



Statistika Machine Learning

#1 Meeting

Introduction Machine Learning

Ferdian Bangkit Wijaya, S.Stat., M.Si
NIP. 199005202024061001





Evolusi Machine Learning



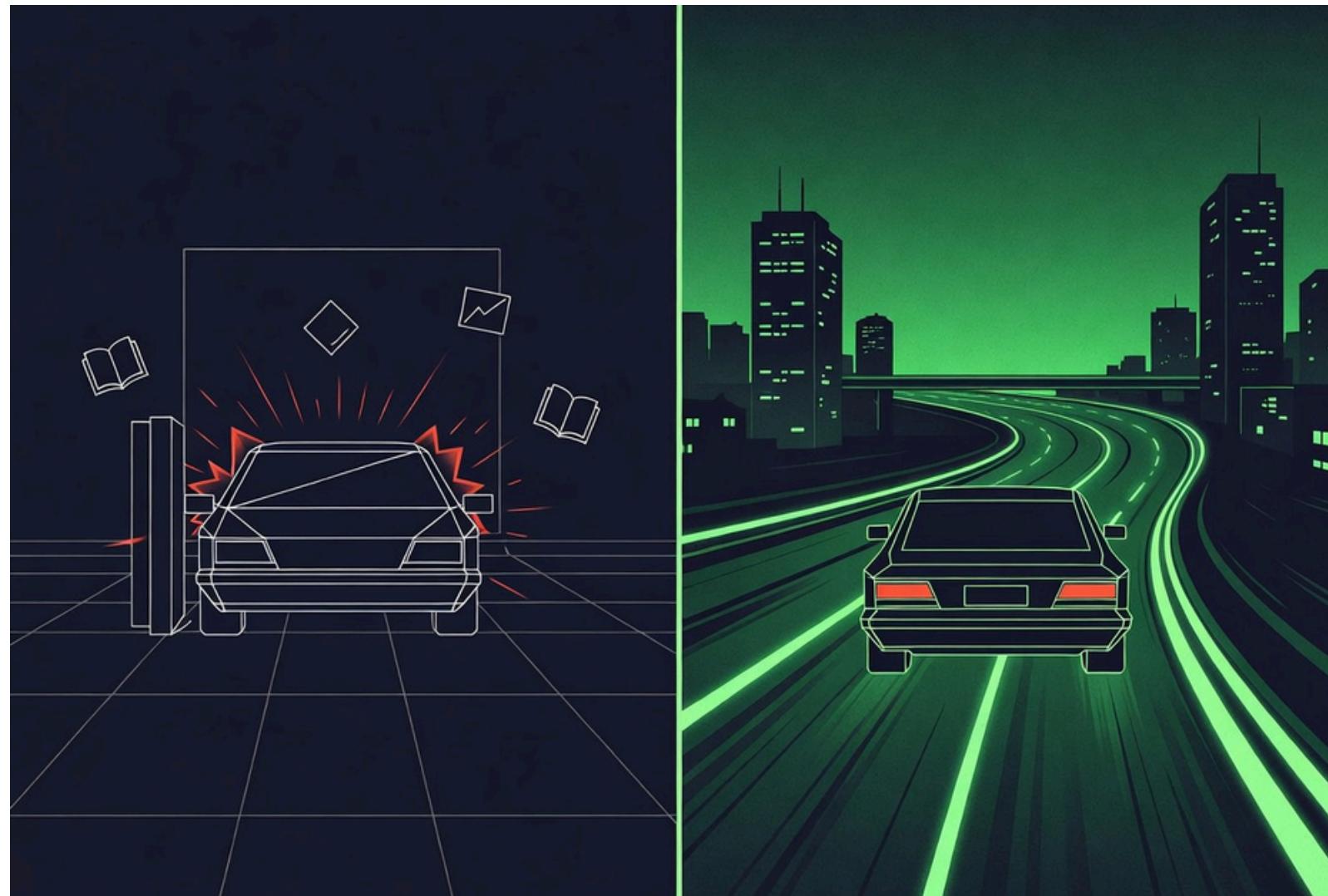
- Definisi: Bagian dari Artificial Intelligence (AI) yang memberikan kemampuan sistem untuk mengekstraksi **pengetahuan dari data**.
- Transisi Metodologi: Pergeseran dari Pemrograman Tradisional (Instruksi Eksplisit) ke **Automated Learning** (Mesin menghasilkan aturannya sendiri)
- Fondasi Statistika: Pemanfaatan teknik statistika dan algoritma untuk mengidentifikasi **pola sistematis** dalam dataset.
- Otonomi Sistem: Proses perbaikan performa secara berkelanjutan melalui **intervensi manusia yang minimal** (Minimum Human Intervention).

Kapabilitas Machine Learning



- Simbiosis Algoritma & Data: Kemampuan model untuk **mengekstraksi hubungan fungsional** antara variabel input dan target dari data historis.
- Analisis Data Kompleks: Kapasitas dalam menangani **unstructured data** (citra, teks, sensor) yang sulit diproses pemrograman konvensional.
- Generalisasi Prediktif: Kemampuan sistem dalam memberikan keputusan atau prediksi akurat pada **data baru yang belum pernah ditemui** (Unseen Data).
- Akurasi Berbasis Kualitas: Integritas hasil model ditentukan oleh **kualitas (quality)** dan **volume (quantity)** data input (Prinsip Garbage In, Garbage Out).

Mekanisme Machine Learning



- Fase Pelatihan (Learning Phase): **Proses pemetaan** (mapping) data input terhadap output menggunakan algoritma tertentu.
- Iterasi Perbaikan: **Penyesuaian parameter** internal model secara otomatis untuk meminimalkan tingkat kesalahan (error rate).
- Eksplorasi Pola Tersembunyi: **Mengidentifikasi tren** makro, perilaku subjek, dan dinamika pasar dalam dataset skala besar.
- Proses Inference: Tahapan di mana model yang telah terlatih digunakan untuk menghasilkan keputusan atau **prediksi secara instan**.

Implementasi Machine Learning



- Finansial & Perbankan: Deteksi anomali pada transaksi (fraud detection) dan estimasi pergerakan harga instrumen keuangan.
- Sektor Kesehatan: Diagnosis penyakit berbasis rekam medis dan proyeksi persebaran wabah secara preventif.
- Personalized Experience: Pengembangan sistem rekomendasi konten dan produk pada platform hiburan serta e-commerce.
- Efisiensi Strategis: Otomasi tugas repetitif dan pengenalan pola untuk reduksi biaya operasional di industri.

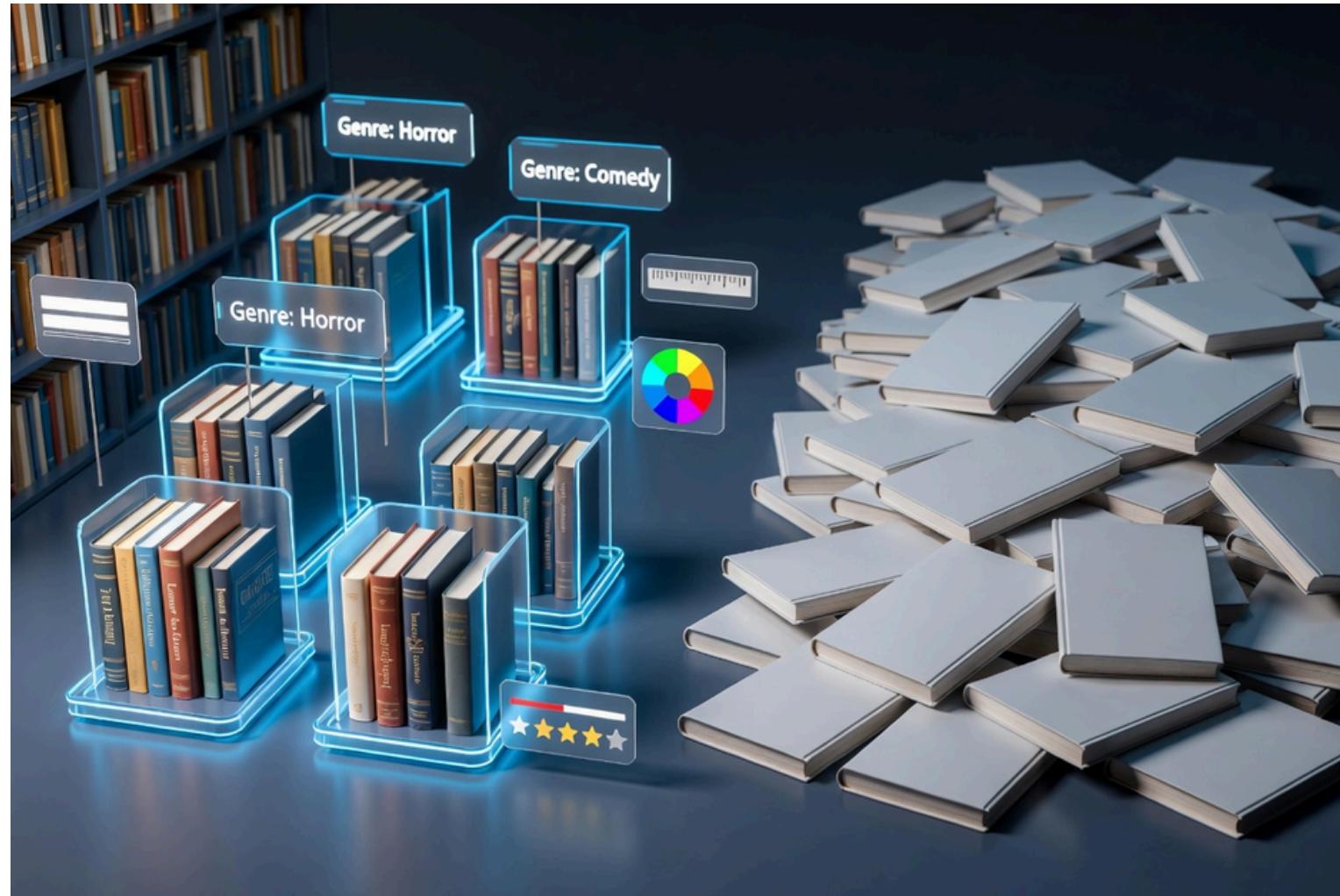


Arsitektur Alur Kerja ML



- Data Acquisition: Proses **pengumpulan data mentah** dari berbagai sumber (database, API, sensor, atau survei).
- Data Wrangling: Transformasi data mentah menjadi **format yang bersih** dan siap olah melalui proses cleaning dan formatting.
- Model Building: Pemilihan **algoritma yang sesuai** dan proses training menggunakan training set.
- Model Evaluation: Pengujian performa model menggunakan **metrik statistik** untuk memastikan akurasi dan reliabilitas.
- Deployment & Monitoring: Implementasi model ke sistem **produksi** dan **pemantauan performa** secara berkelanjutan di dunia nyata.

Data Acquisition :Fungsi Data



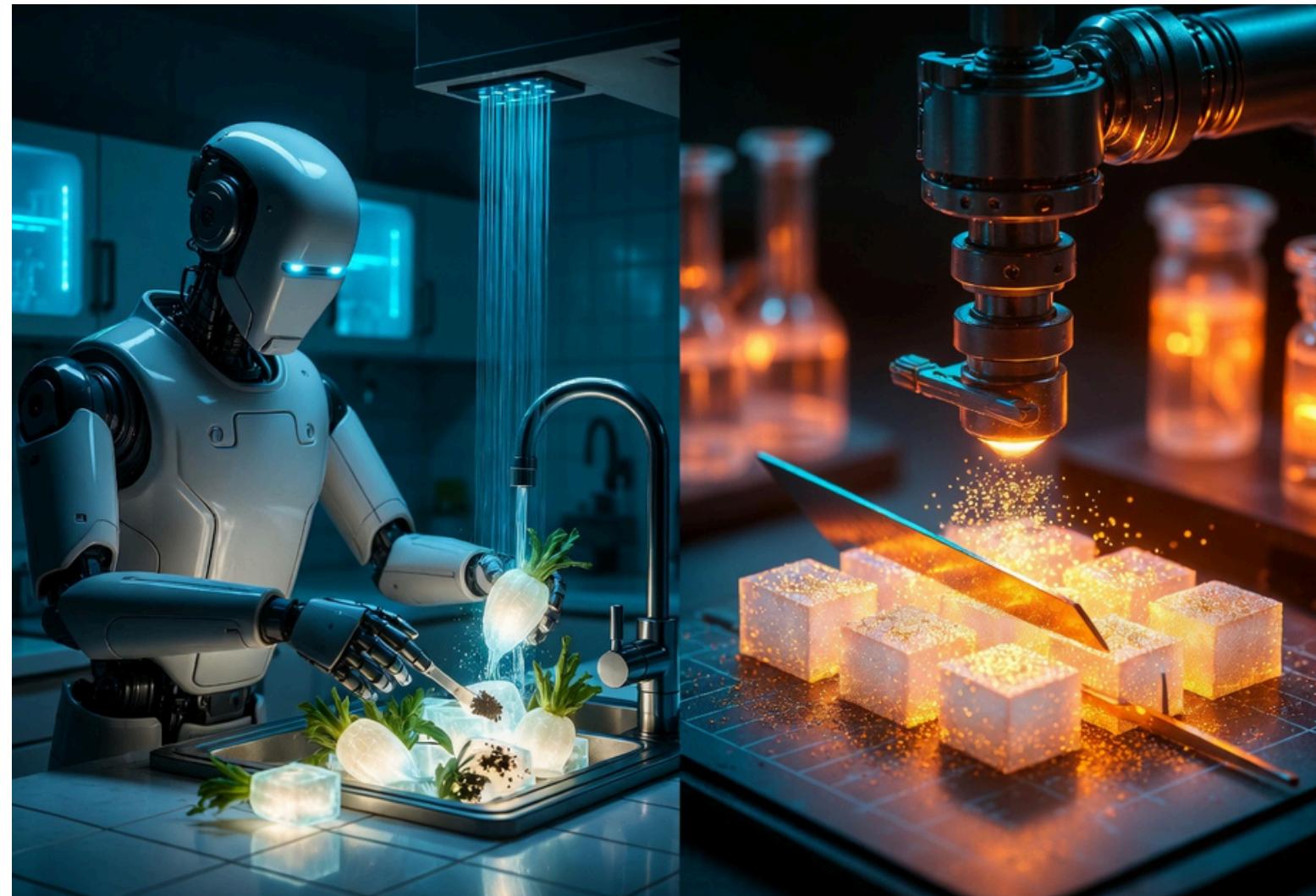
- Fondasi: Data sebagai unit observasi dasar untuk **ekstraksi pola** dan pembentukan fungsi pemetaan model.
- Data Berdasarkan Label:
 - Labeled Data: Dataset yang memiliki variabel **target** (ground truth) untuk tugas Supervised Learning.
 - Unlabeled Data: Dataset **tanpa variabel target**, digunakan untuk eksplorasi struktur dalam Unsupervised Learning.
- Tipe Data Berdasarkan Sifat:
 - Numerical: Data kuantitatif (interval & rasio).
 - Categorical: Data kualitatif (nominal & ordinal)

Data Acquisition : Pro/Cons



- Data Advantages:
 - Automation: Efisiensi dalam pengambilan keputusan berulang secara masif.
 - Personalization: Penciptaan profil unik pengguna (YouTube/Netflix Recommendations).
- Data Disadvantages (Risiko & Tantangan):
 - Algorithmic Bias: Risiko diskriminasi jika data latih **tidak representatif**.
 - Privacy & Security: Kerentanan terhadap **penyalahgunaan data pribadi** dan serangan siber.
 - Interpretability: Tantangan dalam menjelaskan **logika keputusan** pada model kompleks (Black Box).

Data Wrangling : Preprocessing Data



- Data Cleaning: **Bersih-bersih** data meliputi penanganan missing values, penghapusan data duplikat, dan identifikasi outliers.
- Normalization & Scaling: Transformasi fitur ke dalam **skala yang seragam** untuk mempercepat konvergensi algoritma.
- Feature Selection: Proses pemilihan **variabel yang paling relevan** guna mengurangi kompleksitas model dan komputasi.
- Feature Engineering: Konstruksi **fitur baru** dari data mentah untuk meningkatkan daya prediksi model.



Model Building : Data Splitting



- Training Set: Subset data terbesar yang digunakan algoritma untuk **mempelajari hubungan** antar fitur dan target.
- Validation Set: Digunakan untuk **evaluasi sementara** guna optimasi hyperparameters dan mencegah overfitting.
- Test Set: Data independen yang hanya digunakan satu kali di akhir untuk **menguji kemampuan generalisasi** model pada Unseen Data.
- Golden Rule: Larangan keras menggunakan data uji (Test Set) dalam proses pelatihan untuk menghindari kebocoran informasi (Data Leakage).



Model Buliding : Learning Model



- Esensi Learning: Proses matematis untuk meminimalkan **perbedaan antara prediksi model dengan kenyataan** (Ground Truth).
- Komponen Belajar:
 - Input (Features): Informasi yang dipelajari.
 - Target (Label): Jawaban benar yang ingin dicapai.
- Model: "Mesin" yang mencoba **menebak** hubungan keduanya.
- Tujuan Akhir: Mendapatkan **bobot (weights)** paling **optimal** agar prediksi setepat mungkin.



Model Buliding : Loss Function & Optimization



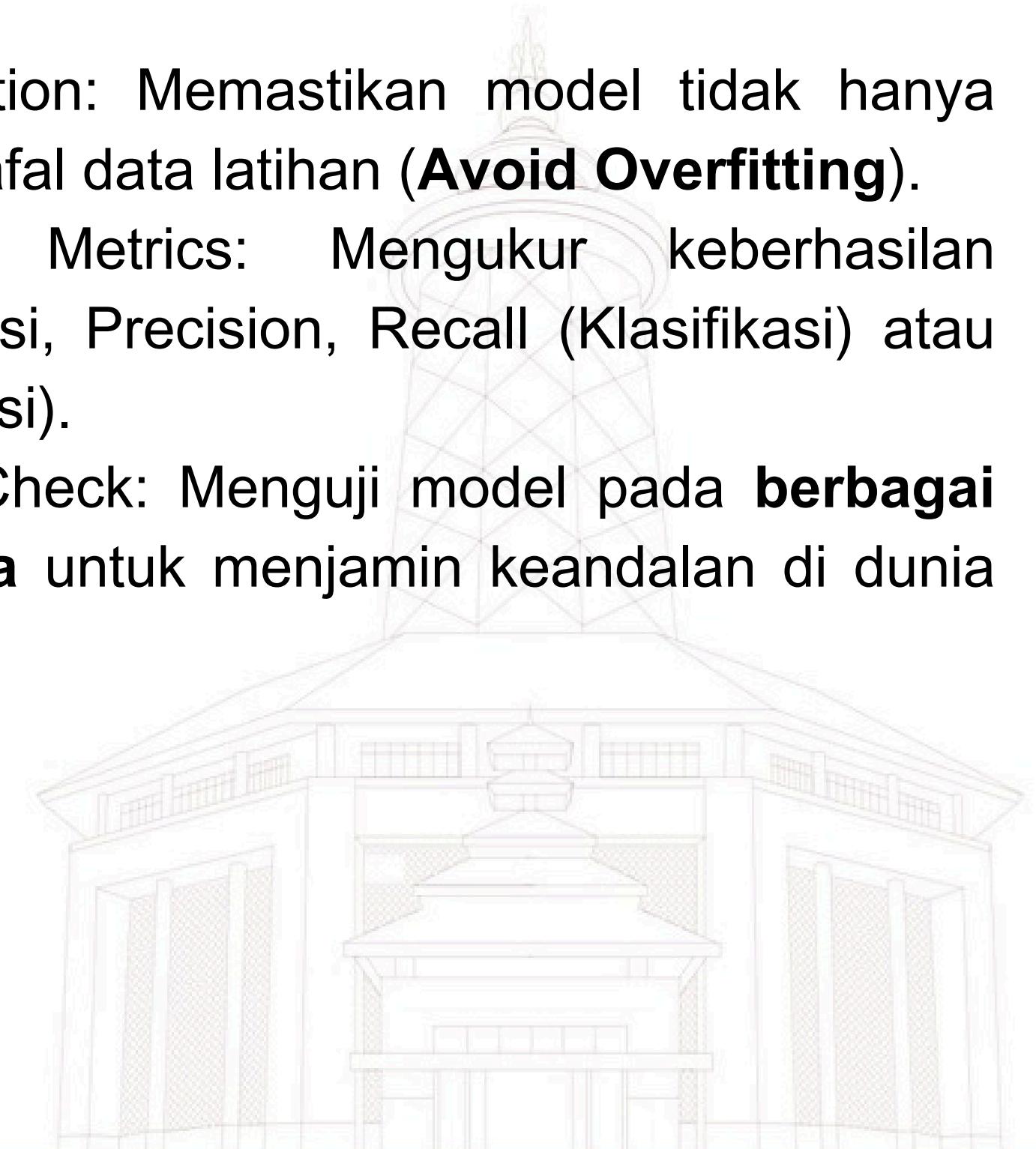
- Loss Function: Metrik penalti yang mengukur deviasi atau "**jarak**" antara prediksi model dengan nilai aktual (Ground Truth).
- Optimization (Gradient Descent): Algoritma matematis untuk **menemukan arah** perubahan parameter yang dapat meminimalkan nilai Loss.
- Weight Update: **Penyesuaian bobot** internal model secara iteratif berdasarkan gradien yang ditemukan selama proses optimasi.
- Convergence: Titik di mana algoritma mencapai nilai **Loss minimum** (optimal) dan tidak ada perbaikan signifikan lagi pada performa model.



Model Buliding : Model Evaluation



- Model Validation: Memastikan model tidak hanya pintar menghafal data latihan (**Avoid Overfitting**).
- Performance Metrics: Mengukur keberhasilan melalui Akurasi, Precision, Recall (Klasifikasi) atau RMSE (Regresi).
- Robustness Check: Menguji model pada **berbagai skenario data** untuk menjamin keandalan di dunia nyata.



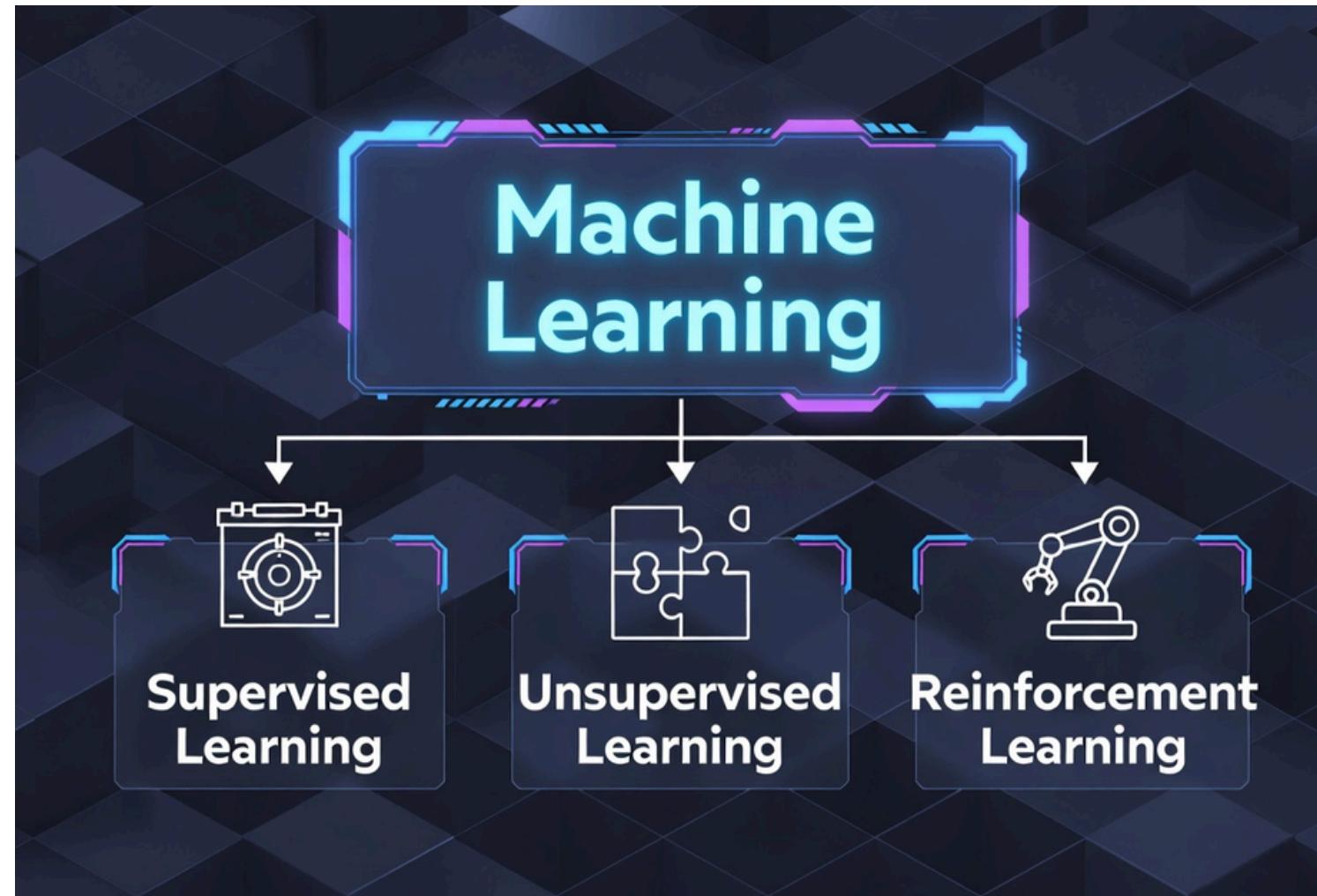
Deploying & Monitoring



- Model Deployment: Proses integrasi model yang telah teruji ke dalam **infrastruktur produksi** (Web, Aplikasi Mobile, atau Embedded Systems).
- Real-time Monitoring: Pengawasan berkelanjutan terhadap **metrik operasional** (latensi, penggunaan memori) dan **metrik performa** (akurasi, deteksi anomali).
- Data & Concept Drift: Identifikasi penurunan performa model akibat **perubahan pola data** di dunia nyata seiring berjalannya waktu.
- Feedback Loop & Retraining: Mekanisme pembaruan model **secara periodik** menggunakan data terbaru untuk menjaga relevansi prediksi.

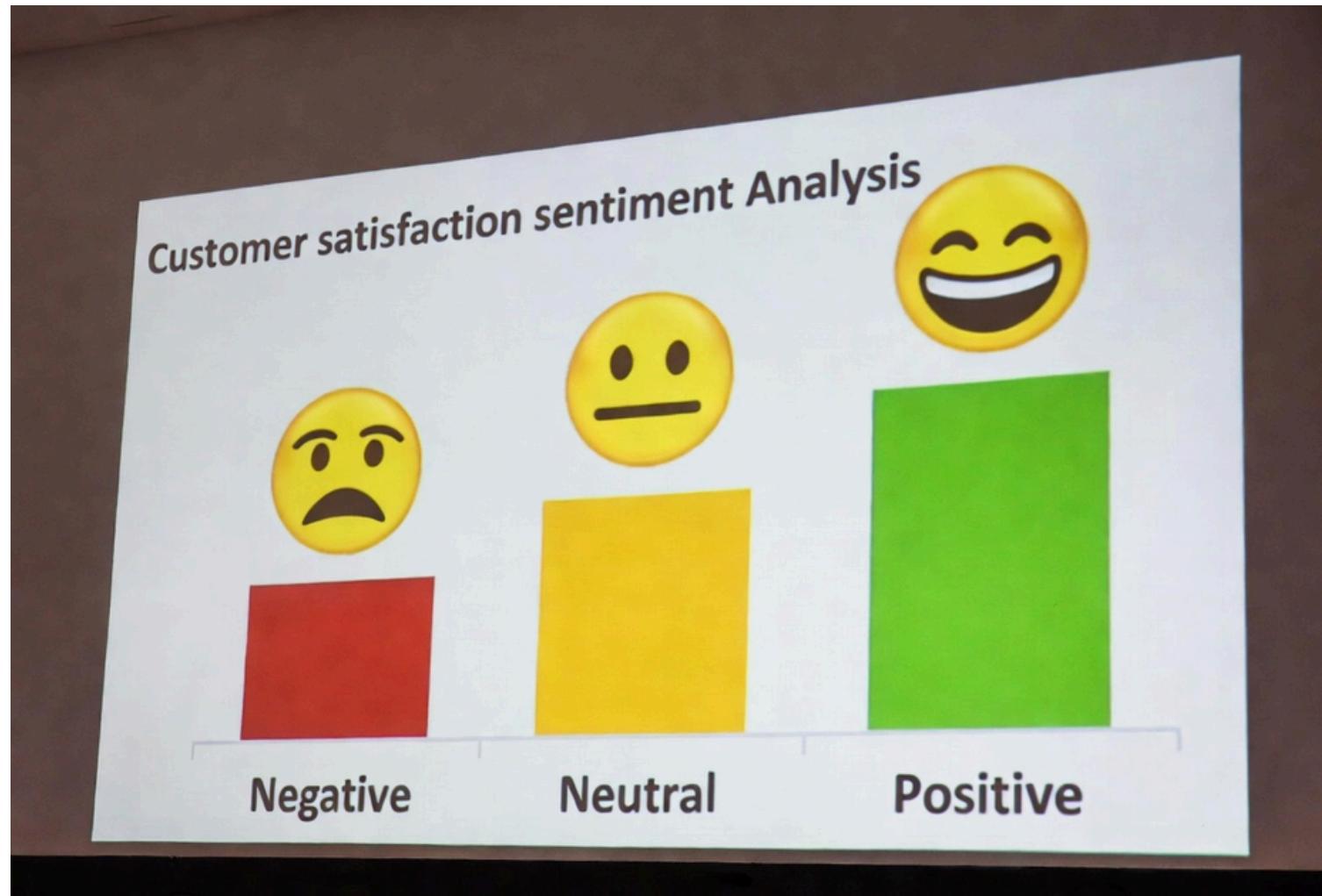


Pilar Machine Learning



- Kategorisasi Model: Klasifikasi algoritma berdasarkan cara sistem memproses informasi dan ketersediaan label (target variable).
- Tiga Pilar Utama:
 - Supervised Learning: Pembelajaran terarah dengan panduan label yang jelas.
 - Unsupervised Learning: Eksplorasi mandiri untuk menemukan struktur tersembunyi dalam data.
 - Reinforcement Learning: Pembelajaran berbasis interaksi agen dengan lingkungan melalui reward dan punishment.

Supervised Learning



- Mekanisme: Proses pelatihan model menggunakan dataset yang sudah memiliki variabel target (Ground Truth) sebagai referensi utama.
- Klasifikasi (Classification): Memprediksi kategori diskret atau label kualitatif.
 - Credit Scoring: Menentukan nasabah "Layak" atau "Tidak Layak" menerima pinjaman.
 - Customer Churn: Memprediksi apakah pelanggan akan "Berhenti" atau "Tetap" berlangganan.
 - Medical Diagnosis: Mengklasifikasi hasil biopsi sebagai "Tumor Jinak" atau "Ganas".
 - Sentiment Analysis: Mengkategorikan ulasan produk sebagai "Positif", "Negatif", atau "Netral".
 - Object Detection: Mengidentifikasi jenis kendaraan di jalan raya (Mobil, Motor, Bus, Truk).



Supervised Learning

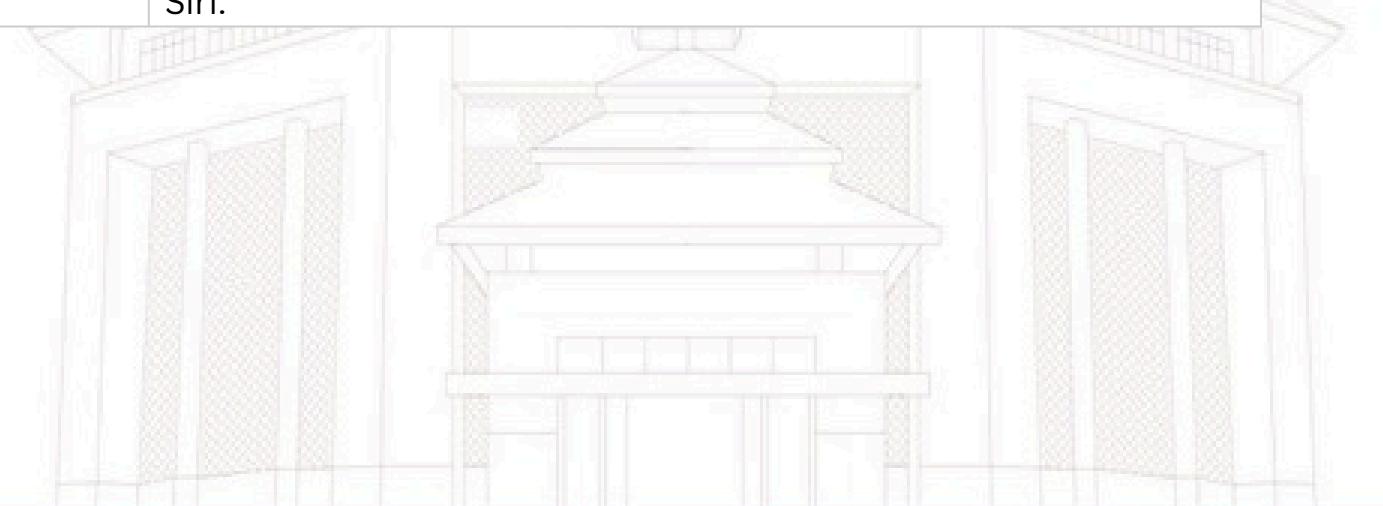


- Regresi (Regression): Mengestimasi nilai numerik kontinu atau kuantitatif.
 - Stock Price Prediction: Estimasi nilai harga penutupan saham di masa depan.
 - Demand Forecasting: Memprediksi jumlah stok barang yang harus disediakan (Unit).
 - Real Estate Valuation: Estimasi harga properti berdasarkan luas tanah dan lokasi.
 - Energy Consumption: Proyeksi penggunaan listrik bulanan suatu wilayah (Kilowatt).
 - Weather Analysis: Prediksi suhu udara atau curah hujan (Derajat/Milimeter).



Supervised Learning : Metode

Kategori	Algoritma	Penjelasan Teknis	Contoh Kasus Riil
Klasifikasi	Naive Bayes	Berbasis probabilitas menggunakan Teorema Bayes dengan asumsi independensi antar fitur.	Analisis sentimen ulasan produk atau filter spam email.
Klasifikasi	Logistic Regression	Memprediksi probabilitas variabel target biner menggunakan fungsi sigmoid.	Prediksi customer churn (pelanggan berhenti langganan).
Klasifikasi	k-Nearest Neighbors (KNN)	Klasifikasi berdasarkan mayoritas kelas pada k tetangga terdekat dalam ruang fitur.	Identifikasi segmen pelanggan untuk target pemasaran.
Klasifikasi	Support Vector Machine (SVM)	Mencari hyperplane optimal yang memisahkan kelas dengan margin maksimal.	Deteksi transaksi kartu kredit palsu (fraud) atau diagnosa medis.
Regresi	Linear & Polynomial	Memodelkan hubungan variabel melalui garis lurus atau kurva polinomial derajat tertentu.	Estimasi harga properti atau pemodelan tren harga saham.
Regresi	Ridge & Lasso Regression	Teknik regresi dengan regularisasi (L1/L2) untuk mencegah overfitting pada data kompleks.	Prediksi indikator ekonomi dengan jumlah variabel yang sangat banyak.
Hibrida	Decision Tree & Random Forest	Struktur hierarkis yang bisa digunakan baik untuk tugas Klasifikasi maupun Regresi.	Prediksi konsumsi energi bangunan atau klasifikasi citra medis.
Lanjut	Neural Networks	Terinspirasi dari otak manusia, mampu mempelajari pola non-linear yang sangat kompleks.	Pengenalan suara (speech recognition) dan asisten virtual seperti Siri.





Supervised Learning : Evaluasi

		Predicted Class 0	Predicted Class 1
Actual Class 0	20	5	
	3	15	

- Matrik Evaluasi Untuk Klasifikasi:
 - Accuracy: Persentase total prediksi benar.
 - Precision & Recall: Fokus pada ketepatan kelas positif dan kemampuan menemukan seluruh data positif.
 - F1-Score: Rata-rata harmonik antara Precision dan Recall.
 - Confusion Matrix: Tabel rincian hasil prediksi vs data aktual.
- Matrik Evaluasi Untuk Regresi:
 - MSE (Mean Squared Error): Rata-rata kuadrat kesalahan prediksi.
 - RMSE (Root Mean Squared Error): Akar kuadrat dari MSE.
 - R-Squared: Koefisien determinasi yang mengukur kecocokan model.



Unsupervised Learning



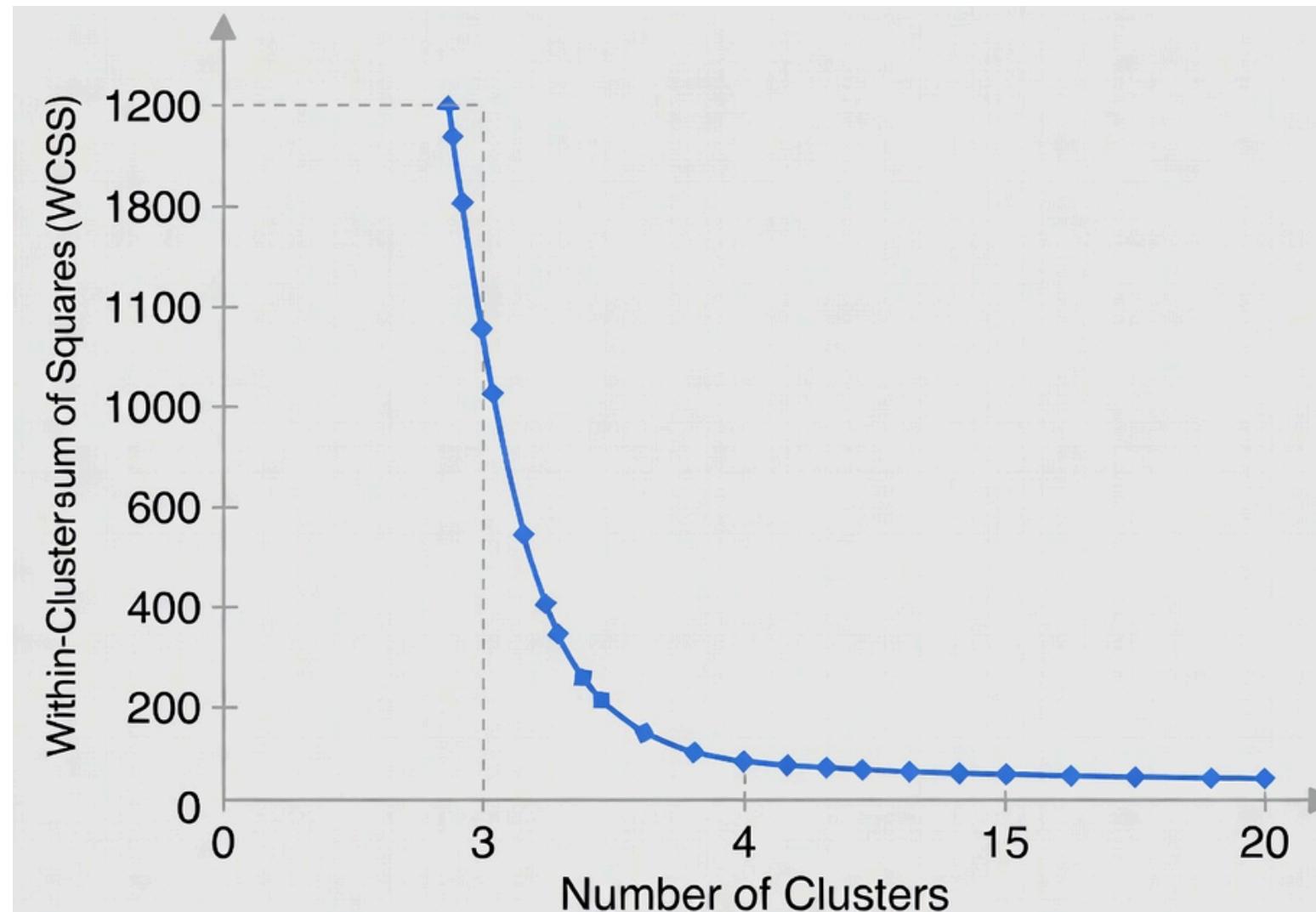
- Mekanisme Operasional: Algoritma dilatih menggunakan dataset yang tidak memiliki variabel target (Unlabeled Data) untuk mengeksplorasi struktur atau pola laten secara mandiri.
 - Clustering (Klasterisasi): Teknik pengelompokan data berdasarkan kemiripan fitur atau pola yang muncul secara alami dari dataset.
 - Segmentasi Pelanggan: Mengelompokkan pembeli berdasarkan perilaku belanja untuk strategi pemasaran yang personal.
 - Anomaly Detection (Deteksi Anomali): Identifikasi observasi langka atau unik yang menyimpang secara signifikan dari pola mayoritas data.
 - Deteksi Fraud: Mengidentifikasi transaksi kartu kredit yang mencurigakan atau tidak lazim.
 - Dimensionality Reduction (Reduksi Dimensi): Metodologi penyederhanaan variabel data yang kompleks tanpa menghilangkan informasi substansial guna efisiensi komputasi.
 - Visualisasi Data: Menyederhanakan data genetika atau citra satelit yang rumit agar mudah dipahami.



Unsupervised Learning : Metode & Evaluasi

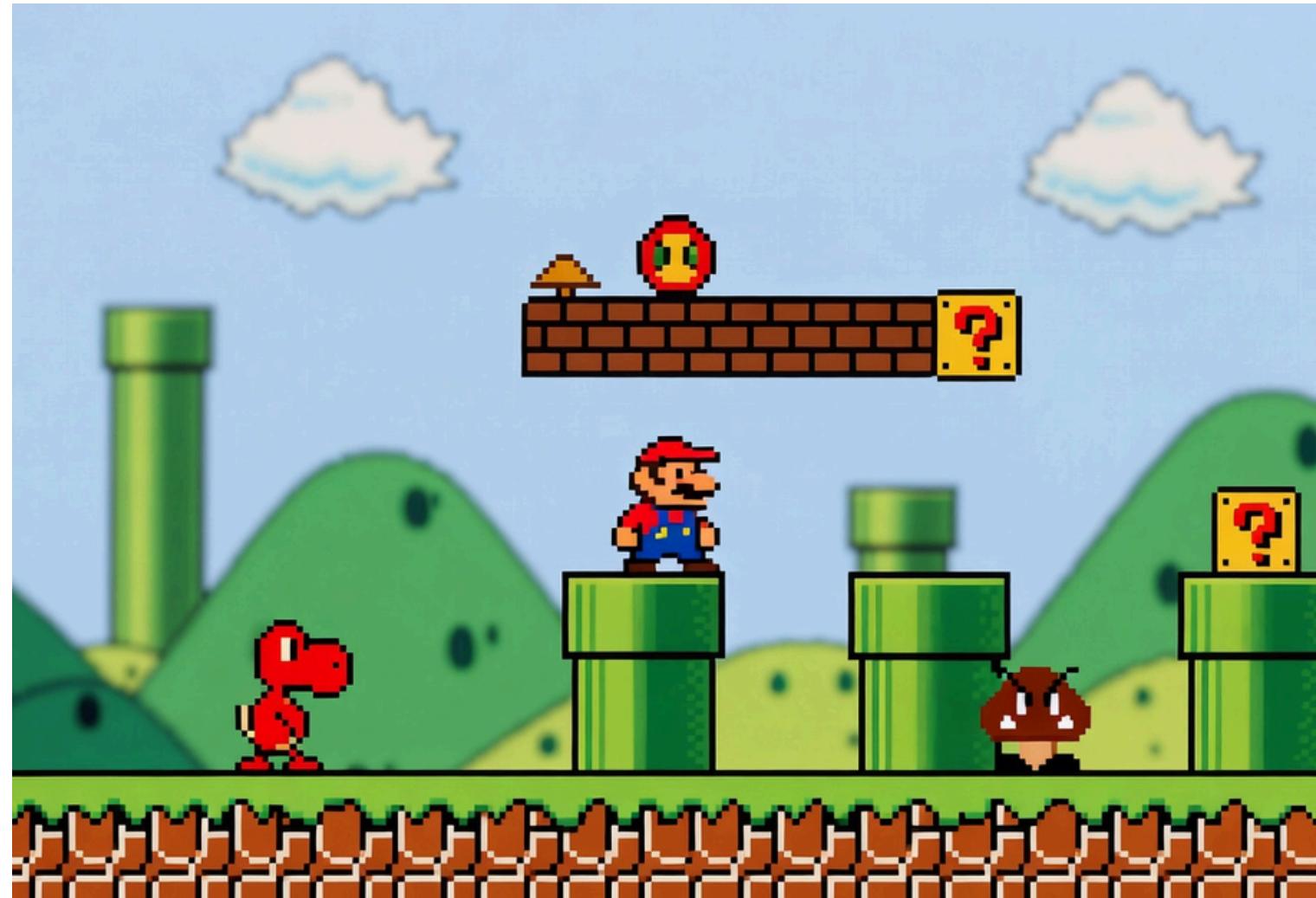
Kategori	Algoritma	Penjelasan Teknis	Contoh Kasus
Clustering	K-Means Clustering	Membagi data ke dalam k kelompok berdasarkan jarak ke titik pusat (centroid).	Segmentasi pelanggan berdasarkan perilaku belanja.
Clustering	Hierarchical Clustering	Membangun hierarki kelompok baik secara bottom-up (agglomerative) atau top-down.	Pemetaan hubungan evolusi antar spesies (Filogeni).
Clustering	DBSCAN	Mengelompokkan data berdasarkan kerapatan titik; sangat efektif untuk data dengan bentuk tidak beraturan.	Identifikasi area rawan kriminalitas di peta perkotaan.
Clustering	Gaussian Mixture Models	Mengasumsikan data berasal dari campuran beberapa distribusi probabilitas Gaussian.	Pemisahan suara pembicara dalam rekaman audio (speaker diarization).
Dim. Reduction	Principal Component Analysis (PCA)	Mentransformasi variabel berkorelasi menjadi sekumpulan variabel linear yang tidak berkorelasi.	Kompresi data citra atau penyederhanaan fitur survei yang sangat banyak.
Dim. Reduction	t-SNE / UMAP	Teknik visualisasi data dimensi tinggi ke dalam ruang 2D atau 3D dengan menjaga struktur lokal.	Visualisasi kemiripan ekspresi gen dalam data biomedis.
Association	Apriori Algorithm	Menemukan aturan asosiasi atau hubungan antar item dalam dataset besar.	Market Basket Analysis (misal: pembeli roti cenderung membeli selai).
Anomaly Detection	Isolation Forest	Mendeteksi anomali dengan cara mengisolasi titik data menggunakan struktur pohon.	Deteksi penyusupan pada jaringan komputer (intrusion detection).

Unsupervised Learning : Evaluasi



- Matrik Evaluasi Untuk Unsupervised:
 - Silhouette Coefficient: Mengukur seberapa mirip suatu objek dengan klaster sendiri dibandingkan dengan klaster lain (Rentang -1 hingga 1).
 - Davies-Bouldin Index: Menilai kualitas klaster berdasarkan rasio jarak dalam klaster dan jarak antar klaster (Semakin kecil semakin baik).
 - Calinski-Harabasz Index: Rasio antara dispersi antar-klaster dengan dispersi dalam-klaster.
 - Elbow Method (WCSS): Menentukan jumlah k optimal dengan melihat titik "siku" pada grafik penurunan Within-Cluster Sum of Squares.
 - Perplexity: Sering digunakan untuk mengevaluasi kualitas model topik (LDA) atau stabilitas visualisasi t-SNE.

Reinforcement Learning



- Mekanisme Inti: Pembelajaran mandiri di mana sebuah Agent belajar mengambil keputusan optimal melalui interaksi dinamis dengan Environment.
- Siklus Interaksi:
 - Action: Tindakan yang diambil oleh Agen.
 - State: Kondisi atau situasi Agen dalam lingkungan saat ini.
 - Reward/Penalty: Umpan balik numerik (hadiah atau hukuman) berdasarkan kualitas tindakan.
- Tujuan Strategis: Menemukan Policy (Kebijakan) yang mampu memaksimalkan akumulasi reward dalam jangka panjang.
- Aplikasi Utama: Navigasi kendaraan otonom, optimasi algoritma trading frekuensi tinggi, robotika industri, dan penguasaan permainan kompleks (AlphaGo).



Reinforcement Learning : Metode & Evaluasi

Komponen	Deskripsi	Metrik Evaluasi
Q-Learning	Algoritma berbasis nilai untuk mencari tindakan terbaik di setiap situasi.	Cumulative Reward: Total hadiah yang dikumpulkan.
SARSA	Algoritma yang memperbarui kebijakan berdasarkan tindakan nyata yang diambil.	Convergence Speed: Seberapa cepat agen menemukan rute optimal.
Deep Q-Network	Kombinasi RL dengan Neural Networks untuk menangani ruang data raksasa.	Success Rate: Persentase keberhasilan dalam menyelesaikan misi.
Policy Gradient	Fokus langsung pada optimasi strategi tanpa perlu menghitung nilai setiap aksi.	Average Reward per Episode: Rata-rata skor yang didapat dalam tiap sesi.





SEE YOU NEXT WEEK !

Ferdian Bangkit Wijaya, S.Stat., M.Si

NIP. 199005202024061001

ferdian.bangkit@untirta.ac.id

