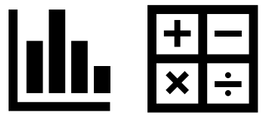
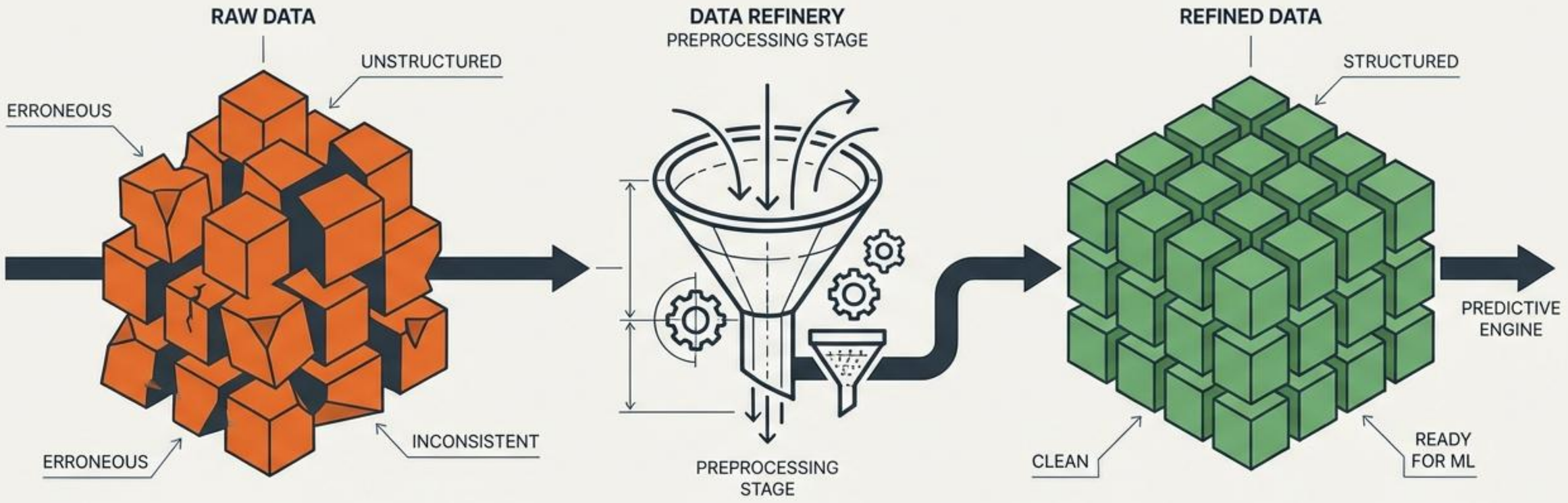
Pertemuan 2 :

DATA PRE-PROCESSING





Data Refinery: Mengubah Data Mentah Menjadi Bahan Bakar Prediktif

Panduan Visual 7 Tahapan Preprocessing Data.

Data mentah di dunia nyata pada dasarnya cacat, beracun, dan tidak dapat digunakan. Preprocessing bukanlah sekadar tugas administratif—ini adalah fondasi rekayasa kritis tempat keandalan machine learning dibangun.

Studi Kasus: Minyak Mentah Digital

Nilai Kosong (NULL)

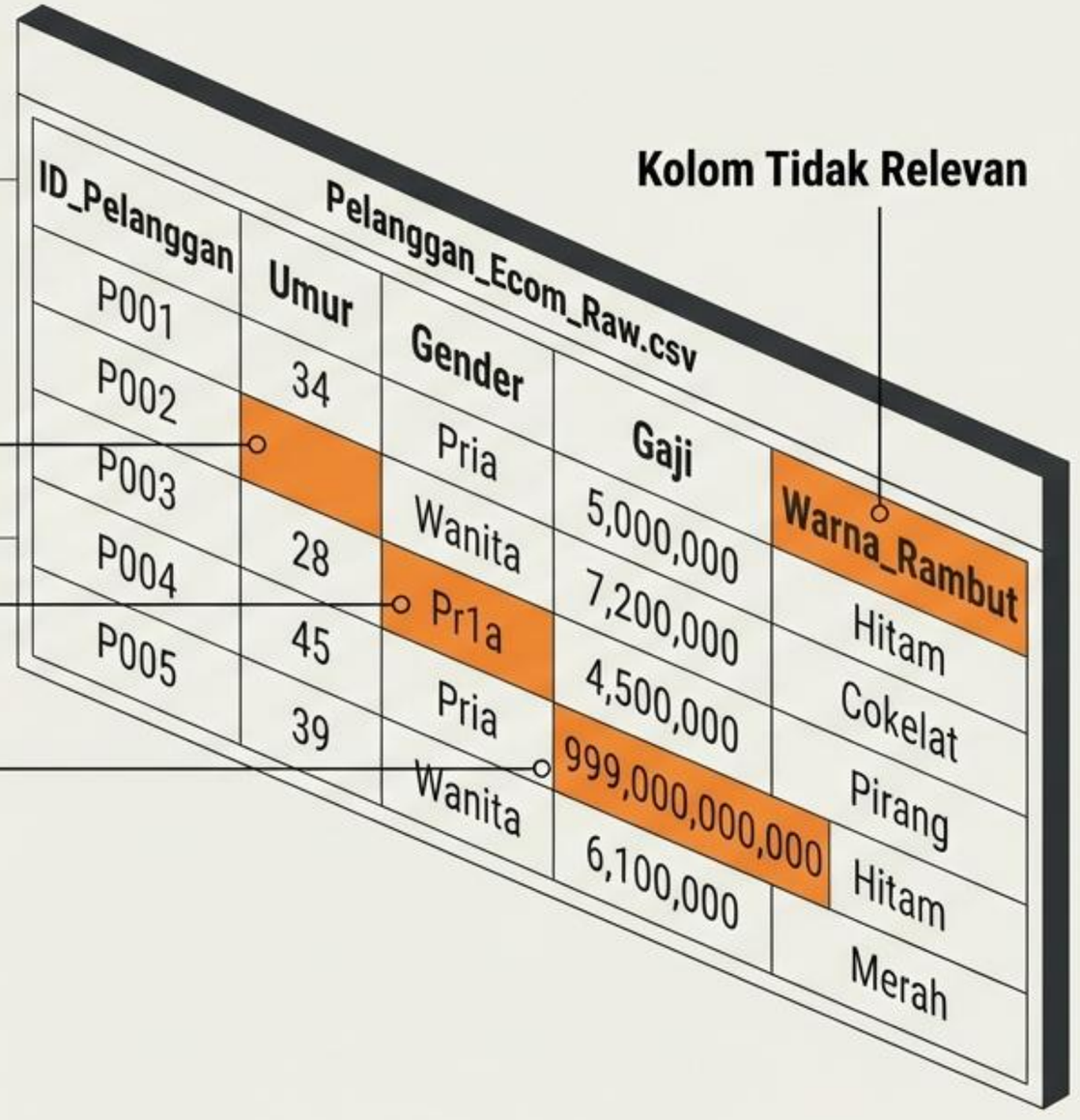
Format Tidak Konsisten

Kolom Tidak Relevan

Skala Ekstrem

Konteks: Membangun model prediksi Churn (berhenti berlangganan) untuk platform E-commerce.

Masalah: Algoritma tidak bisa memproses ambiguitas. Data mentah ini penuh anomali dan membutuhkan diagnosis serta rekayasa sebelum digunakan.



Matriks Diagnostik Preprocessing

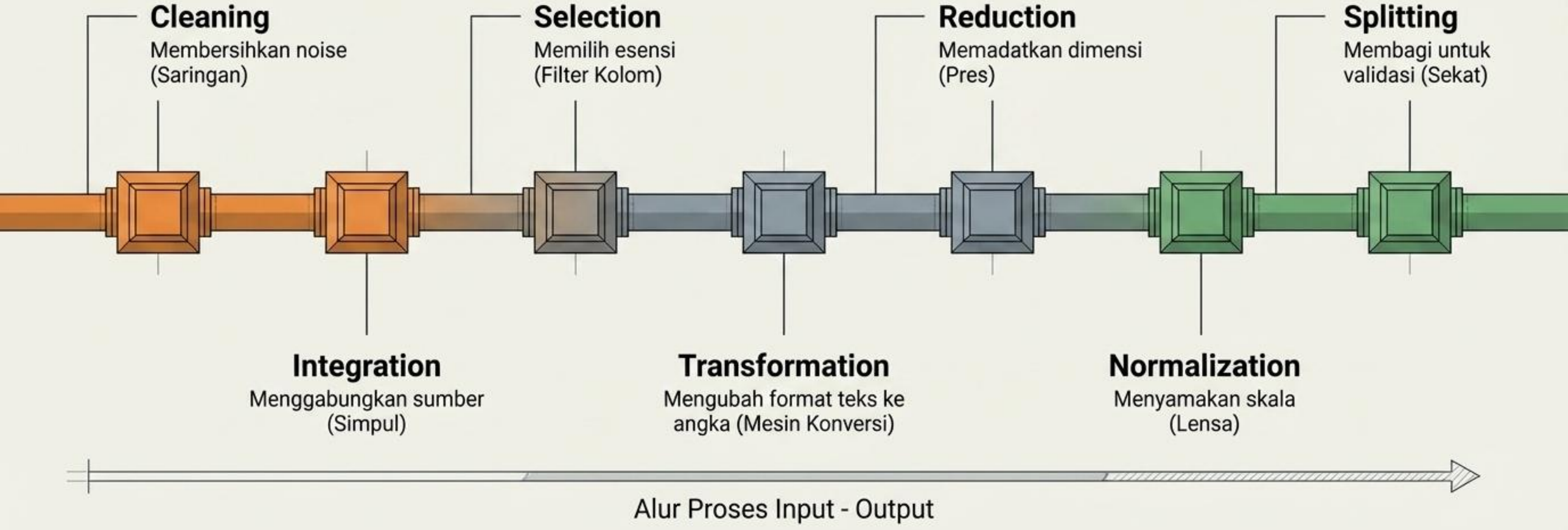
Panduan pemecahan masalah cepat untuk kondisi data yang tidak ideal.

Gejala Data	Tahapan	Tindakan	Hasil Akhir
Banyak sel kosong / NULL	Data Cleaning	Imputasi Mean/Median	Dataset utuh tanpa celah.
Kolom teks (Pria/Wanita)	Data Transformation	One-Hot Encoding	Vektor angka biner (1 dan 0).
Ribuan kolom berisiko overfit	Data Reduction	Ekstraksi PCA	Dimensi padat, esensi terjaga.
Skala angka terlalu jomplang	Data Normalization	Min-Max Scaling	Rentang nilai seragam (0-1).
Model menghafal, gagal di tes	Data Splitting	Pemisahan Train-Test	Evaluasi model objektif.

Data siap digunakan. Kilang pemurnian selesai.
Mesin prediksi siap dihidupkan.

Blueprint Kilang Pemurnian

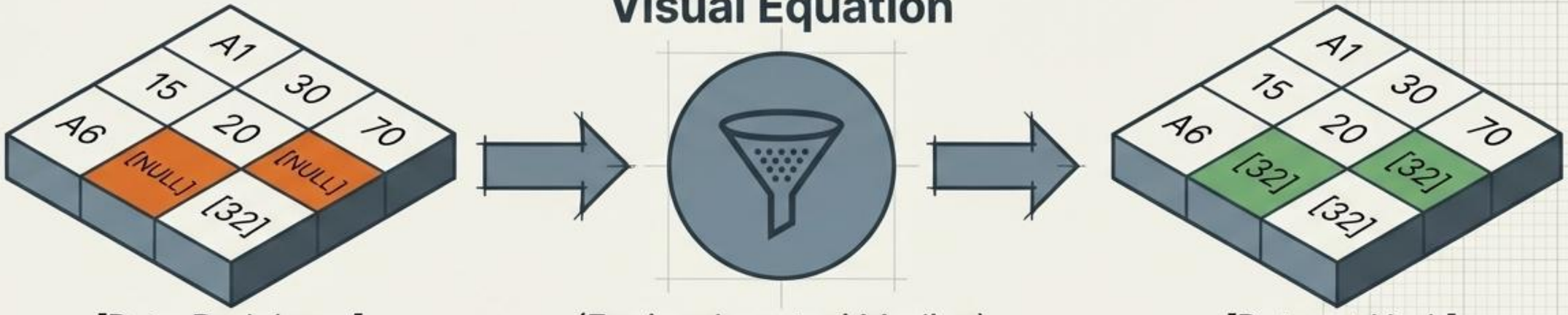
Tujuh stasiun berurutan. Jika data melompati satu stasiun, seluruh jalur pipa akan terkontaminasi.



1. Data Cleaning (Pembersihan)

Menghapus noise, menangani nilai yang hilang (missing values), dan memperbaiki ketidakkonsistenan data agar tidak menyesatkan algoritma.

Visual Equation

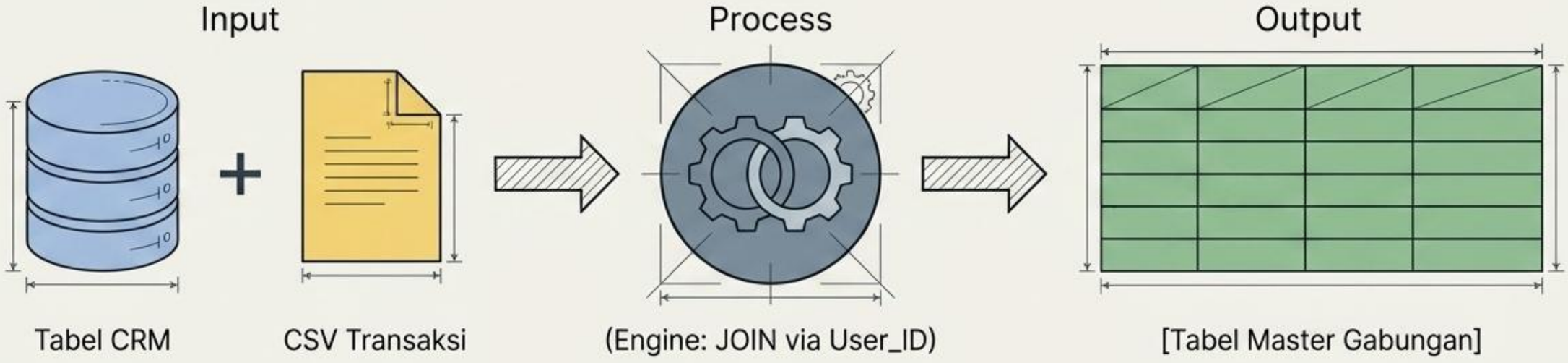


[Data Berlubang] (Engine: Imputasi Median) [Dataset Utuh]

Studi Kasus E-Commerce	
Gejala: Kolom 'Umur' memiliki 15% data kosong. Jika dibiarkan, algoritma akan gagal berjalan.	**Tindakan:** Alih-alih menghapus seluruh baris pelanggan, kita mengisi lubang tersebut dengan nilai tengah (Median = 32) dari seluruh pelanggan lainnya, mempertahankan keutuhan dataset.

2. Data Integration (Integrasi)

Menggabungkan data dari berbagai sumber (database, API, CSV) ke dalam satu dataset master yang koheren.



Sebelum

Tabel A

User_ID	Nama
U01	Budi

Tabel B

User_ID	Total_Beli
U01	Rp 500k

Gejala: Data profil dan log belanja terpisah

Sesudah

Tabel Master

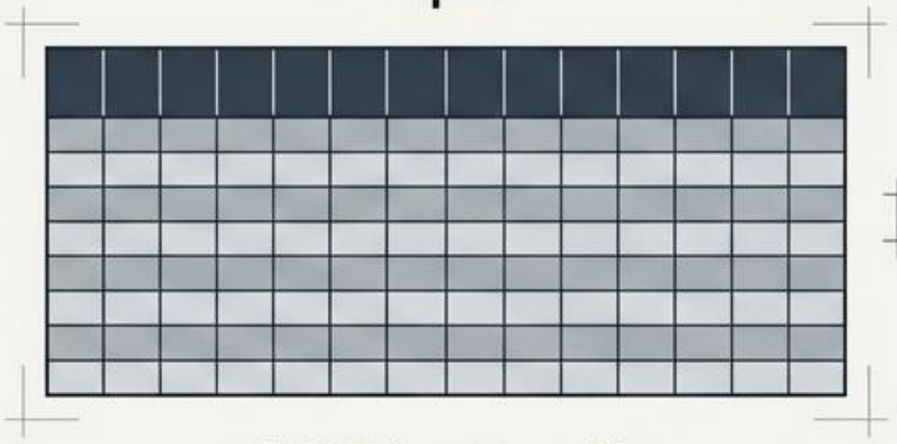
User_ID	Nama	Total_Beli
U01	Budi	Rp 500k

Tindakan: Disatukan menggunakan kunci ID yang sama

3. Data Selection (Seleksi Fitur)

Mengeliminasi kolom (fitur) yang tidak relevan dengan tujuan analisis untuk mengurangi beban komputasi dan mencegah noise tambahan.

Input

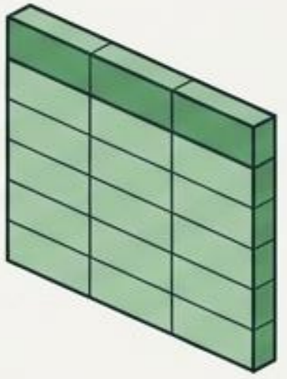


[20 Kolom Mentah]

Visual Equation

(Engine: Filter Relevansi Kausal)

Output



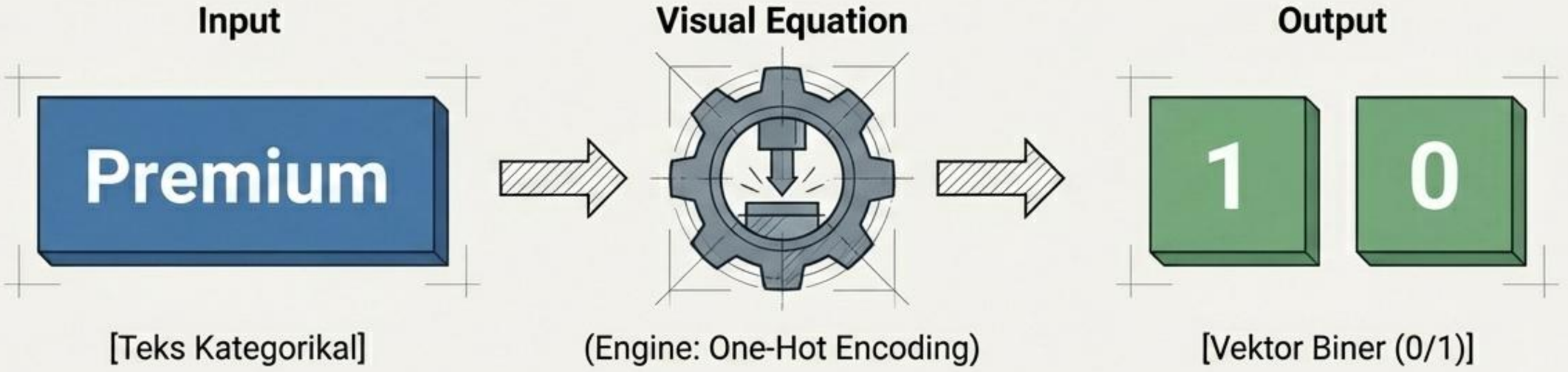
[5 Kolom Inti]

Studi Kasus E-Commerce

Gejala: Model tidak butuh warna sepatu untuk memprediksi churn pelanggan.
Tindakan: Fitur yang tidak memiliki hubungan kausalitas dihapus secara permanen.

4. Data Transformation (Transformasi)

Mengubah wujud data, biasanya dari kategori teks (categorical) menjadi representasi angka (numerical) agar dapat dihitung secara matematis oleh algoritma.



Mentah

Tipe_Akun
Premium
Basic
Basic

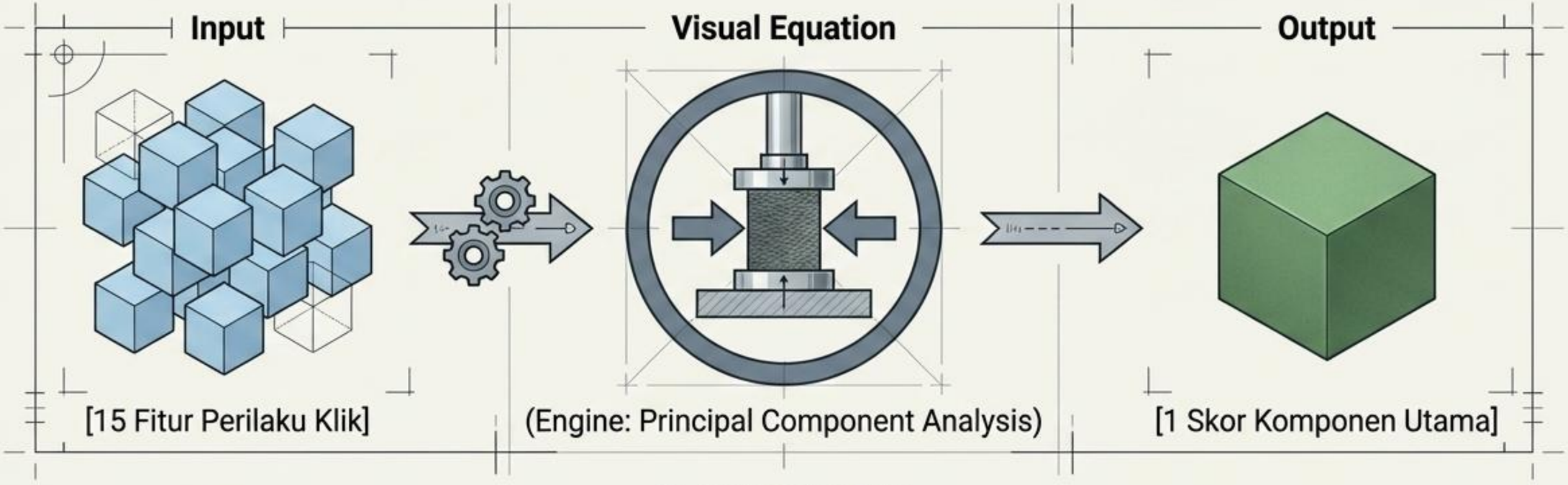
Mesin/Angka

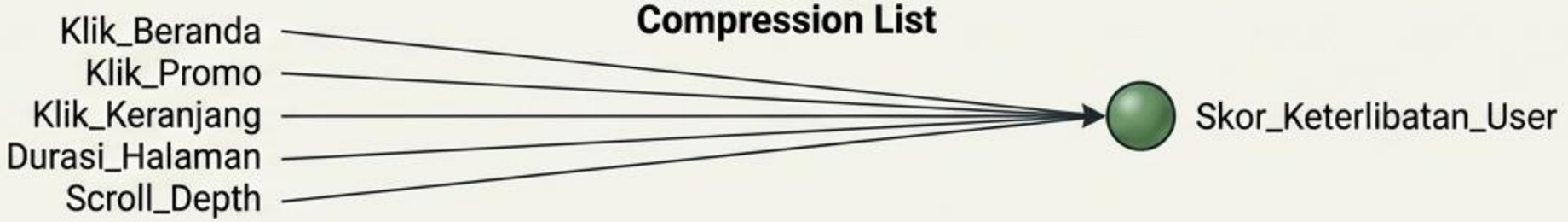
Is_Premium	Is_Basic
1	0
0	1
0	1

Gejala: Model matematika tidak bisa mengalikan kata 'Premium'. **Tindakan:** Konversi teks menjadi kolom indikator biner.

5. Data Reduction (Reduksi Dimensi)

Memadatkan atribut data yang sangat banyak (dimensi tinggi) menjadi representasi yang lebih kecil tanpa menghilangkan esensi pola matematisnya.

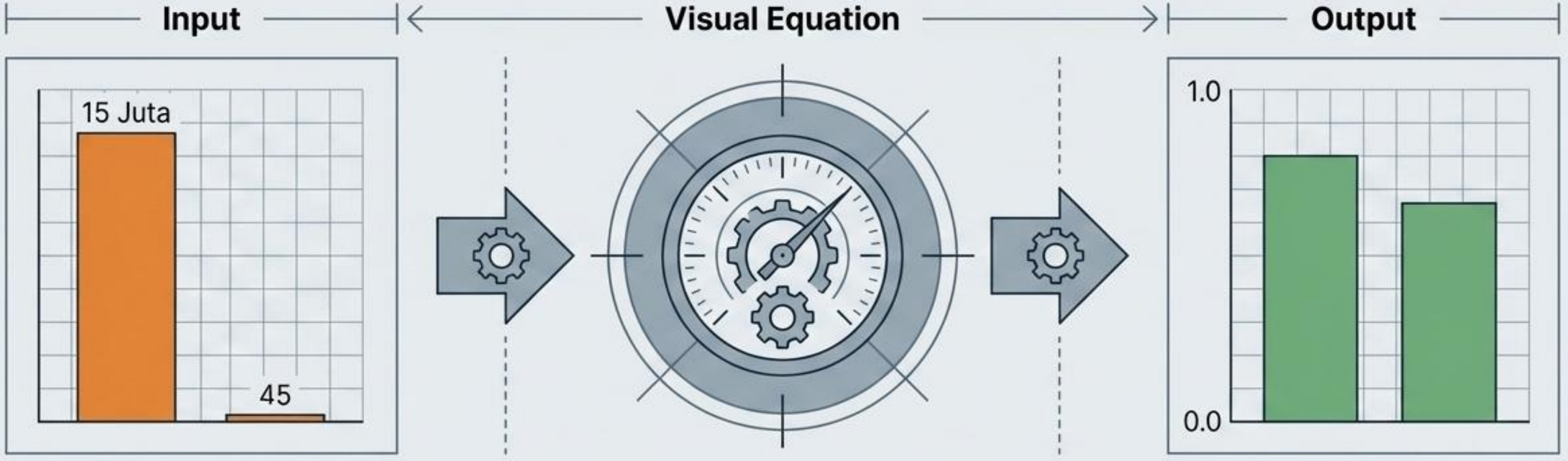




Gejala: Terlalu banyak kolom merekam aktivitas sepele, membuat model lambat dan overfit.
Tindakan: Ekstraksi fitur merangkum 15 metrik klik menjadi 1 skor padat tunggal.

6. Data Normalization (Normalisasi)

Menyesuaikan nilai numerik dari berbagai kolom agar berada dalam rentang skala yang seragam (biasanya 0 hingga 1), mencegah algoritma menjadi bias terhadap angka besar.



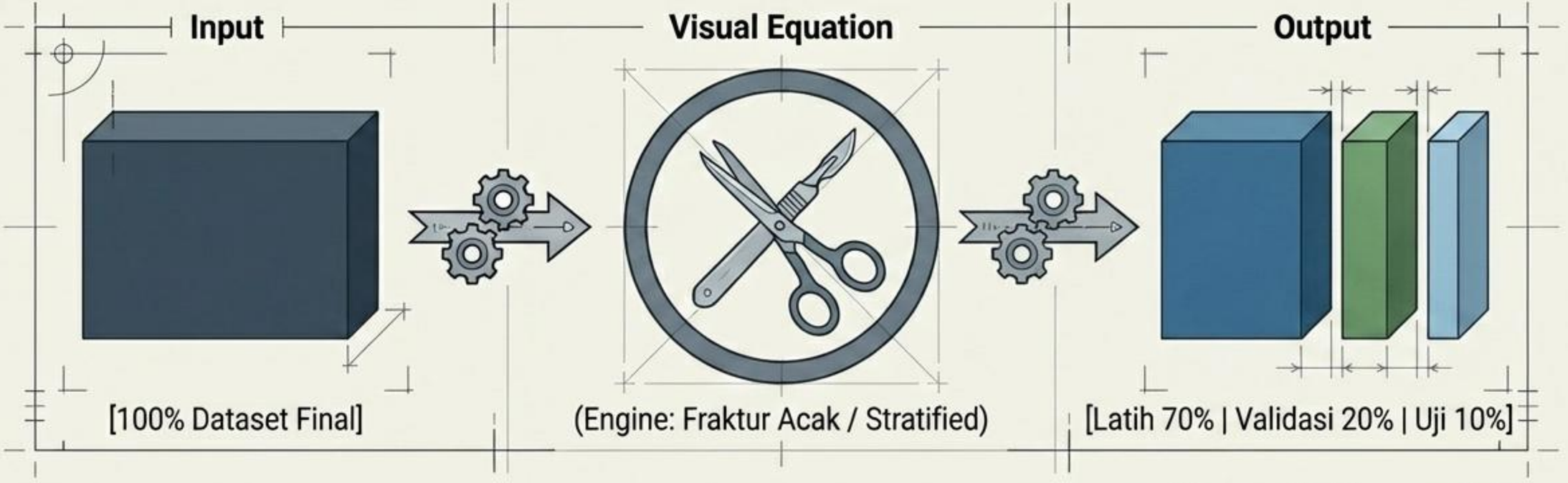
Skala Asli: Kolom 'Total_Belanja' (Rp 15.000.000) akan secara otomatis mendominasi kalkulasi jarak terhadap kolom 'Lama_Sesi' (45 Menit).

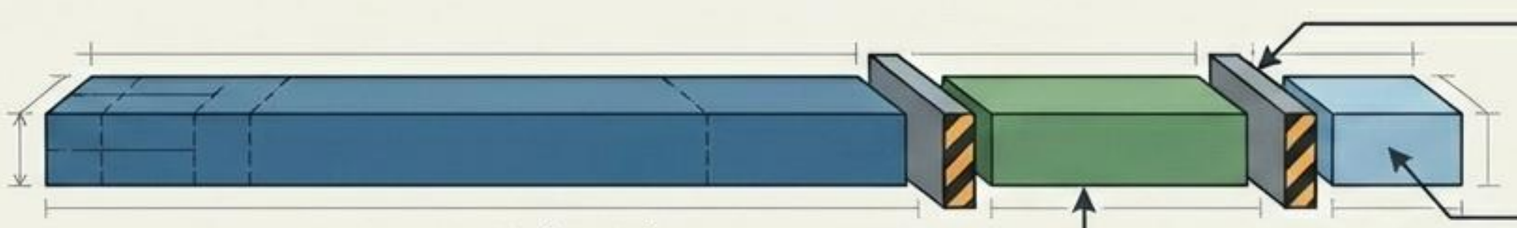
E-Commerce Case Study

Setelah Normalisasi: Belanja diubah menjadi 0.85. Sesi diubah menjadi 0.60. Keduanya kini berbobot setara di mata mesin, tanpa merusak rasio data aslinya.

7. Data Splitting (Pembagian)

Membagi dataset yang sudah bersih menjadi bagian terpisah secara ketat untuk melatih algoritma dan menguji akurasinya pada data yang belum pernah dilihat.





Mencegah Kebocoran Data
(Data Leakage)

Data Latih (Train)
Untuk mengajari AI mengenali pola.

Data Validasi (Val)
Untuk menyetel parameter model.

Data Uji (Test)
Dikunci rapat di brankas.
Ujian akhir.

Anatomi Kesiapan Data

Tujuh langkah teknis bermuara pada empat pertanyaan strategis.
Sebuah model prediktif hanya sekuat pilar terlemahnya.