

Penambangan Data (Data Mining)

PERTEMUAN 1



Materi sebelum UTS

▶	Pengertian dan Konsep Data Mining	01
▶	Eksplorasi Data	02
▶	Pra-Pemrosesan Data	03
▶	Metode-Metode Data Mining dan Text Mining	04
▶	Metode Regresi	05
▶	Klasifikasi dengan Regresi Logistik	06
▶	Klasifikasi dengan Naive Bayes	07
▶	UTS	08

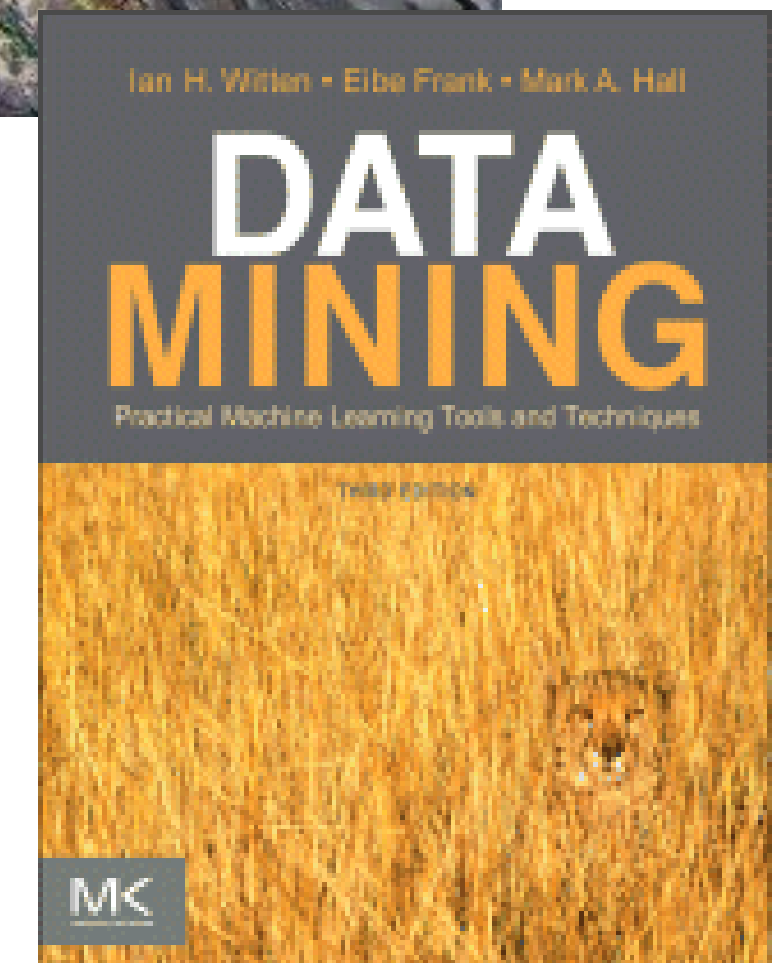
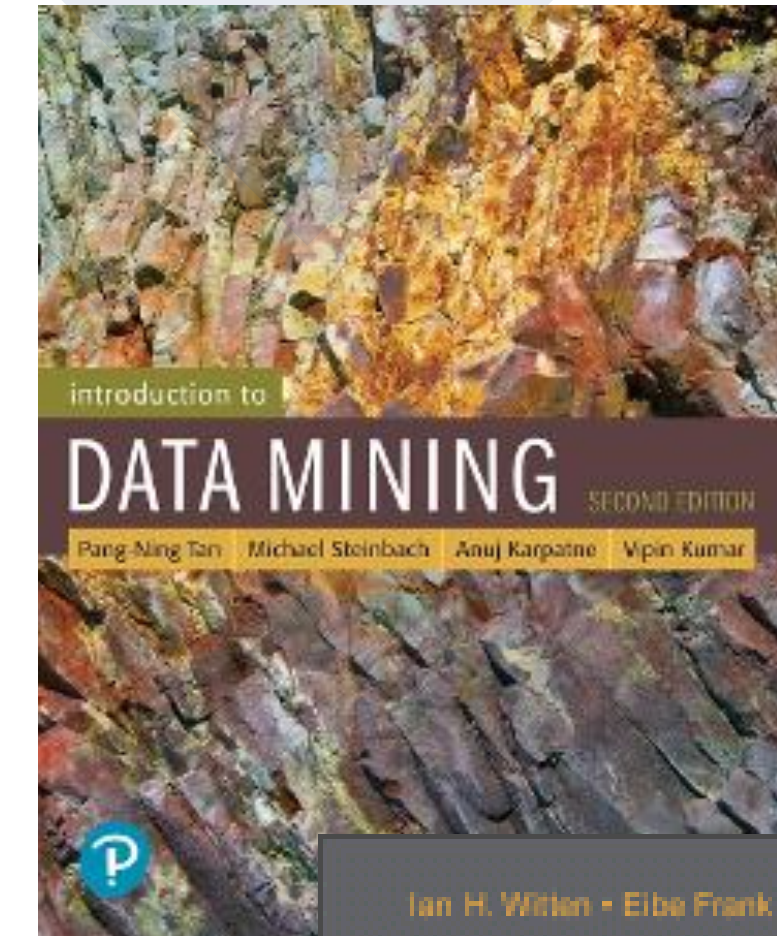
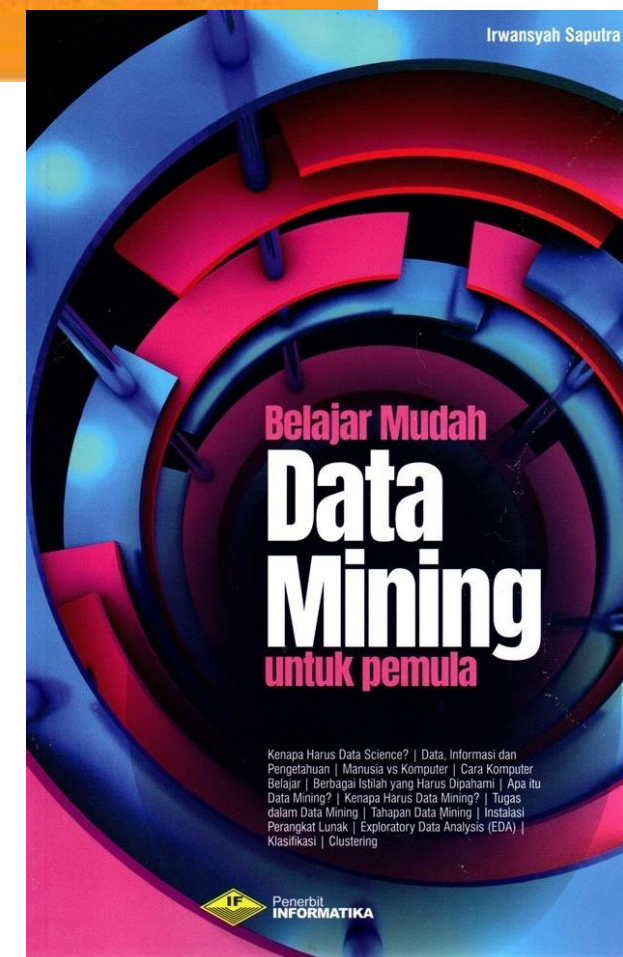
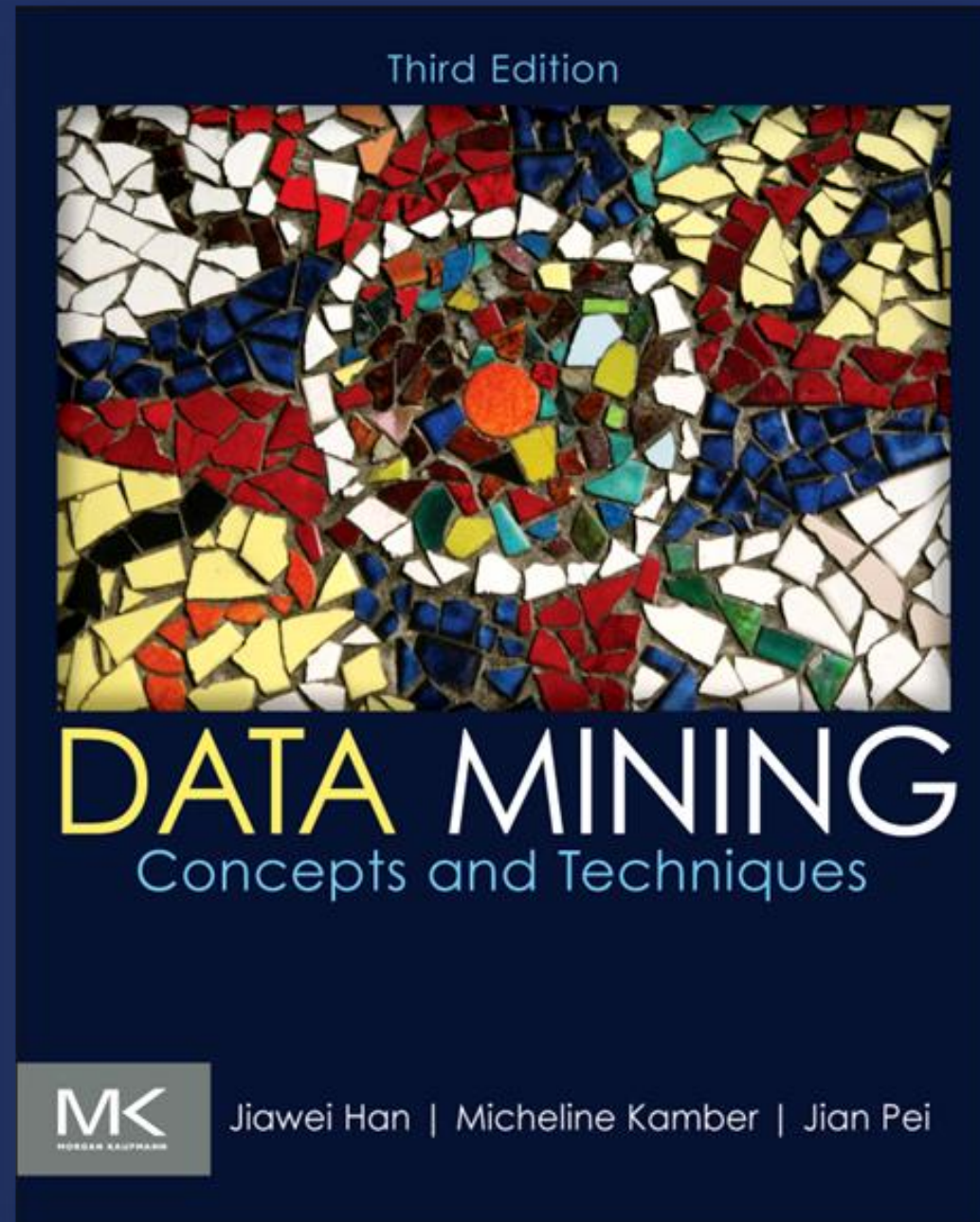


Materi setelah UTS

▶	Klasifikasi dengan Neural Network	09
▶	Klasifikasi dengan Decision Tree	10
▶	Klasifikasi dengan K-Nearest Neighbour (KNN)	11
▶	Klastering dengan K-Means	12
▶	Klastering dengan AHC	13
▶	Asosiasi dengan Apriori	14
▶	Tugas Proyek	15
▶	UAS	16



Buku Referensi



Mengapa Data Mining?

- **Data Over Load**
 - Web data, e-commerce, e-banking
 - Grocery stores
 - Bank/Credit Card Transaction
- **Teknologi komputer menjadi lebih murah dan powerful**
 - Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec
- **Tekanan kompetisi semakin kuat**
 - Dalam bisnis, dapat memberikan layanan yang lebih baik
contoh: Customer Relationship Management



Mengapa Data Mining?

- Data dikumpulkan dan disimpan dengan kecepatan tinggi (GB/hour)

Astronomi

- Sloan Digital Sky Survey
 - New Mexico, 2000
 - 140TB over 10 years
- Large Synoptic Survey Telescope
 - Chile, 2016
 - Will acquire 140TB every five days

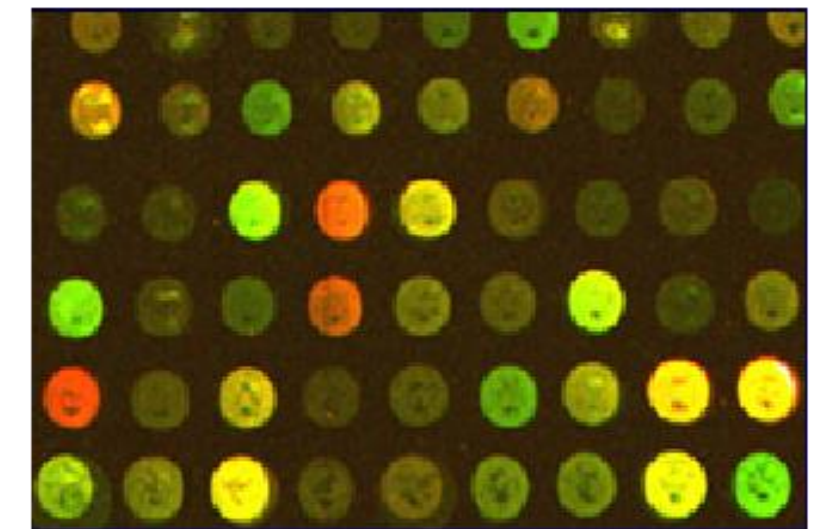
Biologi dan Kedokteran

- European Bioinformatics Institute (EBI)
 - 20PB of data (genomic data doubles in size each year)
 - A single sequenced human genome can be around 140GB in size

kilobyte (kB)	10^3
megabyte (MB)	10^6
gigabyte (GB)	10^9
terabyte (TB)	10^{12}
petabyte (PB)	10^{15}
exabyte (EB)	10^{18}
zettabyte (ZB)	10^{21}
yottabyte (YB)	10^{24}



Sky Survey Data



Gene Expression Data

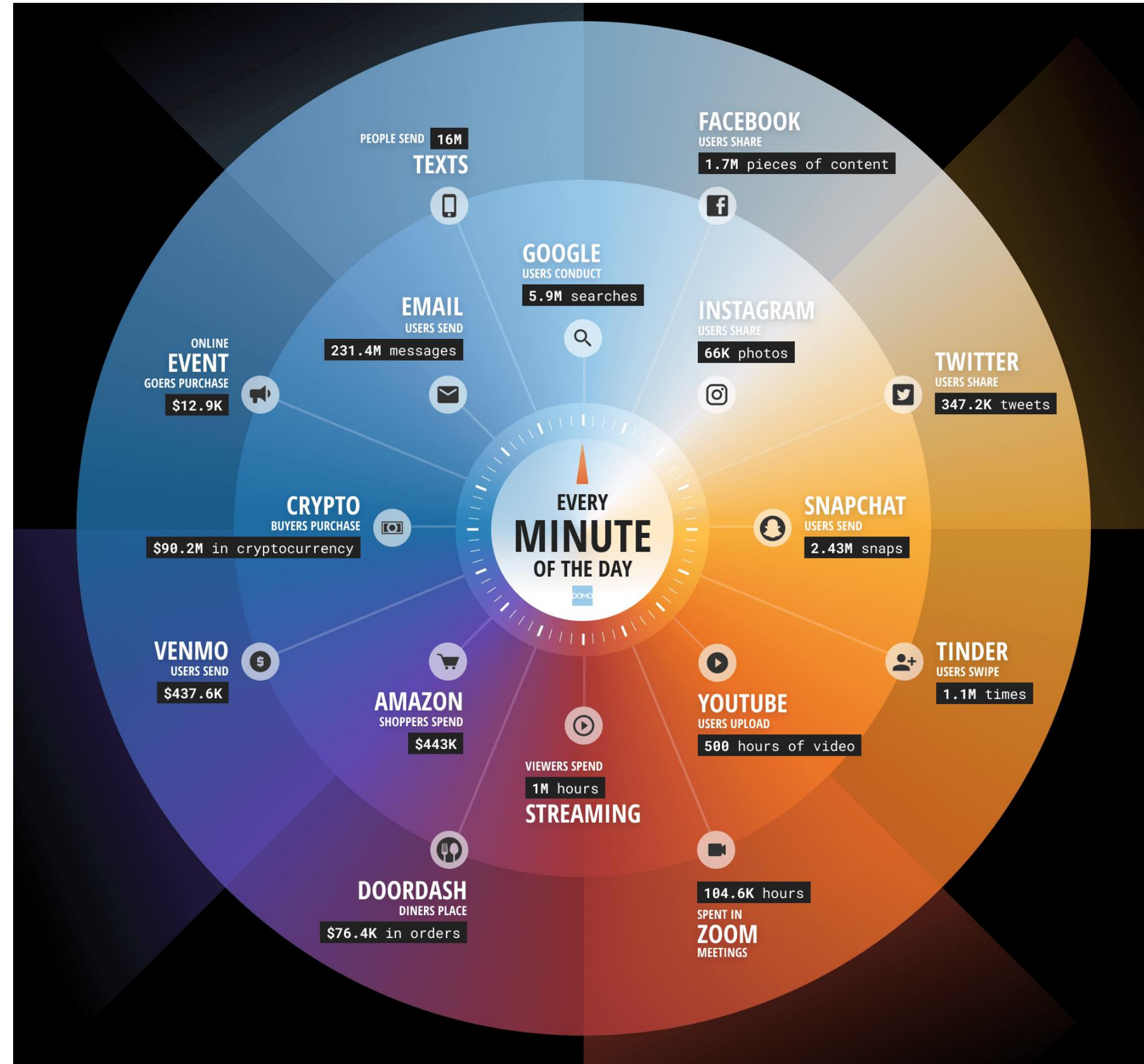
Kebanjiran Data tapi Miskin Pengetahuan

We are **drowning** in data,
but **starving** for knowledge!

(John Naisbitt, Megatrends, 1988)

- Pertumbuhan data sangat cepat dan bersifat eksponensial.

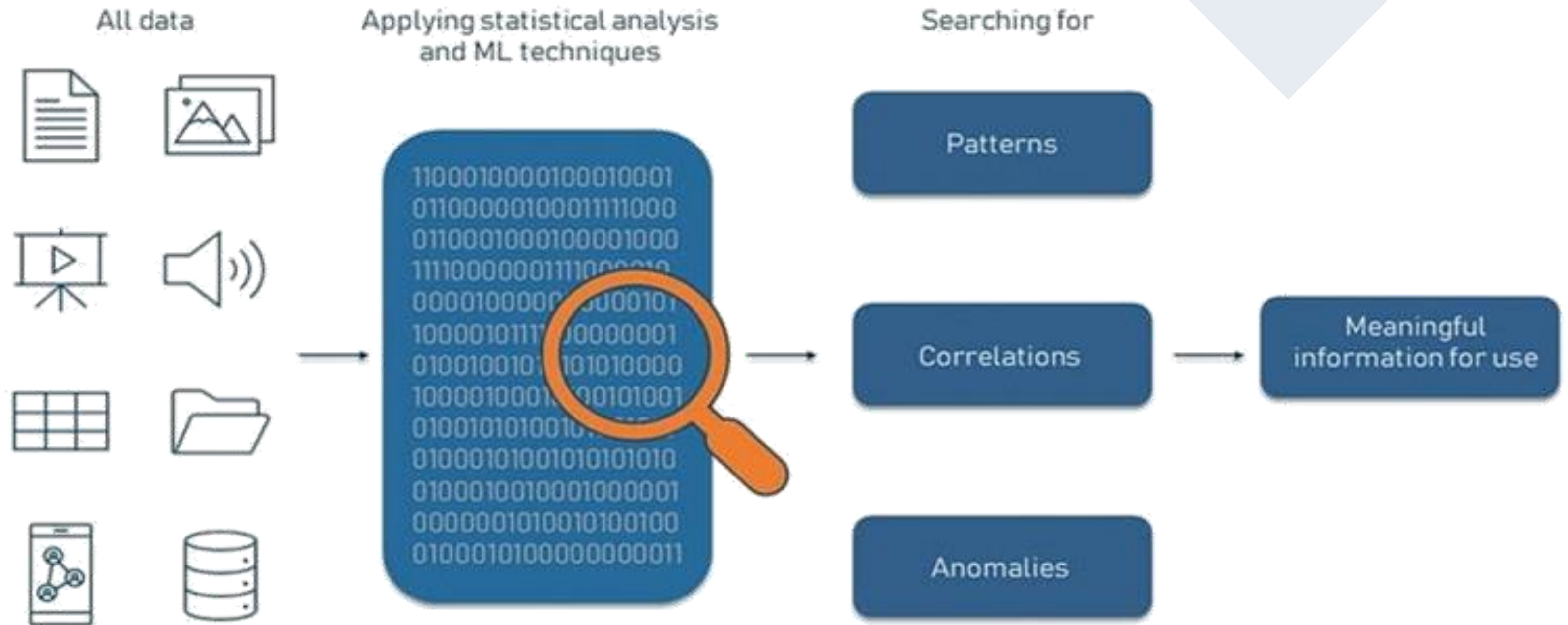
Sumber: <https://www.domo.com/data-never-sleeps#>



Mengubah Data Menjadi Pengetahuan

- Data harus kita olah menjadi **pengetahuan** supaya bisa **bermanfaat** bagi manusia
- Dengan **pengetahuan** tersebut, manusia dapat:
 - Melakukan **estimasi** dan **prediksi** apa yang terjadi di depan
 - Melakukan analisis tentang **asosiasi**, **korelasi** dan **pengelompokan** antar data dan atribut
 - Membantu **pengambilan keputusan** dan **pembuatan kebijakan**

Apa itu Data Mining?



Proses komputasi untuk menemukan pola, korelasi, dan anomali dalam kumpulan data besar dengan menerapkan berbagai teknik analisis statistik dan pembelajaran mesin (ML) untuk mengekstrak informasi dan wawasan yang bermakna dari data.

Apa itu Data Mining?

Data mining (pencarian pengetahuan dari data)

Disiplin ilmu yang mempelajari **metode** untuk **mengekstrak pengetahuan** atau **menemukan pola** dari suatu data yang besar

.

Ekstraksi dari **data** ke **pengetahuan**:

1. **Data**: **fakta yang terekam** dan tidak membawa arti
2. **Informasi**: Rekap, rangkuman, penjelasan dan **statistik dari data**
3. **Pengetahuan**: **pola, rumus**, aturan atau model yang muncul dari data

Nama lain Data Mining

- ☐ Knowledge Discovery in Database (KDD)
- ☐ Knowledge extraction
- ☐ Pattern analysis
- ☐ Information harvesting
- ☐ Business intelligence

Apa Manfaat Data Mining?

- Meningkatkan pengetahuan agar bisa membuat keputusan yang tepat.
- Perusahaan fokus ke informasi yang berharga di datawarehouse/databasenya sehingga dapat meningkatkan strategi bisnis.
- Meramalkan tren masa depan → perusahaan dapat mempersiapkan diri.

Contoh Data Mining

Midwest grocery chain menggunakan Data Mining untuk menganalisis pola pembelian: saat pria membeli popok di hari Kamis dan Sabtu, mereka juga membeli minuman.

Analisis lebih lanjut: pembeli ini belanja di hari kamis dan sabtu, tapi di hari kamis jumlah item lebih sedikit. Kesimpulan yang diambil: pembeli membeli minuman untuk dihabiskan saat weekend.

Tindak lanjut: menjual minuman dengan harga full di hari Kamis dan Sabtu. Mendekatkan posisi popok dan minuman.

Contoh Data Mining

Bank me-mining transaksi customer untuk mengidentifikasi customer yang kemungkinan besar tertarik terhadap produk baru.

Setelah teknik ini digunakan, terjadi peningkatan **20 kali lipat penurunan biaya** dibandingkan dengan cara biasa.

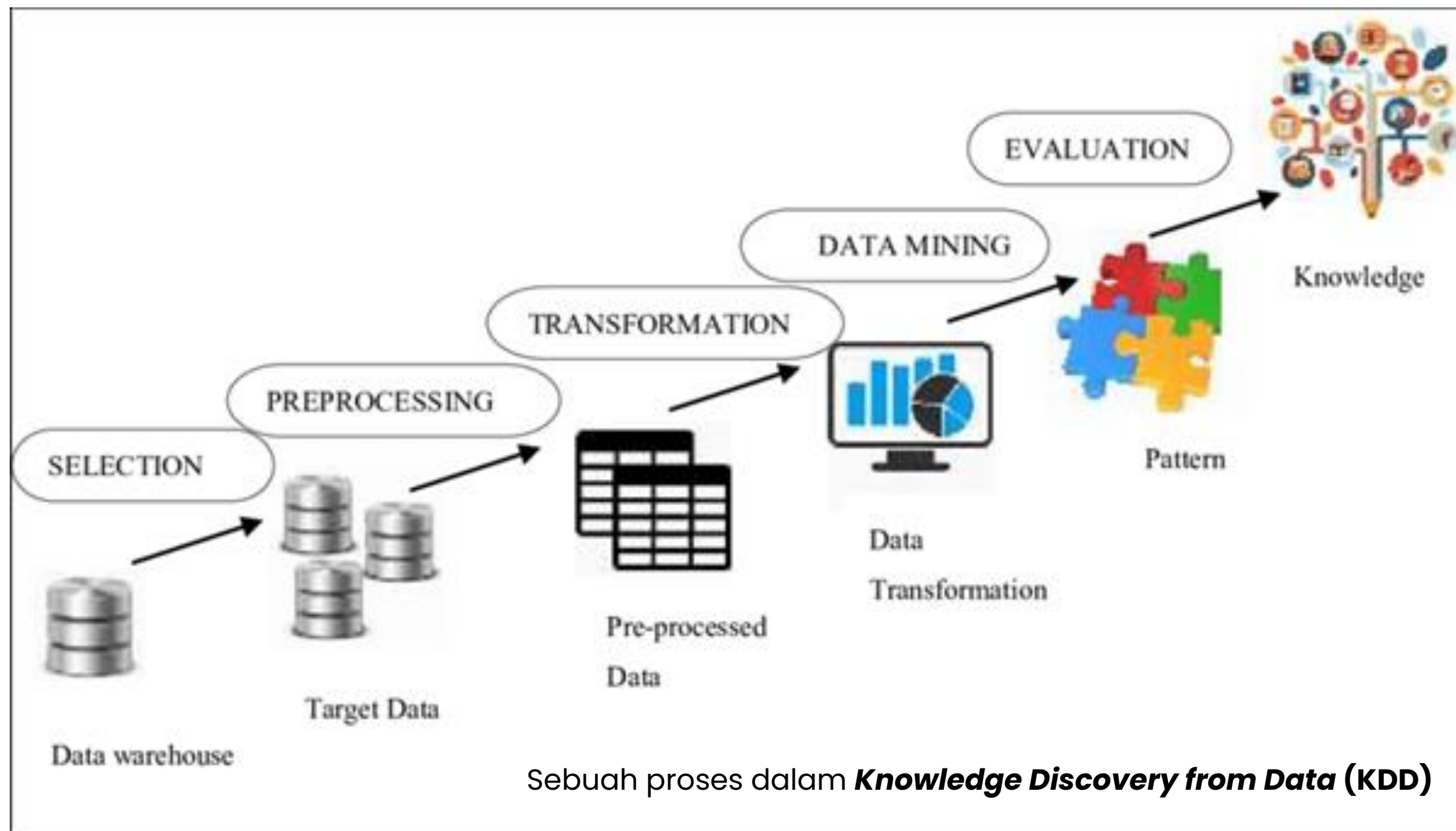


Contoh Data Mining

Perusahaan transportasi memining data customer untuk mengelompokkan customer yang memiliki nilai tinggi yang perlu diprioritaskan.

Proses Data Mining

Knowledge Discovery In Database (KDD) adalah proses yang dibantu oleh komputer untuk mencari dan meneliti sejumlah besar himpunan data dan mengekstrak informasi dan pengetahuan yang berguna.

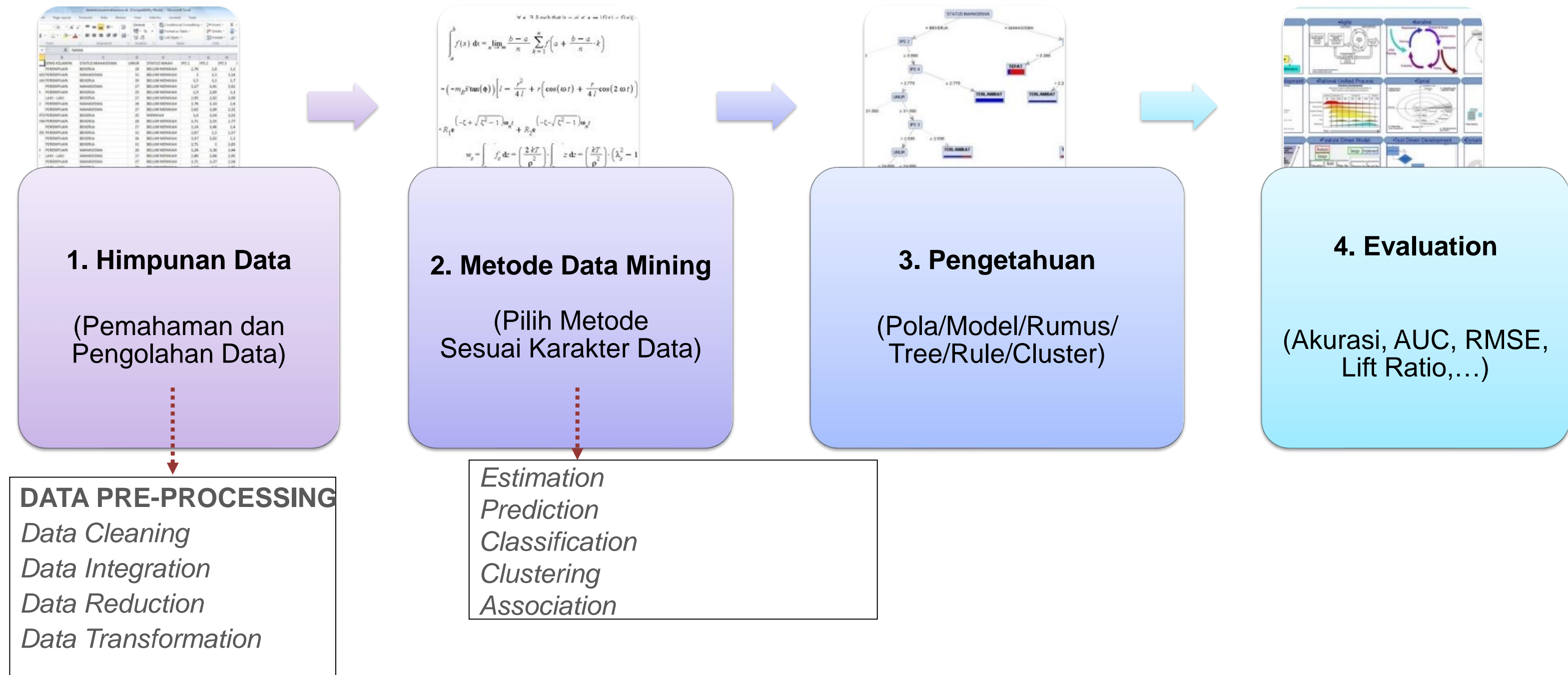


Proses dalam KDD

1. **Pembersihan data:** menghilangkan *noise* dan data yang tidak konsisten.
2. **Pengintegrasian data:** data digabungkan dari berbagai sumber.
3. **Seleksi data:** data yang relevan dengan proses analisis diambil dari basis data.
4. **Transformasi data:** data ditransformasikan atau digabungkan ke dalam bentuk yang sesuai untuk di-*mine* dengan cara dilakukan peringkasan atau operasi agregasi.
5. **Data mining:** merupakan proses yang penting dalam KDD dimana metode-metode cerdas diaplikasikan untuk mengekstrak pola-pola data.
6. **Evaluasi pola:** untuk mengidentifikasi pola-pola yang menarik yang merepresentasikan pengetahuan berdasarkan suatu ukuran kemenarikan.
7. **Presentasi pengetahuan:** merepresentasikan pengetahuan yang telah digali kepada pengguna.

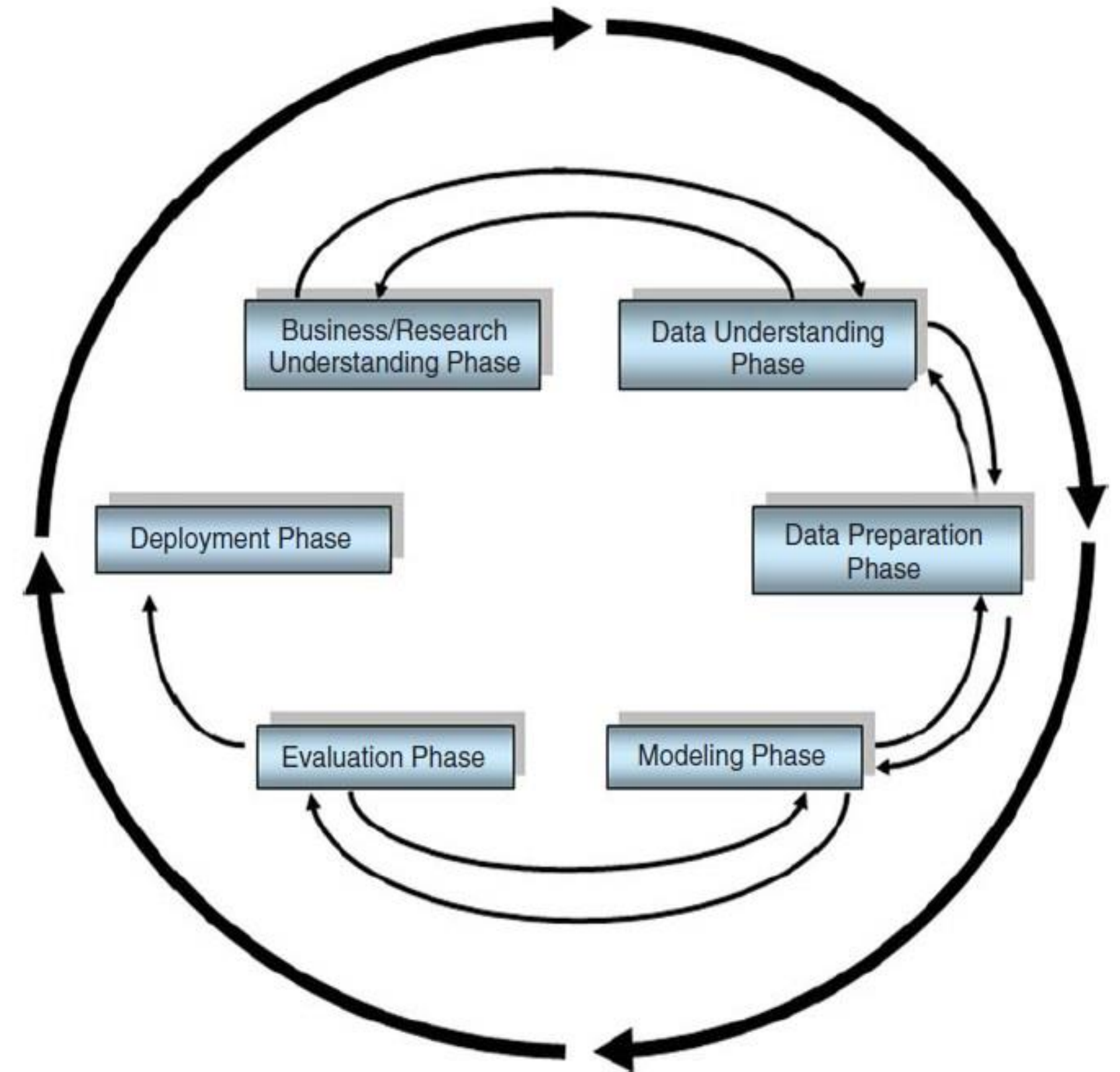
Proses dalam KDD

Typical View dari Machine Learning dan Statistic



Proses Data Mining berbasis Metodologi **CRISP-DM**

- **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*) adalah sebuah standar industri yang biasa digunakan ketika ingin melakukan *data mining* (analisis data).
- Terdapat 6 tahapan yang bisa dilihat melalui ilustrasi di samping.
- Setiap tahapan (sebelum dan setelahnya) bisa saling berinteraksi (maju mundur) untuk mencapai tujuan utama.



Ilustrasi CRISP-DM (La Rose. 2015. Data Mining and Predictive Analytics. Willey)

Tahapan CRISP-DM

1. **Business/Research Understanding**

- Mendefinisikan tujuan *project* dan *requirements* apa saja yang diperlukan untuk sebuah unit bisnis/riset secara keseluruhan.
- Menerjemahkan tujuan unit bisnis dan batasan-batasan yang ada ke formulasi masalah *data mining*.
- Menyiapkan strategi untuk mencapai tujuan ini.

2. **Data Understanding**

- Mengumpulkan data awal yang diperlukan
- Menggunakan teknik EDA (*Exploratory Data Analysis*) untuk menginspeksi dan mengidentifikasi data.

3. **Data Preparation**

- Membersihkan data, melakukan transformasi data, memilih fitur (*feature selection*) yang ingin dianalisis lebih lanjut

4. **Modeling Phase**

- Memilih dan membuat model yang sesuai dengan permasalahan.
- Terkadang bisa kembali ke tahap *data preparation* untuk memastikan data yang diperlukan sesuai dengan kebutuhan modelnya.

Tahapan CRISP-DM (2)

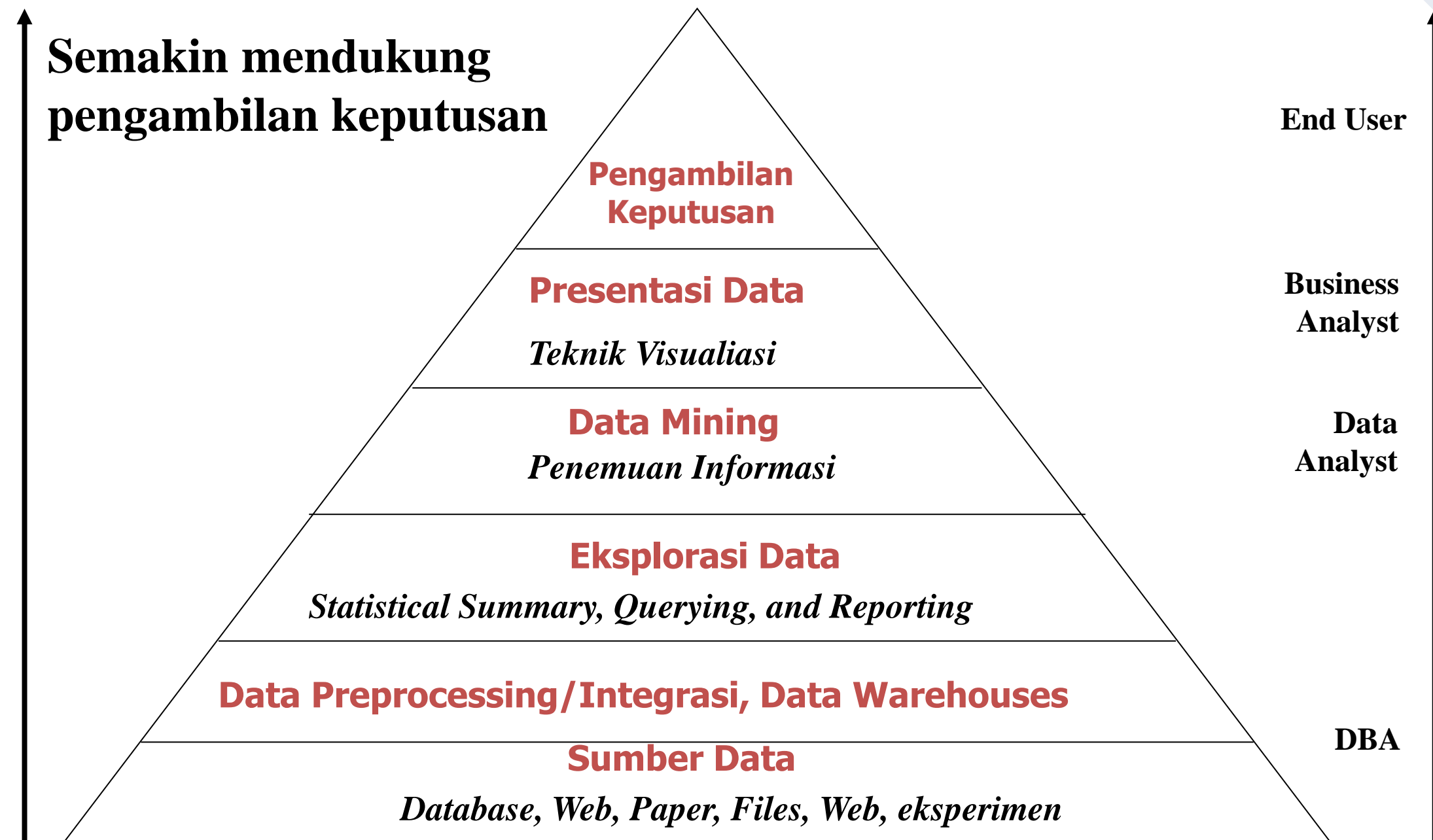
5. Evaluation Phase

- Pada tahapan ini kita mengevaluasi performa dari beberapa model yang sudah dibuat dan memilih model sebelum nanti kita *deploy* (terapkan) di lapangan.
- Pastikan bahwa model yang terpilih berhasil menjawab pertanyaan di tahap pertama (*business understanding*).
- Catat beberapa aspek jika seandainya ada hal-hal yang memang belum bisa dijawab melalui model terpilih.

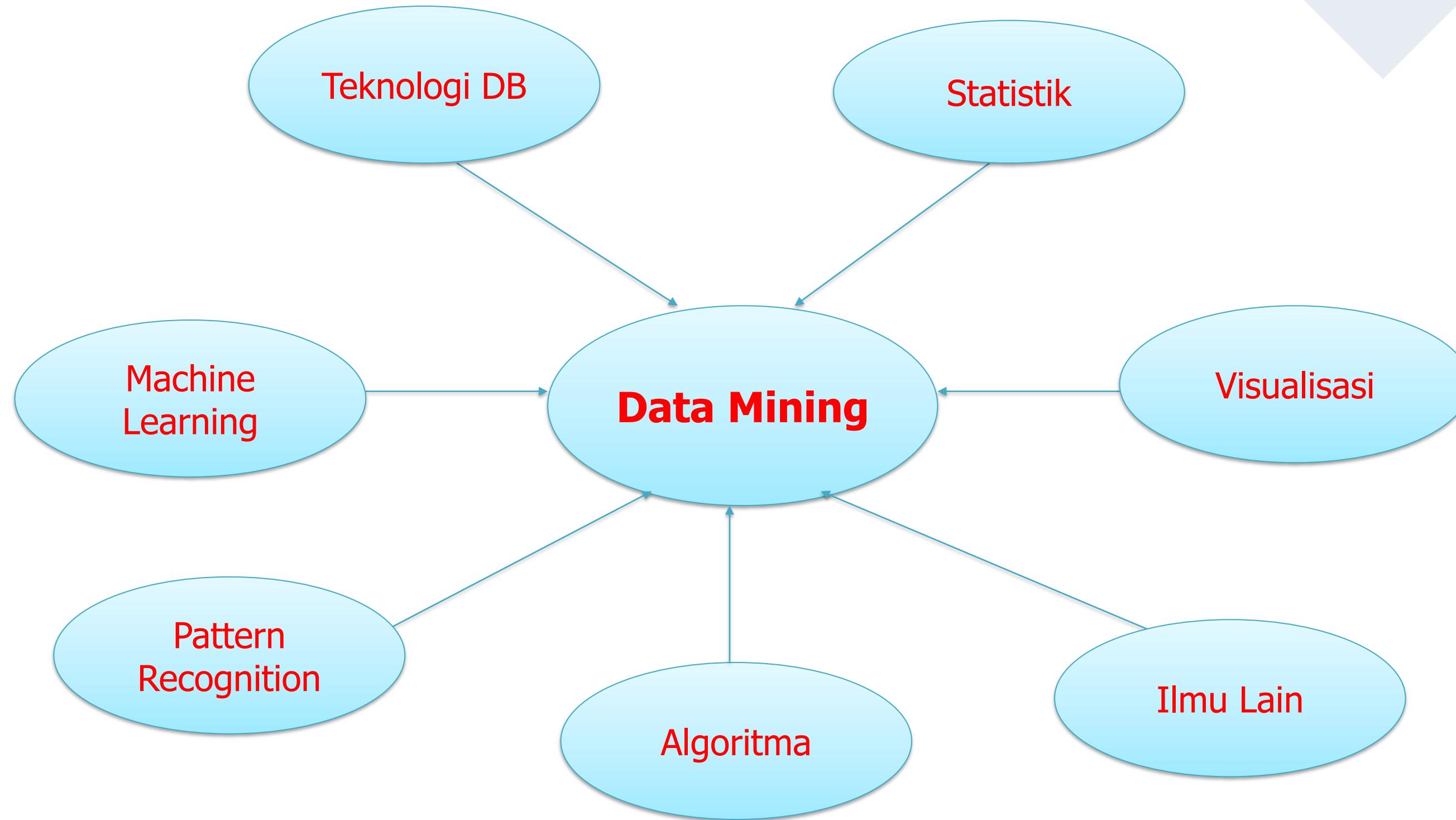
6. Deployment Phase

- Model yang sudah dipilih harus bisa diterapkan/dipakai untuk menjawab tujuan awal.
- Contoh *deployment* yang paling sederhana adalah membuat *report* (laporan) tentang penerapan model di lapangan.
- Contoh *deployment* lainnya adalah penerapan *data mining* secara parallel (bersamaan) di departemen lain.
- Jika target pengguna adalah konsumen, maka konsumen harus bisa menggunakan model kita.

Data Mining dan Business Intelligence



Data Mining : Multi Displin Ilmu



Mengapa tidak analisis data biasa?

- Jumlah data yang sangat besar
 - Algoritma harus scalable untuk menangani data yang sangat besar (tera)
- Dimensi yang sangat besar: ribuan field
- Data Kompleks
 - Aliran data dan sensor
 - Data terstruktur, graph, social network, multi-linked data
 - Database dari berbagai sumber, database lama
 - Spasial (peta), multimedia, text, web
 - Software Simulator

Data Mining dari berbagai sudut pandang

- **Data**

- Relational, datawarehouse, web, transaksional, stream, OO, spasial, text, multimedia

- **Pengetahuan yang akan ditambah**

- Karakteristik, diskriminasi, asosiasi, klasifikasi, clustering, trend, outlier

- **Teknik**

- Database, OLAP, machine learning, statistik, visualiasi

- **Penerapan**

- Retail, telekomunikasi, banking, analisis kejahatan, bio-data mining, saham, text mining, web mining

Permasalahan Pada DM

- **Metodologi**

- Mining beragam pengetahuan dari beragam sumber data
- Kinerja: efesiensi, efektivitas dan skalabilitas
- Evaluasi pola
- Background knowledge
- Noise (gangguan) dan data yang tidak lengkap
- Distributed dan paralel method.
- knowledge fusion (penggabungan)

Permasalahan Pada DM

- **Interaksi pengguna**

- Data mining query languages dan ad-hoc mining
- Visualisasi
- Interactive mining

- **Aplikasi**

- Domain spesifik
- Perlindungan data

Strategi Data Mining

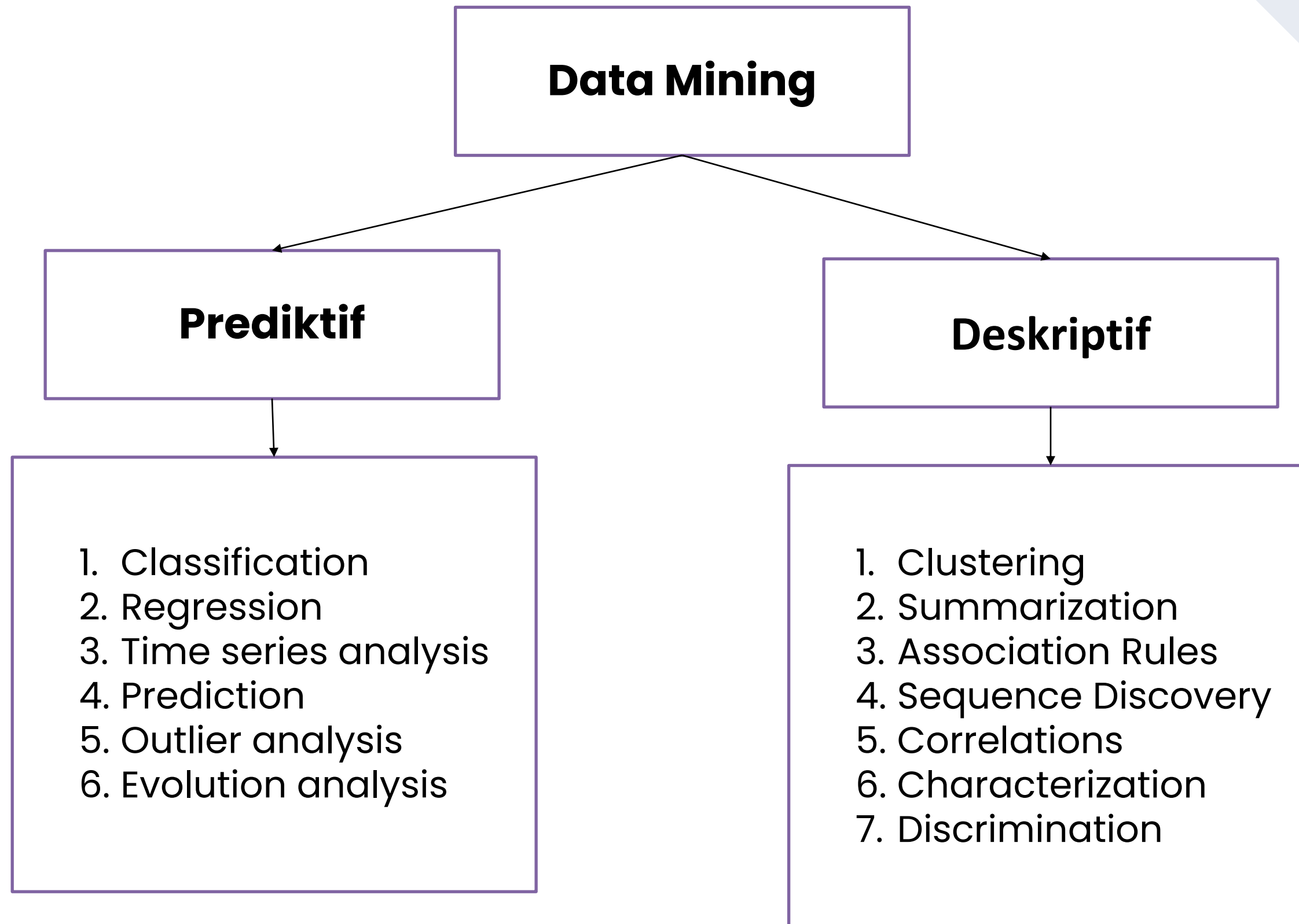
❑ Metode Prediksi

Menggunakan beberapa variabel (atribut) untuk memprediksi nilai yang tidak diketahui atau nilai yang akan datang dari variabel (atribut) lain.

❑ Metode Deskripsi

Menemukan pola-pola (korelasi, *trend*, *cluster*, *trayektori*, dan anomali) yang meringkas hubungan dalam data.

Strategi Data Mining

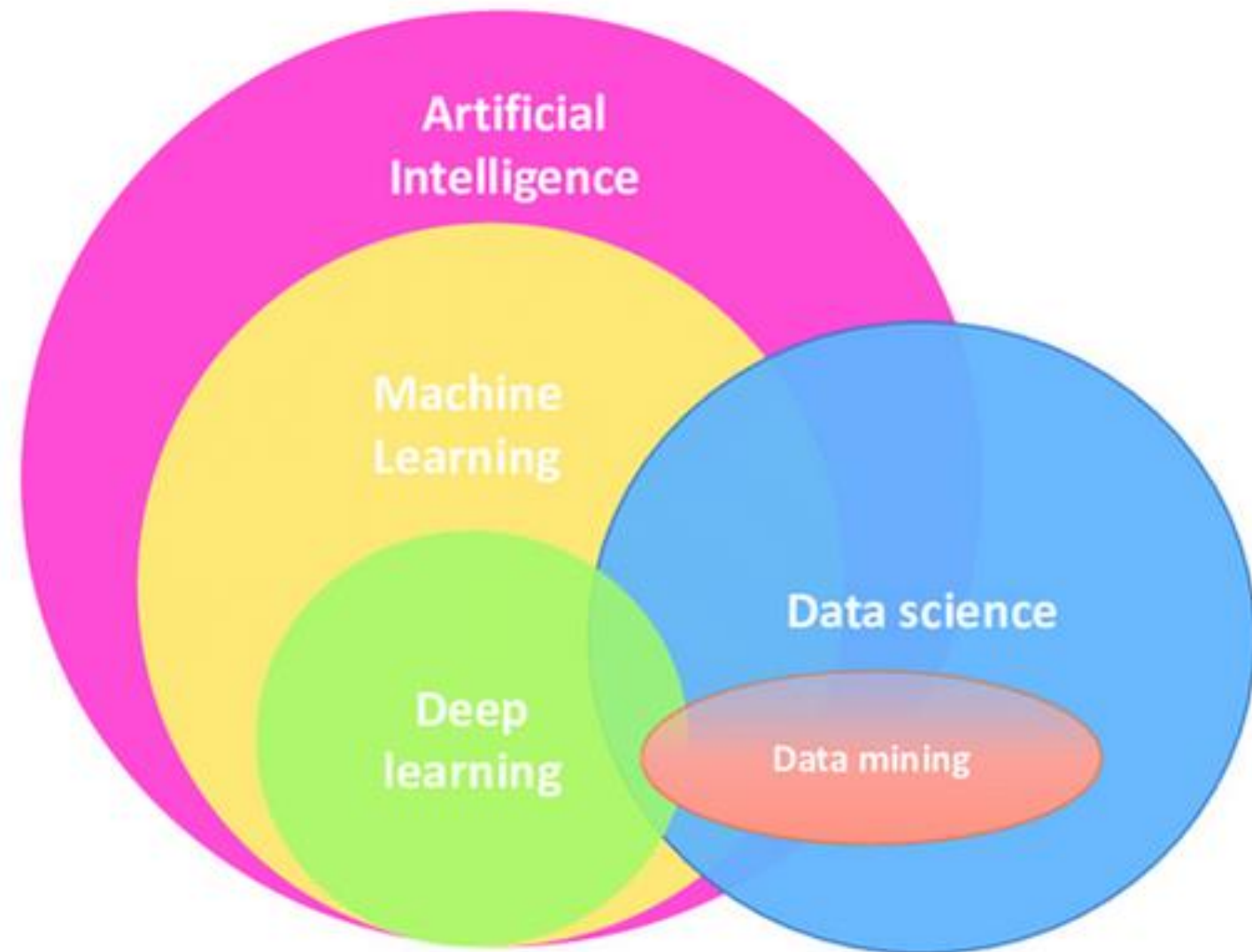


Data Mining vs Bukan Data Mining

Bukan Data Mining	Data Mining
✓ Pencarian informasi tertentu di Internet.	✓ Pengelompokkan informasi yang mirip dalam konteks tertentu pada hasil pencarian.
✓ Pencarian data medis untuk menganalisis catatan penyakit pasien.	✓ Peneliti medis mencari cara pengelompokkan data penyakit pasien berdasarkan data diagnosis, umur dan alamat
✓ Pembuatan laporan tahunan penjualan perusahaan	✓ Pemanfaatan data penjualan perusahaan untuk mendapatkan pola prediksi stok yang sebaiknya disediakan pada tahun berikutnya.

Sumber : Noviandi, Data Mining Pertemuan 1 [Powerpoint Slides]

Data Science vs Machine Learning vs AI vs Deep Learning vs Data Mining



Data Science :

Mengekstraksi Wawasan dari Data: Ilmu Data melibatkan penggalian wawasan dan pengetahuan yang dapat ditindaklanjuti dari sejumlah besar data. Ini mencakup pendekatan multidisiplin, menggabungkan analisis statistik, visualisasi data, pembelajaran mesin, dan keahlian domain. Ilmuwan data menggunakan teknik untuk membersihkan, mengatur, dan memproses data, mengungkap pola dan tren. Mereka menggunakan wawasan ini untuk membuat keputusan berdasarkan data dan memecahkan masalah kompleks di berbagai industri.

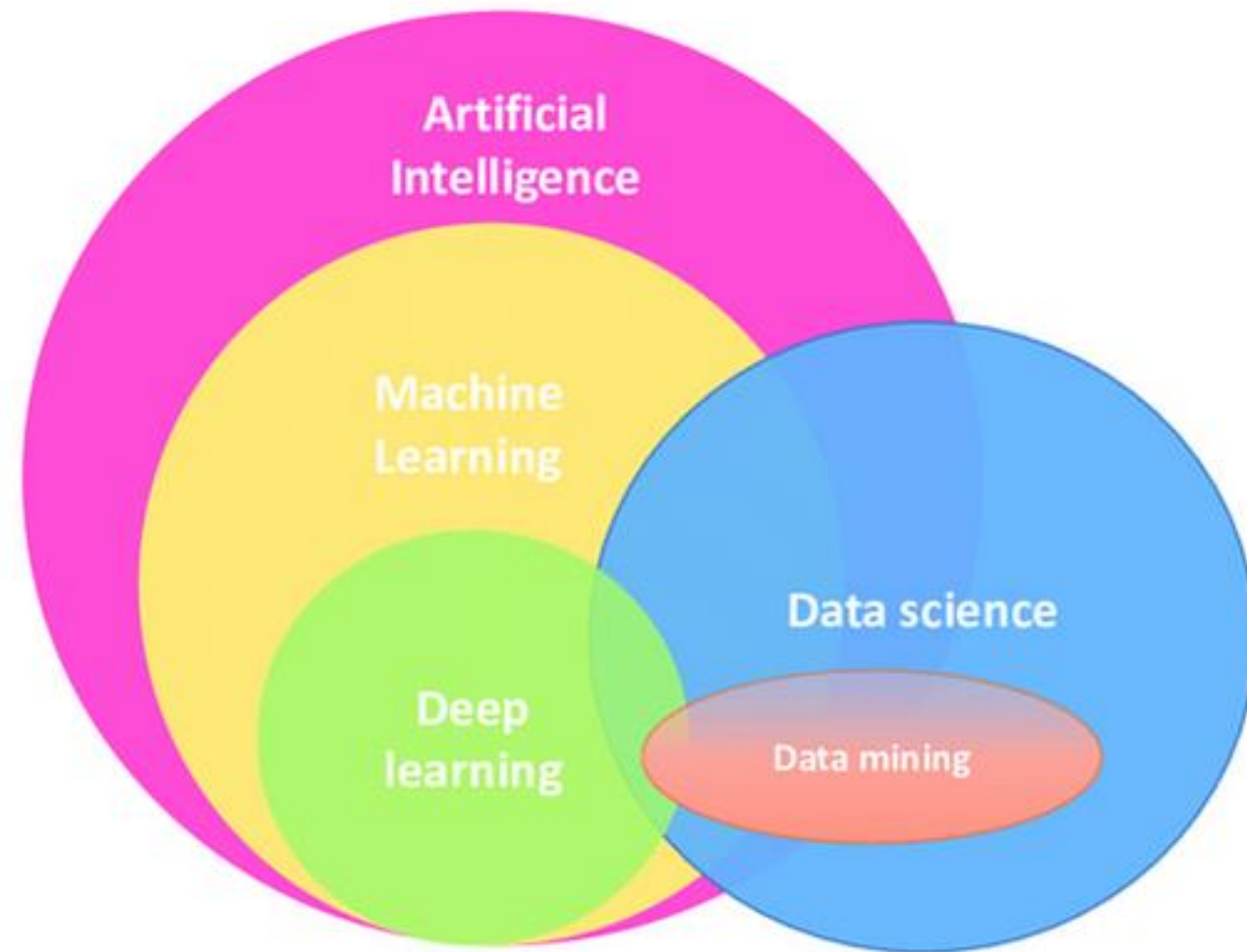
misalnya, sistem rekomendasi yang digunakan untuk memberikan saran yang dipersonalisasi kepada pelanggan berdasarkan riwayat pencarian mereka. Jika, katakanlah, satu pelanggan menelusuri joran dan umpan, sedangkan pelanggan lainnya mencari tali pancing selain produk lainnya, ada kemungkinan besar bahwa pelanggan pertama juga akan tertarik untuk membeli tali pancing.

Data Mining :

Menemukan Pola dalam Data: Data Mining berfokus pada penggalian pola dan pengetahuan yang bermakna dari kumpulan data yang besar. Ini melibatkan analisis data dari perspektif yang berbeda, mengungkap korelasi, asosiasi, atau anomali yang sebelumnya tidak diketahui. Algoritme penambangan data membantu mengidentifikasi pola, tren, dan ketergantungan yang dapat memberikan wawasan berharga untuk tujuan bisnis dan penelitian. Penambangan data sering kali berfungsi sebagai komponen penting dalam bidang ilmu data yang lebih luas.

contoh persediaan ikan, penambangan data adalah tentang mempelajari data 2 tahun terakhir untuk menemukan korelasi antara jumlah penjualan alat pancing sebelum dan selama musim penangkapan ikan di toko-toko yang berlokasi di negara bagian yang berbeda.

Data Science vs Machine Learning vs AI vs Deep Learning vs Data Mining



https://miro.medium.com/v2/resize:fit:786/format:webp/1*6sQHtPMj9Ad0W7ODh8yZTA.png

Machine Learning :

Belajar dari Data: Melatih mesin berdasarkan data historis sehingga mesin tersebut dapat memproses masukan baru berdasarkan pola yang dipelajari tanpa pemrograman eksplisit, artinya tanpa instruksi tertulis secara manual agar sistem dapat melakukan suatu tindakan. Jika bukan karena pembelajaran mesin, mesin rekomendasi yang telah kami sebutkan di atas tidak akan dapat dijangkau karena sulit bagi manusia untuk memproses jutaan kueri penelusuran, suka, dan ulasan untuk menemukan pelanggan mana yang biasanya membeli joran dengan umpan dan mana membeli pancing di atas itu.

Deep Learning :

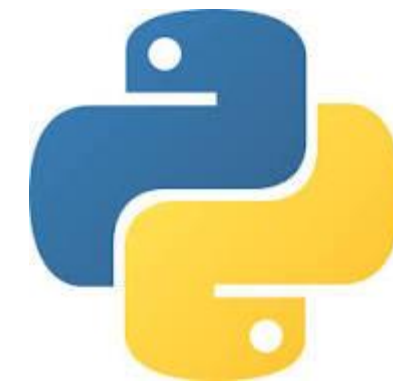
Mensimulasikan Jaringan Syaraf Manusia: cabang pembelajaran mesin paling populer yang menggunakan algoritme kompleks jaringan saraf dalam yang terinspirasi oleh cara kerja otak manusia. Model DL dapat memperoleh hasil yang akurat dari sejumlah besar data masukan tanpa perlu diberi tahu karakteristik data mana yang harus dilihat. Bayangkan Anda perlu menentukan pancing mana yang menghasilkan ulasan online positif di situs web Anda dan mana yang menyebabkan ulasan negatif. Dalam hal ini, jaringan saraf dalam dapat mengekstraksi karakteristik yang bermakna dari ulasan dan melakukan analisis sentimen.

AI:

Meniru Kecerdasan Manusia: Kecerdasan Buatan (AI) mengacu pada pengembangan sistem komputer yang mampu melakukan tugas-tugas yang biasanya membutuhkan kecerdasan manusia. Produk data apa pun di kehidupan nyata dapat disebut AI. Untuk membangun produk AI, Anda perlu menggunakan data mining, pembelajaran mesin, dan terkadang pembelajaran mendalam.

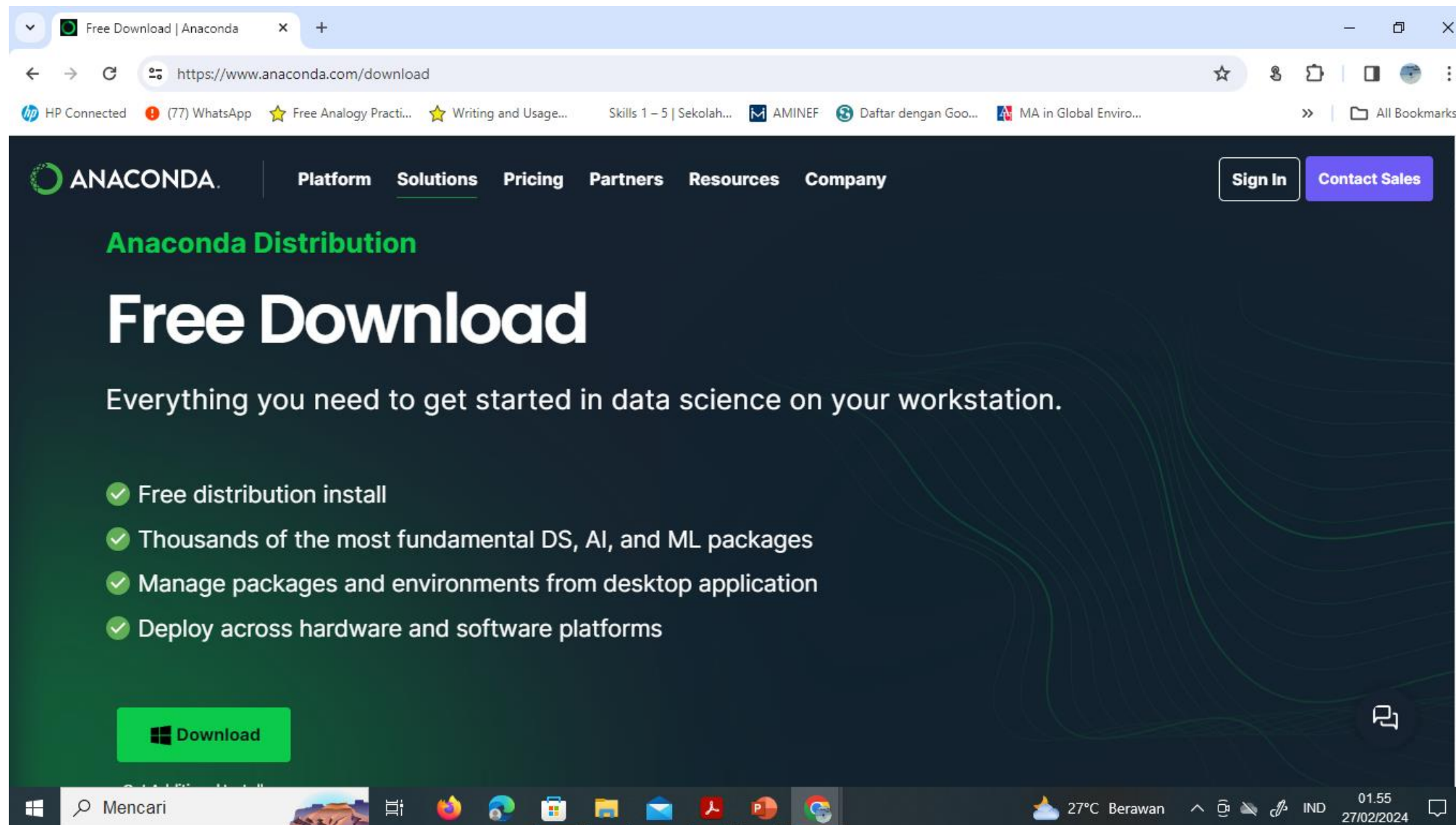
contoh yang terinspirasi dari kegiatan memancing. Anda ingin membeli joran pancing model tertentu tetapi hanya memiliki gambarnya dan tidak mengetahui nama mereknya. Sistem AI adalah produk perangkat lunak yang dapat memeriksa gambar Anda dan memberikan saran mengenai nama produk dan toko tempat Anda dapat membelinya.

Tools untuk Data Mining



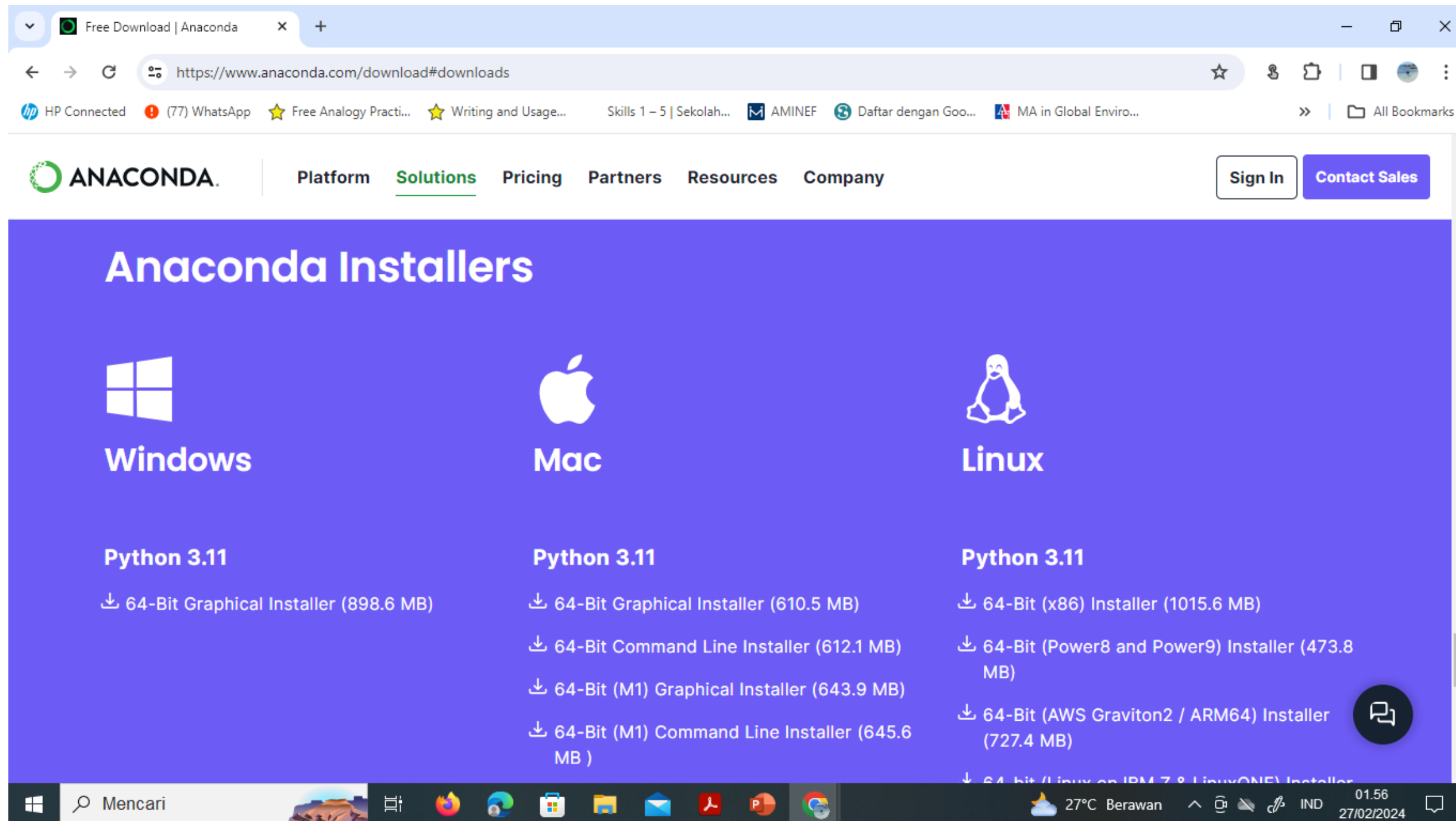
Cara Install Python dan Jupiter

1. Langkah pertama download anaconda.exe pada <https://www.anaconda.com/download>. Anda akan menemukan situs web seperti ini:



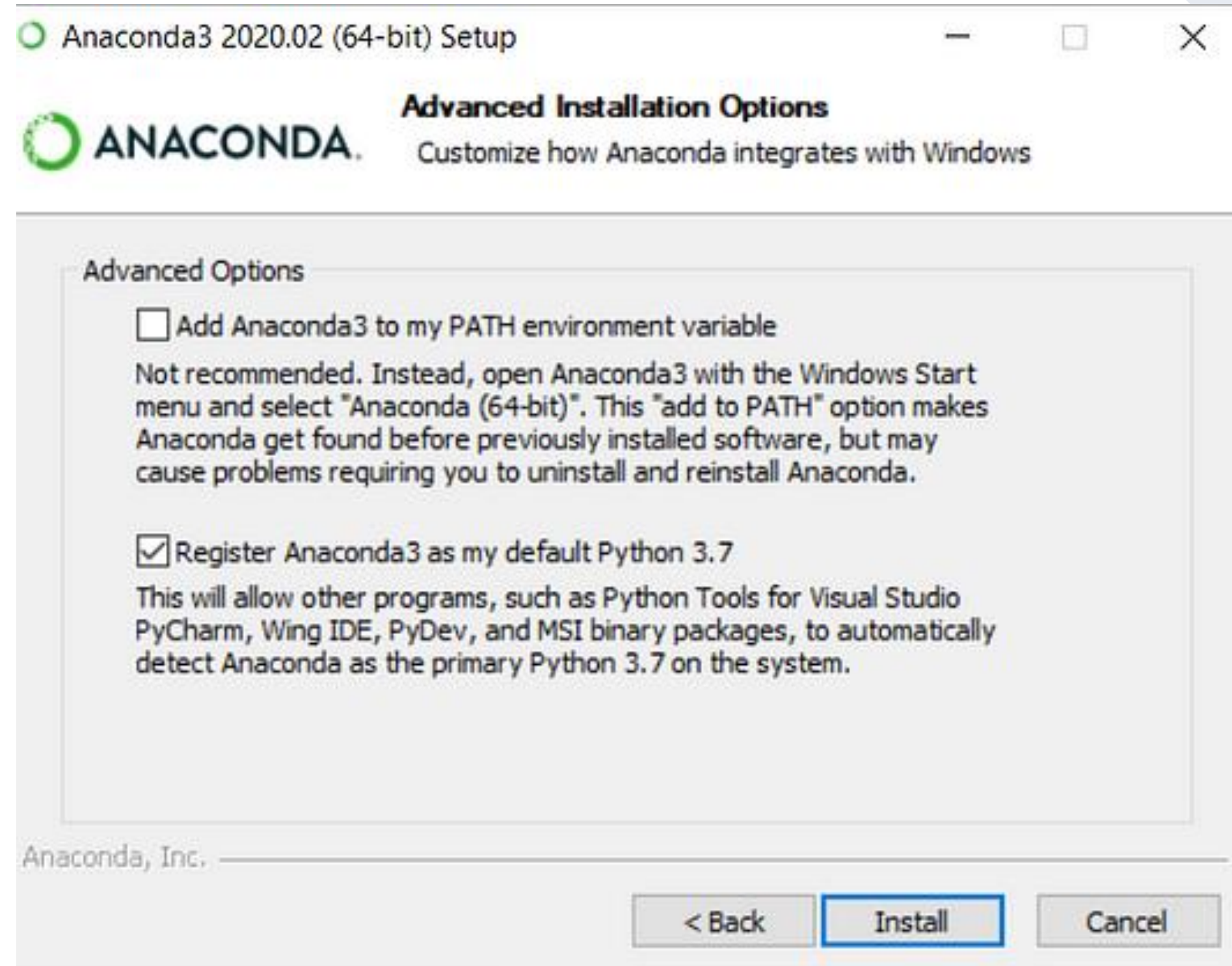
Cara Install Python dan Jupiter

2. lalu klik unduh. maka Anda terlihat seperti ini:



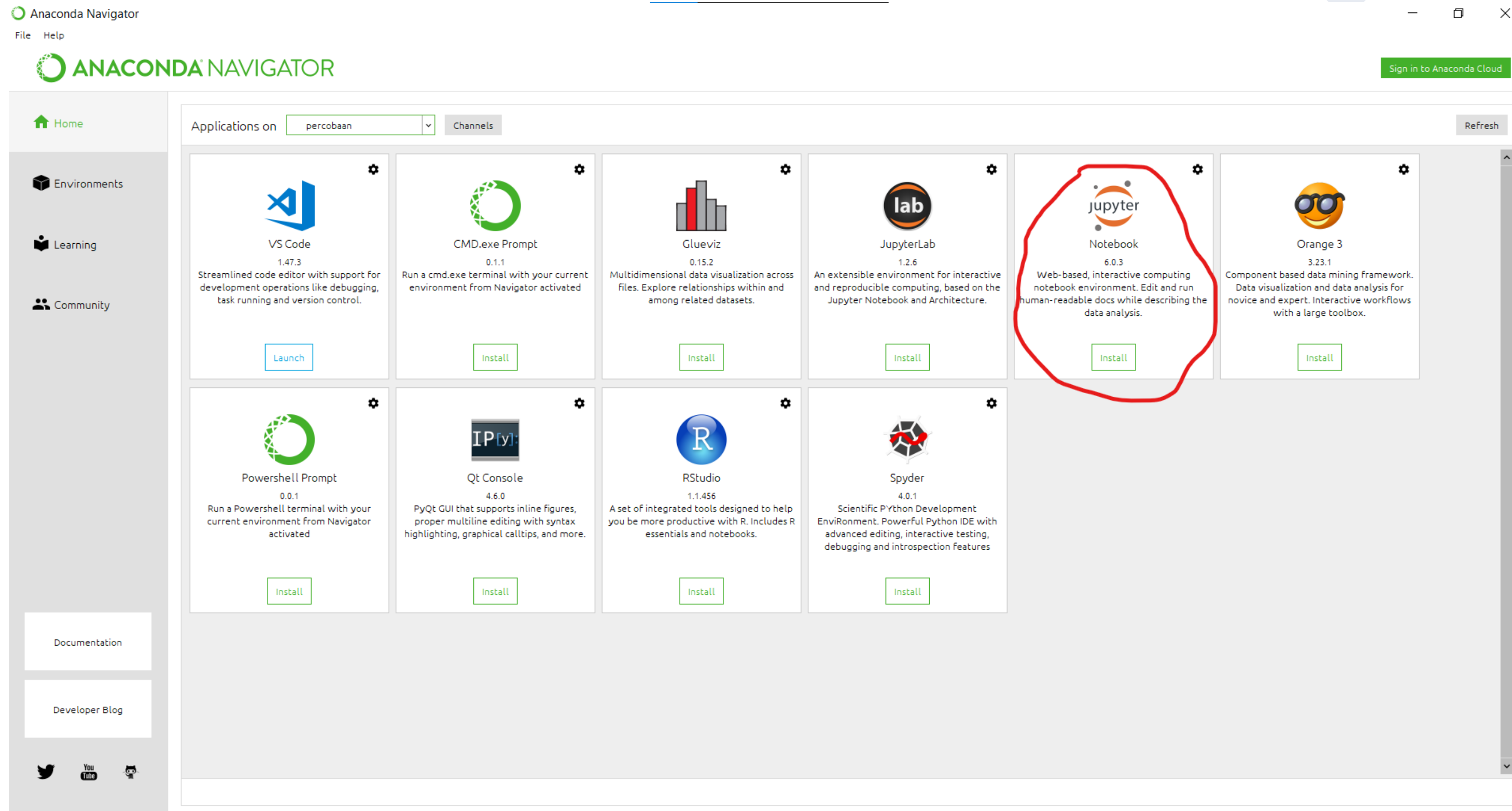
Cara Install Python dan Jupiter

3. Instal Anaconda



Cara Install Python dan Jupiter


4. Instal Jupiter



Latihan

1. Jelaskan pemahaman Anda tentang data mining?
2. Sebutkan dan jelaskan implementasi dari data mining?
3. Bagaimana menurut Anda peran data mining dalam menyelesaikan permasalahan terkait dengan data saat ini?





TERIMA KASIH