

PENGUMPULAN DAN PERSIAPAN DATA ILMIAH

Disusun Oleh:

Dr. Yaniasih, STP, MP.

Modul Pelatihan Jabatan Fungsional Analisis Data Ilmiah (BRIN)

Fokus Implementasi: Lembaga Kebijakan Pengadaan Barang/Jasa
Pemerintah (LKPP)



Profil Narasumber

Dr. Yaniasih, S.TP., M.P.

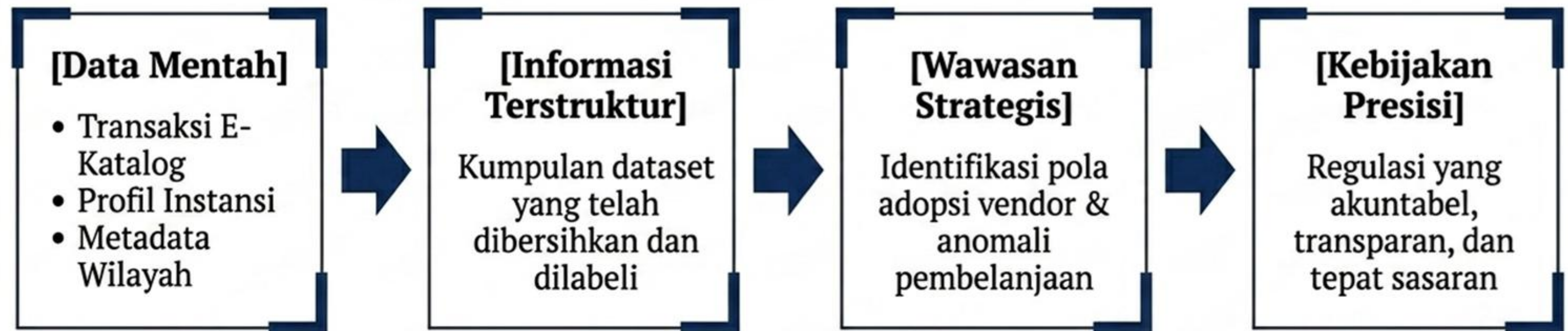
Peneliti, Pusat Riset Sains Data
dan Informasi (BRIN)

Kepakaran & Portofolio Riset:

- **Sains Data & Kecerdasan Artifisial:** Berfokus pada pemrosesan bahasa alami (NLP), Machine Learning, dan Tata Kelola Data (Data Governance).
- **Pemimpin Survei Nasional AI:** Mengorkestrasi pengumpulan dan analitik big data terkait pemanfaatan AI dalam ekosistem riset Indonesia (melibatkan lebih dari 1.800 responden nasional).
- **Pengembang Sistem Triase Multimodal:** Merancang arsitektur Large Language Model (LLM) Human-in-the-Loop untuk pengambilan keputusan medis kritis di Instalasi Gawat Darurat (IGD) Rumah Sakit.

Konteks Analisis Data Ilmiah di KLPP

Di LKPP, data pengadaan bukan sekadar rekaman administratif, melainkan instrumen *early warning* dan fondasi utama *evidence-based policy*.

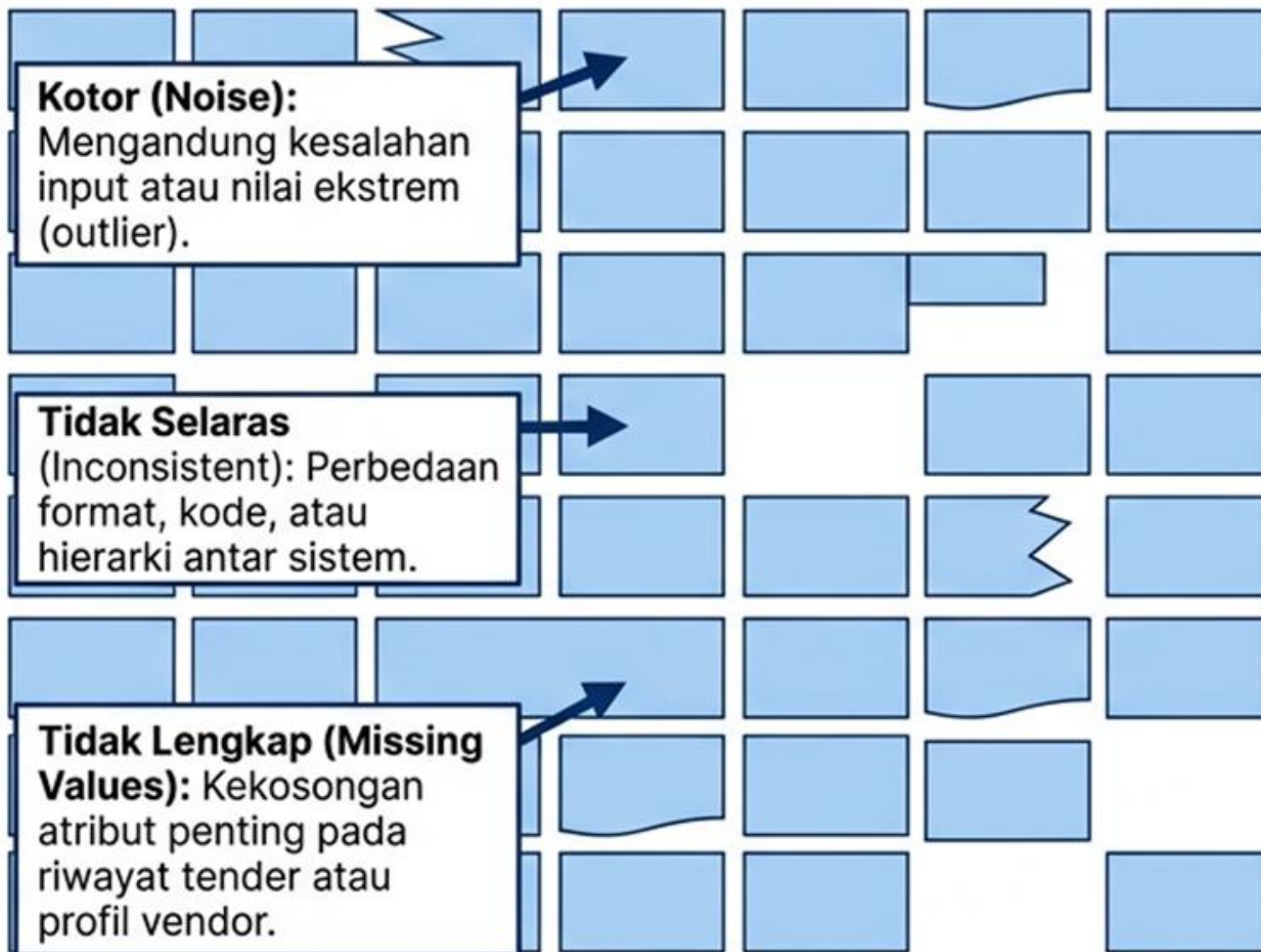


Untuk menjamin kualitas kebijakan di ujung siklus, diperlukan metodologi persiapan data yang ketat dan terstandarisasi di awal siklus.

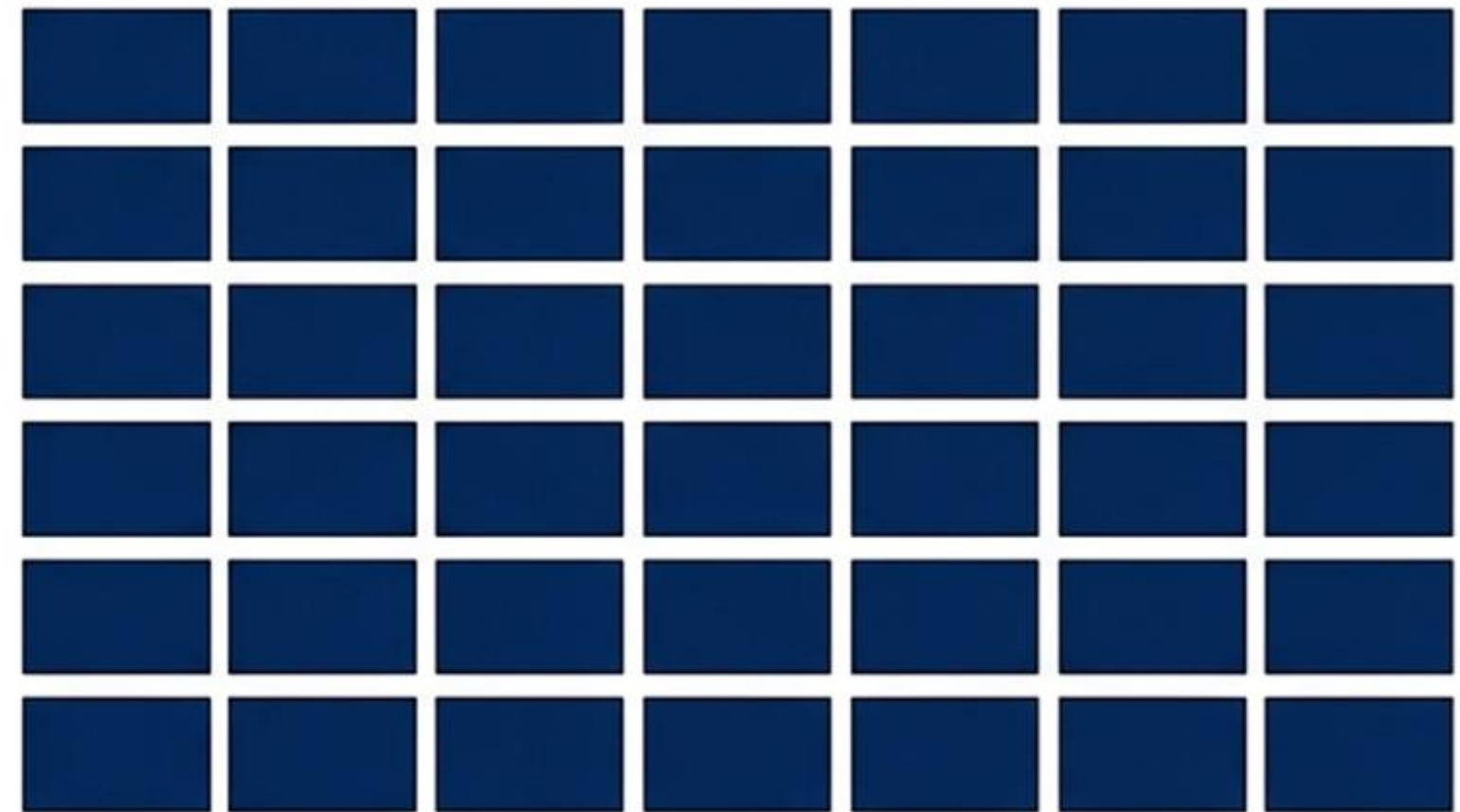
Tantangan Data Pengadaan di Era Digital

Data dalam bentuk aslinya (data dunia nyata) umumnya belum siap untuk digunakan dalam tugas analitik tingkat lanjut.

Karakteristik Data Mentah



Paradigma Baru:



Pra-pemrosesan menyumbang $\pm 80\%$ dari total waktu proyek data. Ini bukan proses administratif ad-hoc, melainkan tahapan ilmiah yang menentukan validitas

Kerangka Kerja Persiapan Data Ilmiah

Berdasarkan Modul Analisis Data Ilmiah (ADI), persiapan data untuk sistem cerdas atau analitik kebijakan publik harus melalui 4 tahapan fundamental:

1. Akuisisi Data:

Ekstraksi terstruktur dari sumber sistem (seperti Existing System SIMRS atau E-Katalog) berdasarkan analisis kebutuhan.

2. Pelabelan Data:

Pemberian konteks, penentuan ground truth, dan penyelarasan semantik.

3. Prapemrosesan Data:

Standardisasi format, pembersihan noise, dan penanganan missing values.

4. Perekayasaan Fitur:

Transformasi variabel operasional mentah menjadi metrik/indikator analitikal strategis.

TAHAP 1: AKUISISI DATA

Pendekatan Sistematis (Top-Down) Sesuai Kebutuhan Analitik

Akuisisi data ilmiah bukanlah sekadar mengunduh dataset (download) atau memperoleh data secara acak. Akuisisi adalah proses pengamatan fenomena (teknis/sosial) dengan menggunakan alat bantu untuk menjawab pertanyaan analitik tertentu.

Tujuan Analitik (Contoh: Deteksi Risiko Gagal Tender)

1. Identifikasi Konteks

Menentukan ruang lingkup fungsionalitas hasil analisis.

2. Integrasi Kebutuhan

Memodelkan aliran informasi dan harmonisasi semantik elemen data.

3. Pemetaan Sumber (Source-to-Target)

Mengidentifikasi sumber data (internal/eksternal) yang berpotensi memenuhi kebutuhan fakta analitik.



Data Mentah Tertarget
(*Targeted Raw Data*)

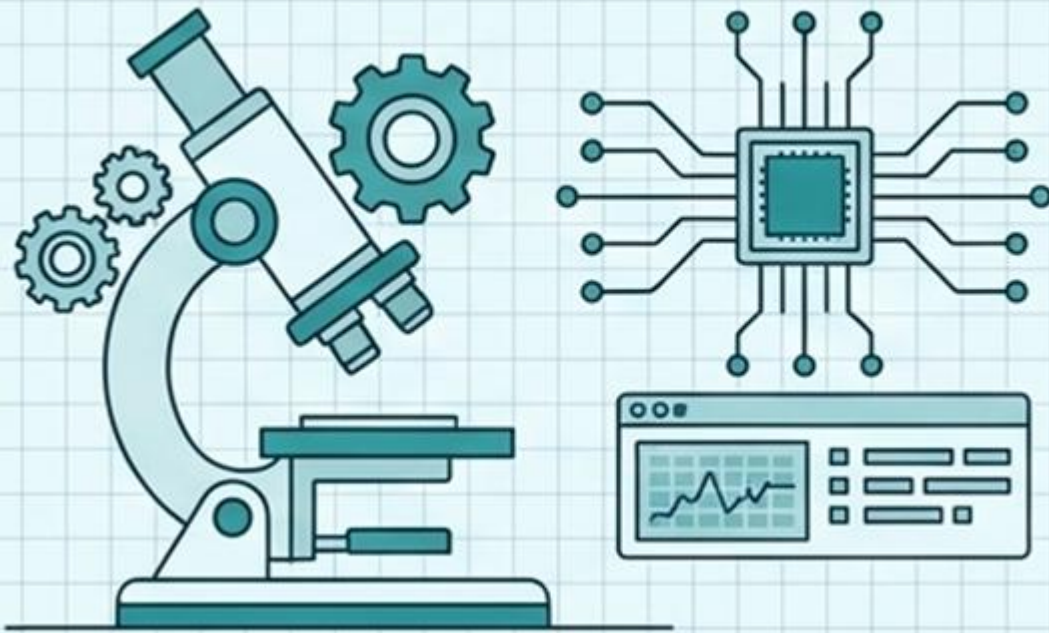
Metode Pengumpulan Data

Metode	Tingkat Kontrol	Instrumen Utama	Keluaran Data
Eksperimen	Kontrol Penuh (Lab)	Mikroskop, Spektroskopi, Sensor	Numerik presisi tinggi, Citra, Audio
Survei	Lingkungan Terstruktur	Kuesioner Digital, Fokus Grup	Kuantitatif & Kualitatif Opini/Perilaku
Web Monitoring	Pasif / Otomatis	Dasbor Mesin Pencari (Alerts/Trends)	Teks skala besar, Tren kata kunci
Media Sosial	Reaktif / Pasif	Dasbor Pihak Ketiga (Analytics)	Sentimen, Metadata, Interaksi (Likes)
Crowdsourcing	Terdistribusi (Masif)	Aplikasi Smartphone (Citizen Science)	Geolokasi, Foto lapangan, Laporan partisipan

Kontrol vs Skala: Eksperimen & Survei

Eksperimen (Lab)

- **Karakteristik:** Lingkungan terkontrol penuh untuk menguji hipotesa kausalitas.
- **Instrumen:** Dari alat dasar (atmometer) hingga sistem kompleks (genome sequencer).



Survei (Digital Form)

- **Karakteristik:** Menangkap opini, perilaku, dan pengetahuan dari populasi secara asinkron.
- **Infrastruktur:** Otomatisasi konversi jawaban responden ke format komputasi terstruktur.

A screenshot of a digital survey form interface. The main question is "How satisfied are you with the service you received?". Below the question are five radio button options: "Very satisfied" (selected), "Somewhat satisfied", "Somewhat satisfied", "Very satisfied", and "Other". To the right of the options is a dropdown menu currently showing "Short answer". Below the dropdown is a list of question types: "Paragraph", "Multiple choice" (selected with a radio button), "Checkboxes" (checked with a checkbox), and "Dropdown". On the right side of the form, there is a "Question summary" panel with the following details: "Question group: About LimeSurvey", "Code: Birthday", "Question: Day of birth", "Type: Date/Time", "Mandatory: No", and "Relevance equation: 1". At the bottom of the summary panel is a button labeled "Add new question to group".

Akuisisi Jejak Digital Otomatis

Menggunakan agregator algoritma untuk mengekstraksi wawasan dari lautan data internet yang masif.

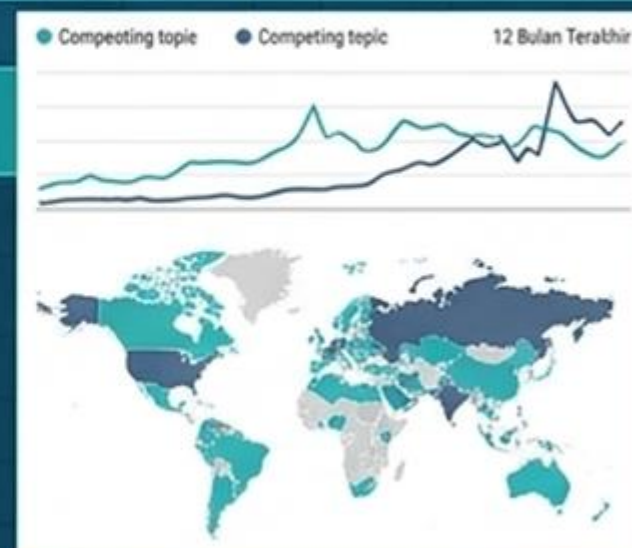
Radar Peringatan (Web)

Bot pelacak kata kunci real-time dengan push notification.

The interface includes a search bar labeled 'Kata Kunci' with a magnifying glass icon and a close button. Below it are several filter options: 'Seberapa sering' (How often) with a dropdown set to 'Sekarang' (Now); 'Sumber' (Source) with a dropdown; 'Bahasa' (Language) with a dropdown set to 'Indonesia'; 'Wilayah' (Region) with a dropdown; 'Kirim ke' (Send to) with a dropdown; and 'Kirim ke' (Send to) with a text input field containing '@gmail.com'. A green button labeled 'Buat Notifikasi' (Create Notification) is at the bottom.

Peta Tren (Web)

Pemetaan polarisasi topik dan popularitas pencarian berdasarkan wilayah.



Ekstraksi Interaksi (Media Sosial)

Menganalisis metrik interaksi (likes, comments, sentimen) dan metadata lintas platform.



Contoh: Akuisisi Data secara Sistematis

Landasan Teori

Definisi (Modul ADI)

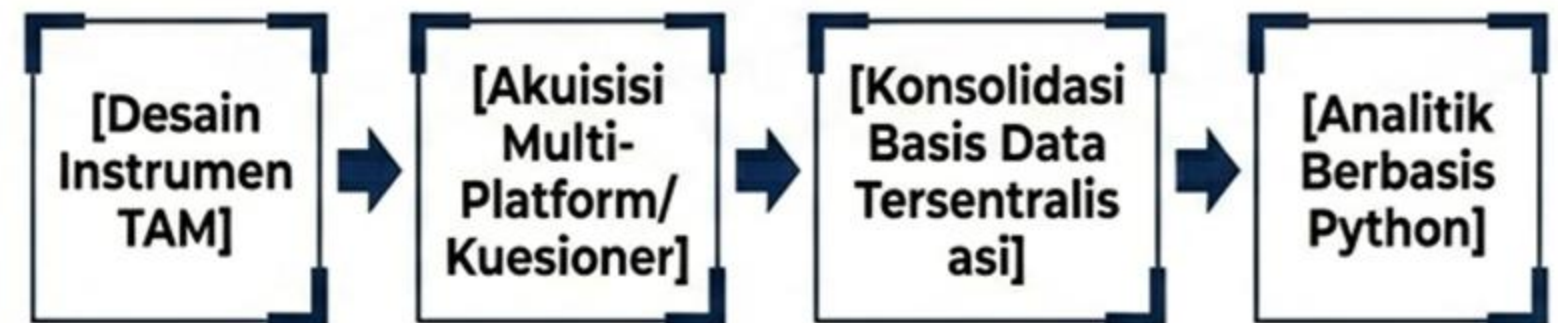
Akuisisi data adalah proses pengamatan fenomena secara langsung maupun tidak langsung untuk menjawab pertanyaan spesifik, dengan mempertimbangkan mempertimbangkan analisis kebutuhan (*context requirement*) dan pemetaan sumber (*source-to-target mapping*). Data tidak boleh dikumpulkan secara acak.

Studi Kasus

Arsitektur Pengumpulan Data: Survei AI Nasional BRIN

- Target: 1.800+ responden (ASN, Peneliti, Akademisi, Industri).
- Metode: Proportional *random sampling* berbasis instrumen Technology Acceptance Model (TAM).

Alur Akuisisi:



Analisis Kebutuhan Data: Studi Kasus E-Katalog LKPP

Akuisisi data di LKPP harus digerakkan oleh kebutuhan kebijakan (*policy-driven*), bukan sekadar mengunduh data yang kebetulan tersedia.

Pertanyaan Kebijakan (Context)	Kebutuhan Data Target (Mapping)
Siapa pengguna terbesar anggaran?	Entitas Instansi / K/L/PD
Apakah terjadi konsentrasi transaksi geografis secara spesifik?	Metadata Wilayah / Kode Pos Pemenang
Bagaimana tingkat pemanfaatan produk lokal vs impor?	Indikator TKDN / Status PDN

Peringatan: Data yang diakuisisi tanpa pertanyaan analitik yang spesifik berisiko tinggi menjadi ‘noise’ bagi infrastruktur analitik institusi.

Menghubungkan Tujuan Analitik dengan Sumber Ekosistem LKPP

Tujuan Analitik Utama: Prediksi Probabilitas Kegagalan Kontrak

Elemen Data yang Dibutuhkan	Sumber Data LKPP (Kandidat)	Harmonisasi Semantik (Target)
Riwayat Transaksi Vendor	Sistem E-Katalog (API / DB)	Penyelarasan ID Vendor antar sistem.
Nilai Pagu & HPS	SPSE (Sistem Pengadaan Secara Elektronik)	Penyamaan format nilai nominal (Rupiah).
Metode Pemilihan & Kategori Barang	Portal data.lkpp.go.id	Standardisasi klasifikasi KBLI.

Latihan 1

5-10 Menit

- Tentukan tujuan analisis data
- Tentukan cara pengumpulan data
- Tulis di room chat, format:
Nama_Tujuan_Cara

TAHAP 2: PELABELAN DATA

Memberikan Konteks untuk Otomatisasi (Supervised Learning)

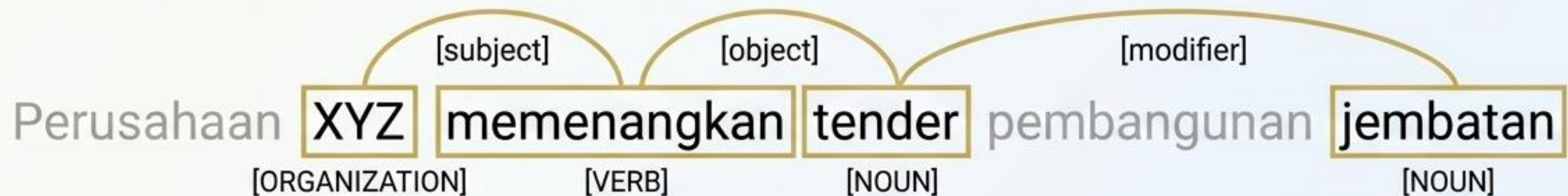
Pelabelan data (data annotation/tagging) adalah proses pemberian informasi tambahan (konteks) pada data mentah agar dapat dimengerti dan diolah dengan benar oleh mesin.

Identifikasi Obyek: Menandai entitas spesifik yang menjadi fokus observasi.

Identifikasi Semantik: Mengatasi ambiguitas makna dalam teks.

Struktur Mesin: Menciptakan data latih (*training data*) untuk algoritma *Supervised Learning*.

Perusahaan XYZ memenangkan tender pembangunan jembatan.



Metode Pelabelan: Manual vs. Otomatisasi

Menentukan taktik pelabelan sangat bergantung pada jenis data dan kompleksitas pengadaan.

	 Pelabelan Manual (Human Expert)	 Pelabelan Otomatis (Rule-Based Machine)
Penerapan Optimal	Dokumen adendum kontrak, identifikasi celah hukum, audit investigatif.	Transaksi harian e-Katalog, klasifikasi kategori barang (KBKI).
Karakteristik	Sangat akurat untuk konteks yang ambigu.	Cepat, skalabel untuk jutaan baris data.
Tantangan	Membutuhkan waktu lama & rentan subjektivitas auditor. 	Kualitas variatif; membutuhkan pakar untuk menyusun rules awal. 
Alat Bantu	VIA Annotator, LabelMe	Natural Language Toolkit (NLTK), spaCy

Contoh Pengalaman di BRIN



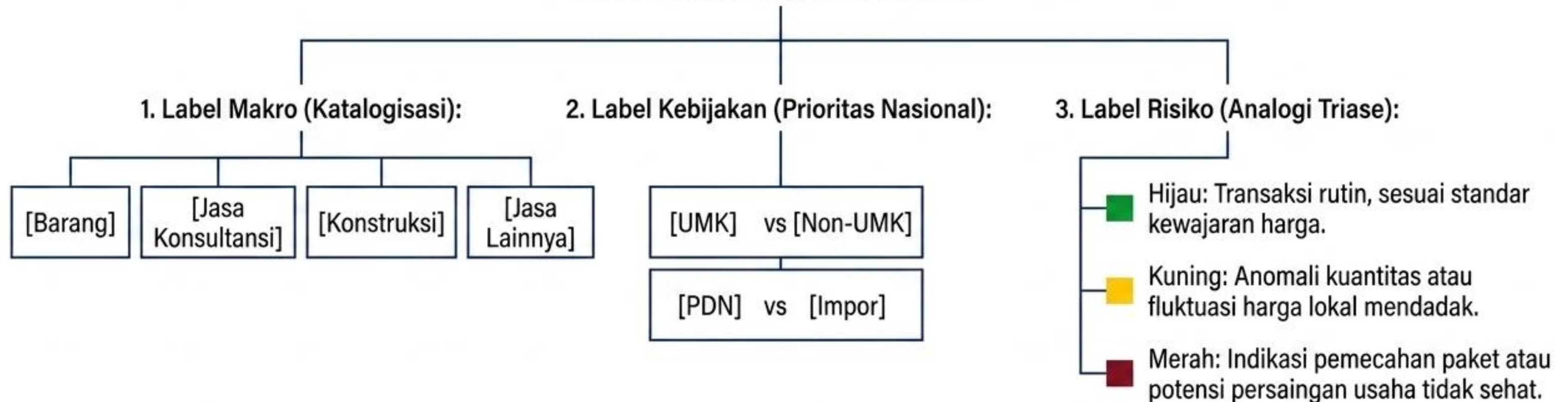
Studi Kasus: Sistem Triase Medis BRIN

- **Konteks:** Menentukan tingkat urgensi penanganan pasien di Instalasi Gawat Darurat.
- **Label Kategori:** Hijau (Non-Urgent), Kuning (Urgent), Merah (Emergency).
- **Prinsip Utama:** AI bertindak sebagai pendukung keputusan klinis objektif, bukan pengganti kepakaran manusia. Validasi akhir tetap dilakukan oleh tenaga medis ahli.

Analogi Pelabelan Triase untuk Vendor & Transaksi LKPP

Sama seperti Sistem Triase yang mengkategorikan risiko klinis untuk menyelamatkan nyawa, LKPP mengkategorikan transaksi untuk mitigasi risiko fraud dan pembinaan efisiensi vendor.

Skema Pelabelan Multidimensi di LKPP

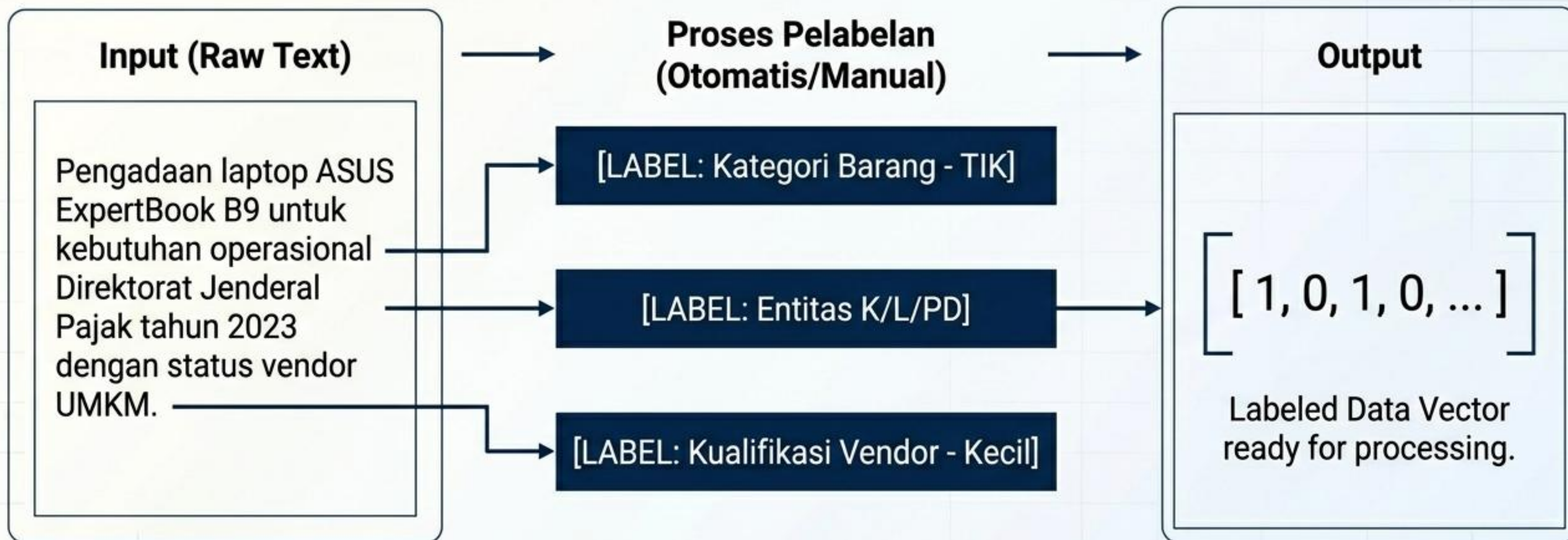


Tantangan utama: Menghindari bias pakar dan menjaga konsistensi anotasi (Inter-annotator agreement).

Studi Kasus 2: Pelabelan Teks Kontrak & Vendor LKPP

Dari Teks Tidak Terstruktur Menjadi Entitas Berlabel

Dalam ekosistem pengadaan, banyak informasi tersimpan dalam deskripsi teks bebas (*unstructured text*). Pelabelan mengekstrak nilai dari teks tersebut.



Latihan 2

10-15 Menit

Dari dataset di <https://data.lkpp.go.id/id/dataset/>
pilih data yang bisa dilakukan pelabelan dan tujuannya untuk apa

A	B	C	D	E
jenis_katalog	pengelola_katalog	nama_etalase	kualifikasi_usaha	jumlah_penyedia
Katalog Lokal	Kab. Kebumen	Petugas Penunjang Kegiatan Pemerintah Daerah Kabupaten Kebumen	Usaha Mikro	1484
Katalog Nasional	LKPP	Peralatan Elektronik dan Pendukungnya	Usaha Kecil	1310
Katalog Nasional	LKPP	Perkakas	Usaha Kecil	1008
Katalog Sektor	Kementerian Kesehatan	Fasilitas Kesehatan Sektor	Usaha Kecil	825
Katalog Nasional	LKPP	Peralatan Elektronik dan Pendukungnya	Usaha Mikro	781
Katalog Lokal	Kab. Magelang	Jasa Tenaga Pendukung Perkantoran Kabupaten Magelang	Usaha Mikro	717
Katalog Sektor	Kementerian Kesehatan	Fasilitas Kesehatan Sektor	Non-Usaha	683
Katalog Nasional	LKPP	Peralatan Pendidikan	Usaha Mikro	592
Katalog Nasional	LKPP	Peralatan Elektronik dan Pendukungnya	Non-Usaha	572
Katalog Nasional	LKPP	Peralatan Perkantoran	Usaha Mikro	565
Katalog Sektor	Kementerian Pekerjaan Umum	Pekerjaan Preservasi Jalan Bidang Bina Marga	Usaha Mikro	543
Katalog Sektor	Kementerian Pekerjaan Umum	Pekerjaan Preservasi Jalan Bidang Bina Marga	Usaha Mikro	447
Katalog Sektor	Kementerian Kesehatan	Fasilitas Kesehatan Sektor	Usaha Mikro	427
Katalog Lokal	Aceh	Rumah Layak Huni	Non-Usaha	419
Katalog Sektor	Kementerian Pekerjaan Umum	Pekerjaan Preservasi Jalan Bidang Bina Marga	Non-Usaha	412
Katalog Nasional	LKPP	Perkakas	Usaha Mikro	409
Katalog Nasional	LKPP	Alat Laboratorium	Usaha Mikro	372
Katalog Nasional	LKPP	Perkakas	Non-Usaha	357

Label jenis lembaga
- Kementerian
- Lembaga

Label kategori barang
- Barang/Jasa
- Kesehatan,
Pendidikan, dll

Label kategori ukuran
- > 1000 banyak
UKM
- < 10 jarang

→ Tujuan analisis
misal indentifikasi
daerah tertentu perlu
meningkatkan jumlah
UKM bidang tertentu

TAHAP 3: PRAPEMROSESAN DATA

Sains Kualitas Data: Validasi, Pembersihan, dan Standarisasi

Data dunia nyata seringkali tidak siap analitik. Prapemrosesan mengatasi tiga masalah kualitas utama:



1. Ketidaklengkapan (*Incompleteness*)

Kehilangan nilai atribut (*missing values*). Diatasi dengan teknik Imputasi (mengisi dengan mean/median/modus) atau Eliminasi record.



2. Gangguan (*Noise & Outliers*)

Anomali nilai ekstrem yang mendistorsi model. Diatasi dengan deteksi statistik (Z-Score) dan teknik smoothing (binning, regresi).



3. Ketidakkonsistenan (*Inconsistency*)

Perubahan sistem pengkodean atau format (misal: 'Jkt' vs 'Jakarta'). Diatasi dengan transformasi standarisasi.



Studi Kasus 3: Matriks Diagnostik Prapemrosesan LKPP

Tindakan Korektif pada Anomali Data Pengadaan

Gejala pada Data LKPP (Symptom)	Identifikasi Ilmiah (Method)	Tindakan Prapemrosesan (Action)
Nilai Pagu/HPS tercatat Rp 1.500.000.000.000 untuk pengadaan ATK (Kesalahan ketik/Typo).	Deteksi Outlier (Penyimpangan Ekstrem dari Distribusi Standar).	Smoothing menggunakan rata-rata historis pengadaan serupa atau Karantina data.
Kolom 'Masa Berlaku Kontrak' kosong pada dataset E-Katalog.	Identifikasi Missing Value.	Imputasi nilai median, atau Eliminasi record jika tidak relevan.
Format input 'NPWP' vendor bervariasi (dengan tanda baca '-' vs tanpa tanda baca).	Deteksi Ketidakkonsistenan Format (String Mismatch).	Standarisasi menggunakan fungsi pembersihan string (Regex) menjadi format baku 15 digit.

Kritikalitas Deteksi Anomali (Outlier)

Satu kesalahan input dapat menghancurkan seluruh algoritma dan mencondongkan model prediktif.



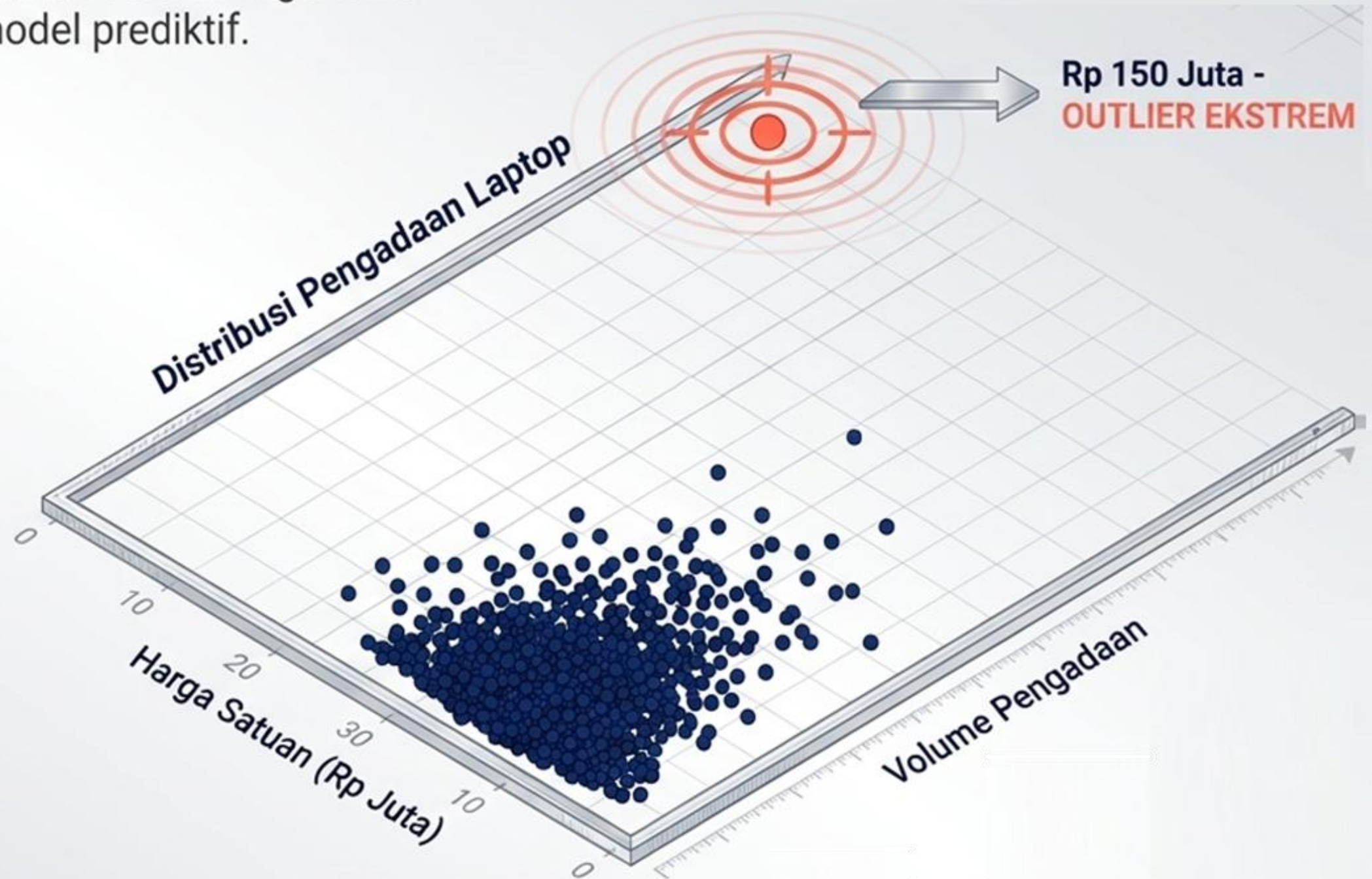
Merusak Model

Satu outlier ekstrem menarik rata-rata (mean) belanja nasional secara drastis, **mendistorsi kewajaran harga.**



Sinyal Investigasi (**Fraud**)

Apakah mungkin menjadi indikasi adanya "mark-up" misalnya sehingga perlu pendalaman lebih lanjut?



Implementasi Prapemrosesan pada Dataset LKPP

Masalah klasik dalam data birokrasi adalah Entity Resolution (Standardisasi Nomenklatur) dan penanganan Outlier.

Kasus 1: Standardisasi Entitas (Data Cleaning)	
Raw Data (Kotor / Variatif)	Clean Data (Standar Mesin)
Kemenkeu	Kementerian Keuangan
Kementerian Keu.	
Kementrian Keuangan RI	

Kasus 2: Deteksi Outlier (Bukan Sekadar Menghapus)
- Skenario Transaksi: Pembelian pengadaan pulpen tunggal tercatat senilai Rp 1.500.000.000 (indikasi human error input digit).
- Tindakan Presisi: Sistem mendeteksi ini sebagai <u>outlier statistik</u> . Alih-alih menghapusnya secara otomatis (yang menghilangkan jejak error), analisis <u>mengkarantina data</u> tersebut untuk verifikasi silang.

Latihan 3

10-15 Menit

Dari dataset di <https://data.lkpp.go.id/id/dataset/>

Cek apakah ditemukan data yang outlier, missing value, atau inkonsisten

→ rekomendasi perbaikan data

TAHAP 4: PEREKAYASAAN FITUR (FEATURE ENGINEERING)

Mentransformasi Atribut Menjadi Representasi Informatif

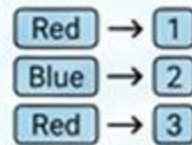
Berbeda dengan pembersihan data (*cleaning*) yang mengembalikan data ke kondisi normal, perekayasa fitur (*organizing*) mentransformasi data menjadi bentuk yang secara matematis optimal untuk algoritma Machine Learning.

Pemahaman Level Data



Nominal, Ordinal, Interval, Rasio (menentukan operasi matematis yang diizinkan).

Transformasi Data Kategorikal



Mengubah teks kategori menjadi format numerik yang dapat dikomputasi mesin.

Seleksi Fitur



Memilih subset atribut terbaik untuk menghindari kutukan dimensi menggunakan teknik *Filtering*, *Wrapper*, atau *Embedded*.

Cleaning

Dec 16 9am: server shutting down & restarting
Dec 16 10am: server resuming all operations

Cleaning

Dec 16 9am: server shutting down & restarting
Dec 16 10am: server resuming all operations

Organizing

Dec 16 9am: server shutting down & restarting
Dec 16 10am: server resuming all operations

Organizing

Time	Event Code
Dec 16 9am	SD&R
Dec 16 10am	SRAO

Proses mengekstraksi, menggabungkan, dan mentransformasi data mentah tingkat rendah menjadi fitur metrik strategis yang langsung merepresentasikan permasalahan analitik.



Transformasi Konseptual (Bukti Empiris BRIN):

Survei AI Nasional:



Sistem Triase Klinis:



Fitur (indikator) yang dikonstruksi dengan pemahaman domain yang baik lebih berharga daripada penerapan algoritma yang sekadar kompleks.

Rekayasa Fitur untuk Pengawasan Pengadaan Nasional

Bagaimana kita mentransformasi baris transaksi tabular E-Katalog yang masif menjadi **dasbor wawasan operasional** pimpinan LKPP?

[Kuadran 1] Rasio Keberpihakan Lokal

Formula: $(\text{Total Belanja Produk Dalam Negeri (PDN)} / \text{Total Anggaran Belanja Instansi}) \times 100\%$

[Kuadran 2] Indeks Kompetisi Vendor

Konsep: Menghitung **jumlah vendor unik** yang **memenangkan kontrak** dalam satu instansi per **kuartal/tahun** (mendeteksi monopoli).

[Kuadran 3] Velocity Transaksi (Kecepatan Adopsi)

Konsep: Mengukur waktu rata-rata (lead time) dari tanggal produk tayang di E-Katalog hingga terjadinya klik persetujuan pembelian.

[Kuadran 4] Indeks Konsentrasi Geografis

Konsep: Persentase dominasi serapan vendor dari satu provinsi tertentu terhadap total volume transaksi nasional.

Latihan 4

10-15 Menit

Dari dataset di <https://data.lkpp.go.id/id/dataset/> atau data lain yang dimiliki
Buat konsep rekayasa fitur yang bisa dilakukan dan output yang direncanakan

Matriks Komparasi Paradigma Sains Data Lintas Sektor

Sintesis implementasi **metodologi persiapan data**: Dari domain medis dan riset akademik menuju pengawasan pengadaan publik.

Tahapan Analitik Dasar	Sistem Triase Klinis (BRIN)	Survei AI Nasional (BRIN)	Sistem Pengadaan Publik (LKPP)
Akuisisi	Rekam Medis (Teks, Tanda Vital, Citra)	Kuesioner Skala Nasional (1.800 Responden)	Ekstraksi Database E-Katalog, SIKaP, SIRUP
Pelabelan (Ground Truth)	Tingkat Kegawatan Prioritas (Merah/Kuning/Hijau)	Klaster Adopsi & Literasi Teknologi	Status Kapabilitas Vendor (UMK/Non), Level Risiko Audit
Prapemrosesan	Anonimisasi Pasien, Validasi Teks Anamnesis	Pembersihan Duplikasi & Penyaringan Responden Noise	Standarisasi Nomenklatur KLPD, Penanganan Missing Values
Perekayaan Fitur	Metrik Cumulative Risk Score	Indeks Kesiapan AI (AI Readiness Index)	Rasio Adopsi PDN, Indeks Konsentrasi Transaksi Geografis

Data Governance sebagai Fondasi Kebijakan Berbasis Bukti

"Garbage In, Garbage Out" (GIGO).

Kecerdasan Buatan (AI) maupun **Dasbor Analitik** (BI) terancang tidak akan pernah menghasilkan kebijakan yang valid jika data dasarnya kotor, bias, dan tidak terstruktur.

Pilar Tata Kelola Data Pengadaan

- | | | |
|--|---|--|
| 1. Kualitas Data Fundamental: Validitas, kelengkapan, dan akurasi sejak titik entri (input) pertama kali di sistem. | 2. Metadata Terstandarisasi: Pemahaman semantik yang sama antar unit kerja di lingkungan LKPP. | 3. Etika & Keamanan Privasi: Pengelolaan data transaksi sensitif dengan prinsip privacy-preserving dan audit log berkelanjutan. |
|--|---|--|

Transformasi digital pengadaan tidak dimulai dari penerapan AI, tetapi dari kedisiplinan tata kelola kualitas data institusi.

Penutup



Dr. Yaniasih, S.TP., M.P.

Pusat Riset Sains Data dan Informasi - Badan Riset dan Inovasi Nasional (**BRIN**)
Mentransformasi Ekosistem Riset menuju Data-Driven Policy