0	0

Menghasilkan N-1 fitur.
Kategori terakhir direpresentasikan oleh ketiadaan (0,0).

Variasi: Effect Encoding menggunakan nilai 1, 0, dan -1.

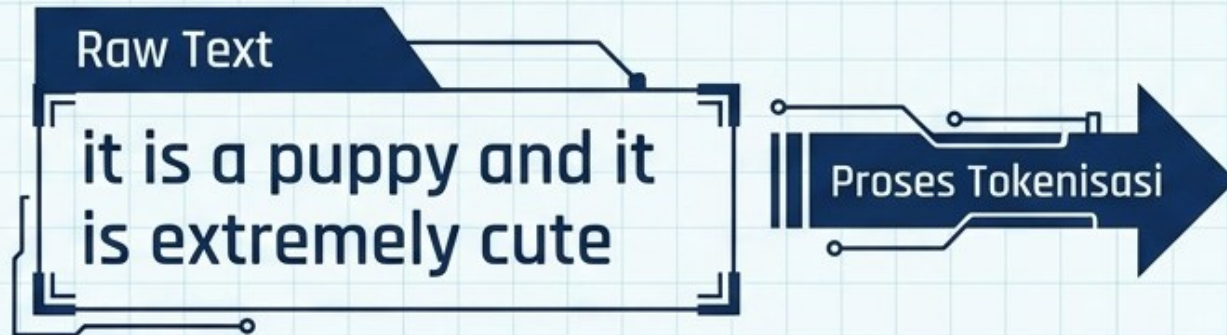
Hash Encoding Mengatasi Skala Kategori yang Masif

One-Hot gagal pada kategori unik yang berjumlah jutaan. Hash Encoding memetakan input secara deterministik ke dalam rentang kelompok (bin) yang terukur.



Pipa Natural Language Processing: Bag-of-Words

Teks mentah dipecah menjadi dimensi vektor berdasarkan frekuensi kemunculan. Urutan tata bahasa diabaikan secara total.

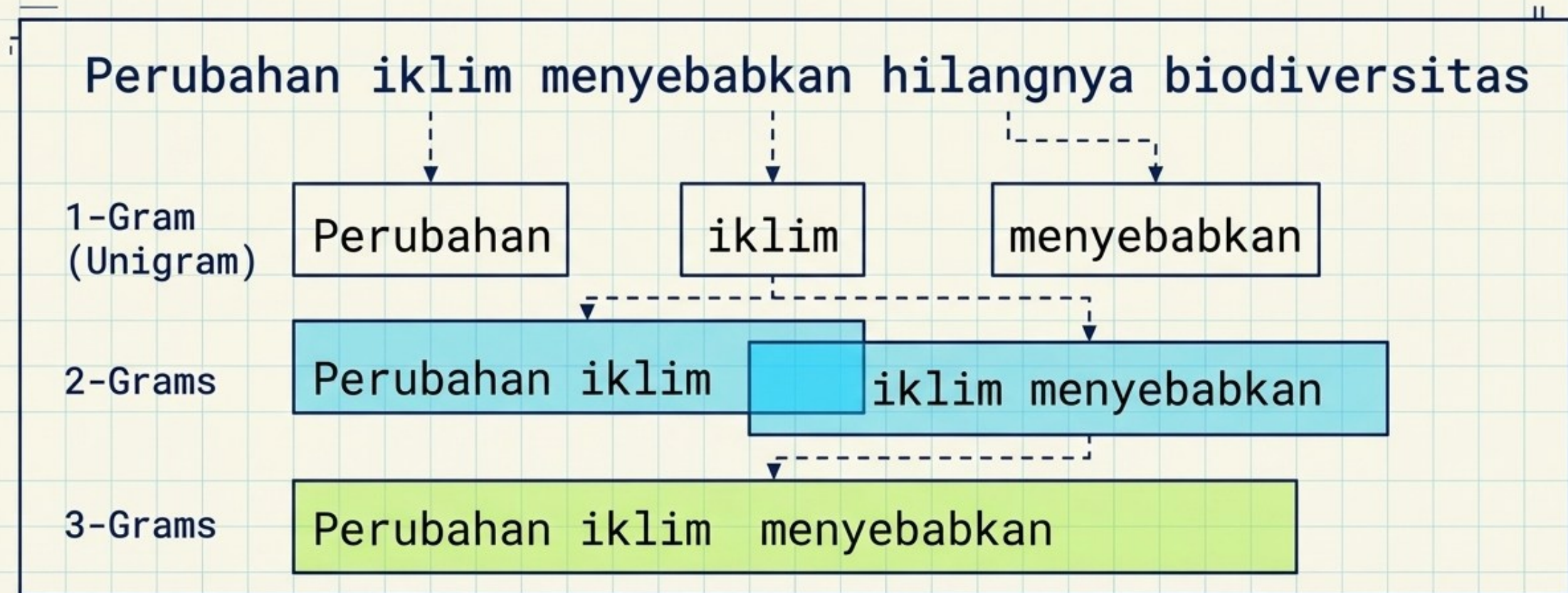


Bag-of-words vector	
it	2
they	0
puppy	1
and	1
cat	0
aardvark	0
cute	1
extremely	1

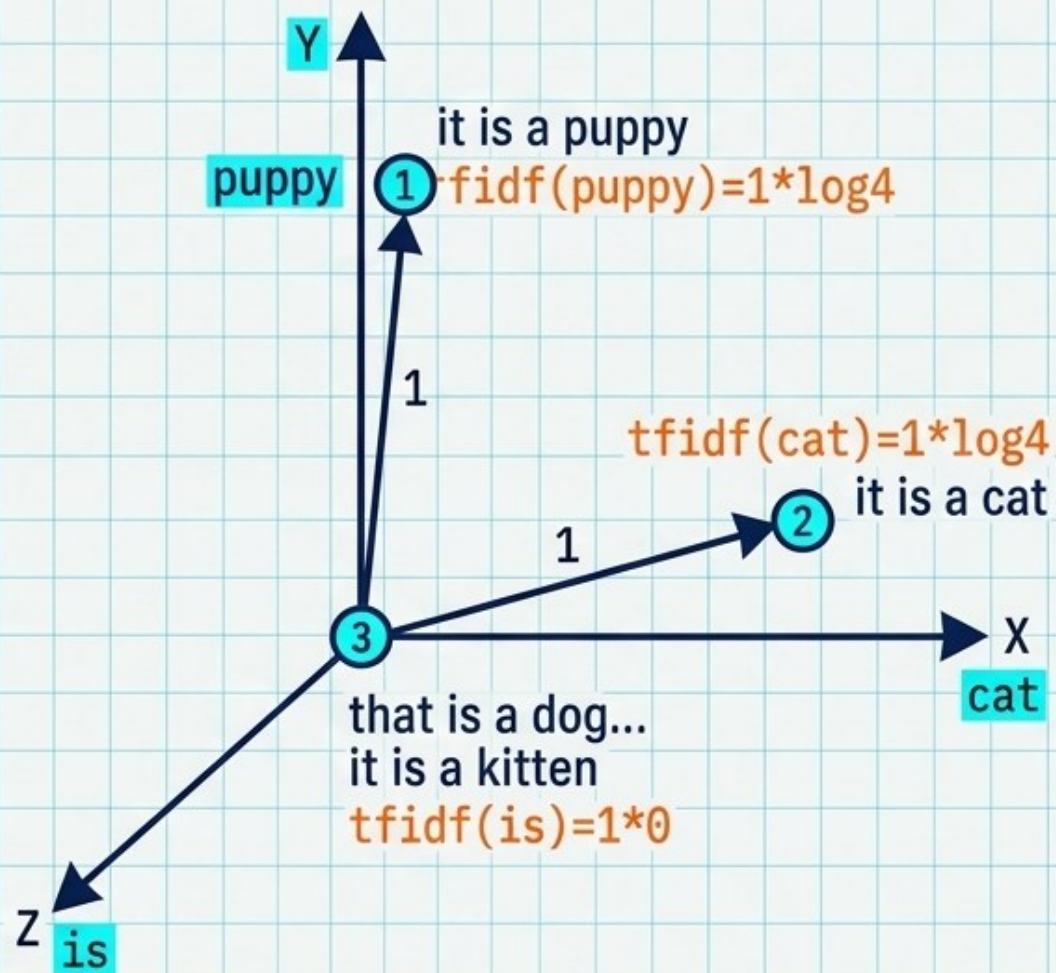
Dimensi array sama dengan total kosakata referensi (N). Nilai 0 menunjukkan ketiadaan kata.

N-Grams Menjaga Konteks Urutan Makna

Mengekstraksi fitur dengan mempertahankan urutan lokal. Fitur lebih kaya, namun menuntut sumber daya komputasi eksponensial.



TF-IDF: Menimbang Nilai Kata dalam Ruang Vektor



$$tf-idf = bow * \log\left(\frac{N}{\text{dokumen containing } w}\right)$$

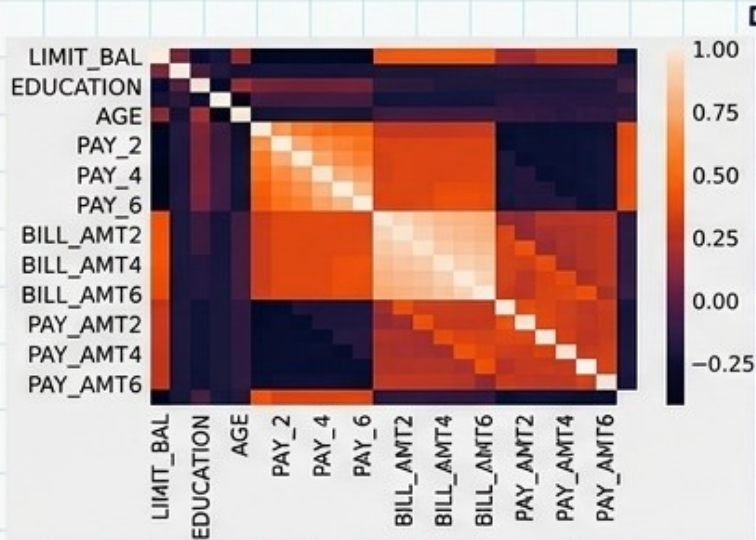
- TF (Term Frequency): Seberapa sering kata muncul di 1 dokumen.
- IDF (Inverse Document Frequency): Seberapa unik kata tersebut di seluruh dataset.



Kata pengganggu umum mendapat bobot 0.0, sementara kata unik mendapat bobot tinggi.

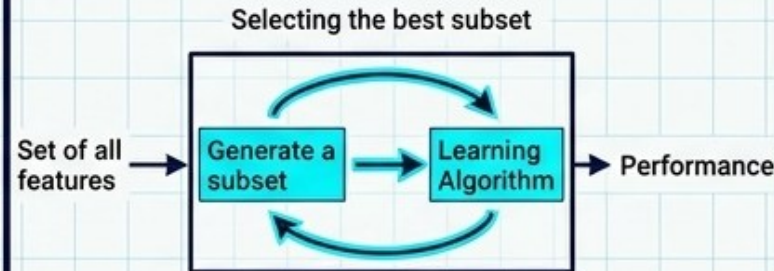
Tiga Pilar Seleksi Fitur untuk Eliminasi Noise

Filtering (Evaluasi Eksternal)



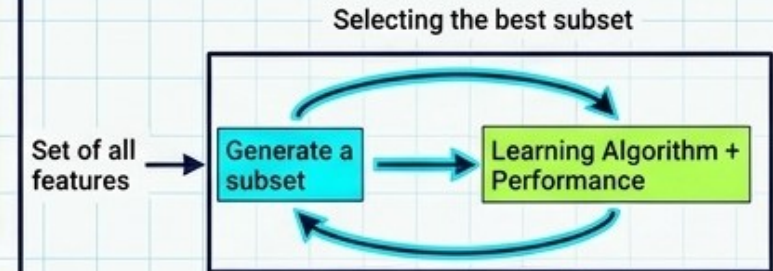
Menghapus fitur individual berkorelasi rendah sebelum masuk model.

Wrapper (Evaluasi Iteratif)



Mencari kombinasi iteratif. Melakukan training berulang untuk menemukan subset performa kumulatif terbaik.

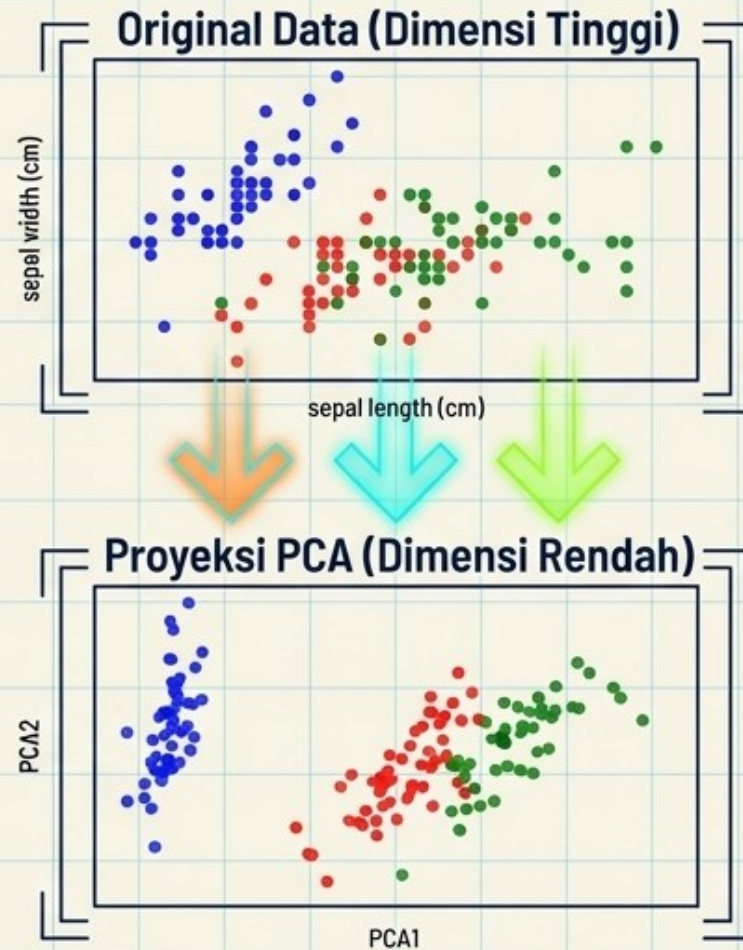
Embedded (Evaluasi Internal)



Seleksi fitur terjadi bersamaan dengan proses learning. Model secara mandiri menyesuaikan bobot signifikansi fitur.

Menghindari Kutukan Dimensionalitas (Transformasi Fitur)

Komputasi melambat drastis saat dimensi melebar. Transformasi menyusutkan ukuran sambil mempertahankan inti informasi.



PCA (Principal Component Analysis)

Fokus menangkap variasi data secara keseluruhan. Memproyeksikan data ke sumbu baru (Principal Components) yang memaksimalkan penyebaran varians.

LDA (Linear Discriminant Analysis)

Fokus menyorot fitur dengan kemampuan pembeda kelas yang paling tajam. Mengoptimalkan separasi antar grup klasifikasi.

The Feature Engineering Blueprint: Sintesis Keseluruhan

Perekayasa fitur dibentuk spesifik oleh arsitektur dan tujuan dari model akhir.

Area Perekayasa Fitur



Anatomi Perekayasaan Fitur

(Anatomy of Feature Engineering)

Memberikan gambaran komprehensif tentang bagaimana data mentah diproses dan ditransformasi menjadi fitur yang optimal untuk algoritma Machine Learning.

Perekayasaan fitur adalah proses komprehensif dalam memilih, mengekstrak, dan mentransformasi data mentah menjadi atribut yang lebih informatif. Tujuannya adalah untuk meningkatkan akurasi dan efisiensi algoritma pembelajaran mesin melalui pemahaman karakteristik data dan penerapan teknik seleksi yang tepat.

Teknik Pengembangan & Seleksi (Lapisan Inti)

Enkoding & Ekstraksi Teksual

One-Hot Encoding

Kategori: Merah, Hijau

Fitur: $\begin{bmatrix} 1, 0, 0, \\ 0, 1, 0 \end{bmatrix}$

N (Jumlah kategori unik), Biner (1 atau 0)

Hashing

longstring

1022365

Bag-of-Words, TF-IDF

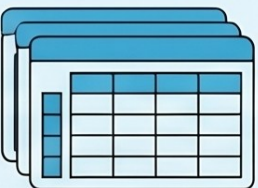
learn machine data

Fitur: $\begin{bmatrix} \text{learn: 1, machine: 1, data: 1, ...} \end{bmatrix}$

Tabol: Perbedaan Metode Enkoding Data Kategorikal

Metode Enkoding	Jumlah Fitur yang Dihasilkan	Representasi Nilai
One-Hot Encoding	N (Jumlah kategori unik)	Biner (1 atau 0)
Dummy Encoding	N - 1	Biner dengan 1 fitur dummy
Effect Encoding	N - 1	Menggunakan 1, 0, dan -1

FITUR OPTIMAL (OPTIMIZED FEATURES)



MODEL MACHINE LEARNING

Tiga Metode Seleksi Fitur

Filtering (Korelasi)

Menggunakan teknik korelasi

Wrapper (Subset)

Menggunakan subset

Embedded (Saat Pelatihan)

Menggunakan saat pelatihan

Transformasi Dimensi (PCA/LDA)

Memproyeksikan data dimensi tinggi ke ruang yang lebih rendah untuk efisiensi.

Fondasi & Pemahaman Fitur (Lapisan Dasar)



DATA MENTAH (RAW DATA)

Karakteristik & Level Data

Data Terstruktur

Tingkatan Data

Nominal

Ordinal

Data Tidak Terstruktur

Interval

Ratio

Signifikansi Berbasis Tujuan



Makna fitur ditentukan oleh tujuan spesifik dari model pembelajaran mesin yang dibangun.

Transformasi Log Server



Tanggal: 2023-10-27
Waktu: 10:30:45
Pesan: User login successful



Terima Kasih

Hanif Fakhurroja ©2026