MATERI MODUL ONLINE DATA MINING
PENGERTIAN DAN KONSEP DASAR DATA MINING
SESI ONLINE 1

Syefira Salsabila

1. Pendahuluan

Dengan kemajuan teknologi informasi dewasa ini, kebutuhan akan informasi yang akurat sangat dibutuhkan dalam kehidupan sehari-hari, sehingga informasi akan menjadi suatu elemen penting dalam perkembangan masyarakat saat ini dan waktu mendatang. Namun kebutuhan informasi yang tinggi kadang tidak diimbangi dengan penyajian informasi yang memadai, sering kali informasi tersebut masih harus di gali ulang dari data yang jumlahnya sangat besar. Kemampuan teknologi informasi untuk mengumpulkan dan menyimpan berbagai tipe data jauh meninggalkan kemampuan untuk menganalisis, meringkas dan mengekstrak pengetahuan dari data. Metode tradisional untuk menganalisis data yang ada, tidak dapat menangani data dalam jumlah besar. Pemanfaatan data yang ada di dalam system informasi untuk menunjang kegiatan pengambilan keputusan, tidak cukup hanya mengandalkan data operasional saja, diperlukan suatu analisis data untuk menggali potensi-potensi informasi yang ada. Para pengambil keputusan berusaha untuk memanfaatkan gudang data yang sudah dimiliki untuk menggali informasi yang berguna membantu mengambil keputusan, hal ini mendorong munculnya cabang ilmu baru untuk mengatasi masalah penggalian informasi atau pola yang penting atau menarik dari data dalam jumlah besar, yang disebut dengan *data mining.*

Sebagai bidang ilmu yang relatif baru, saat ini *Data Mining* (DM) menjadi salah satu pusar perhatian para akademisi maupun praktisi. Beragam riset salah satu pusat perhatian para akademisi maupun praktisi. Beragam riset dan pengembangan DM telah memberikan banyak prosuk yang berguna bagi masyarakat luas, walaupun ada sebagian masyarakat yang merasa dirugikan atau kurang nyaman dengan hadirnya DM.

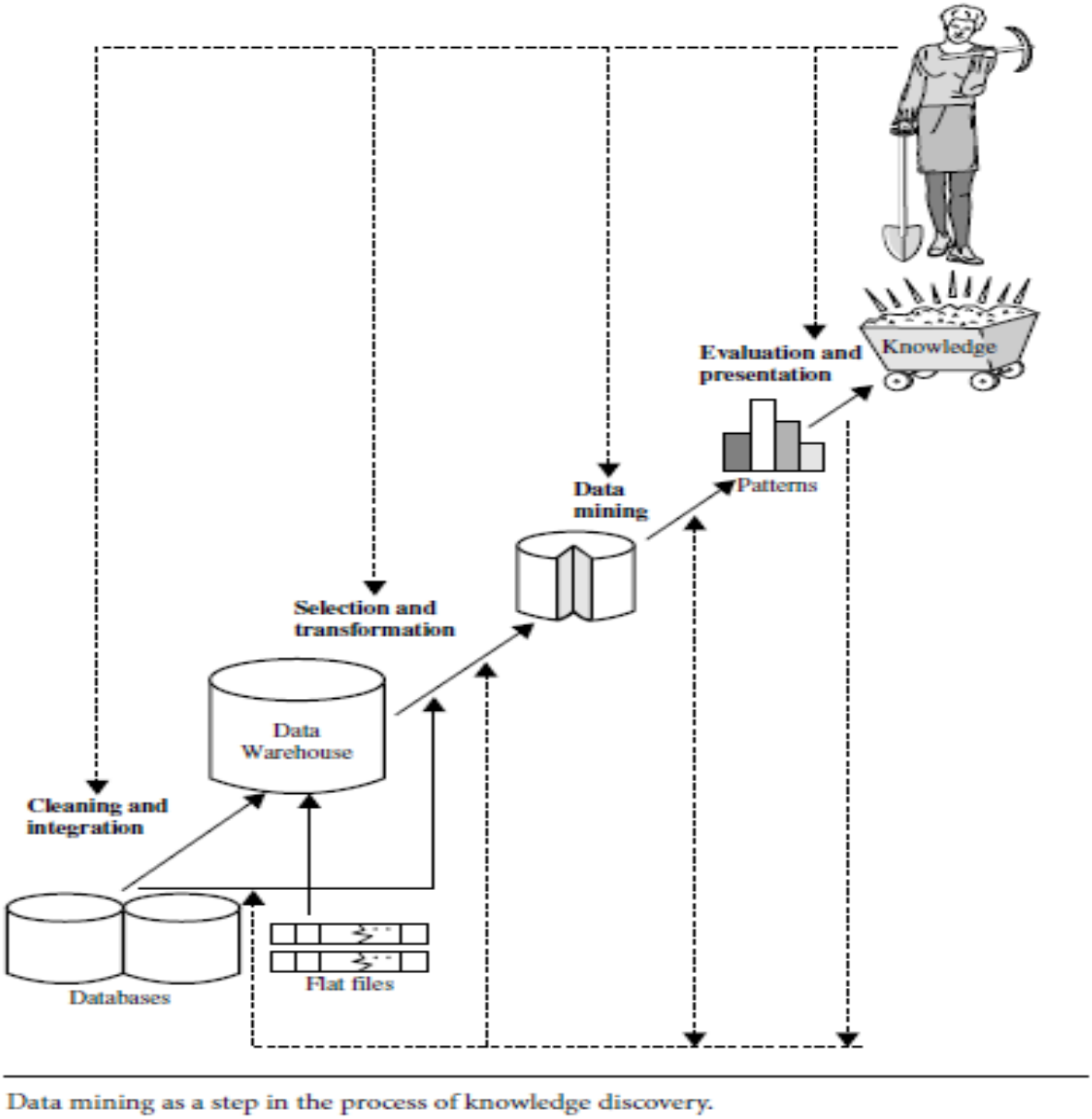
2. Apa itu Data Mining?

Definisi umum dari *data mining* itu sendiri adalah proses pencarian pola-pola yang tersembunyi (*hidden patern*) berupa pengetahuan (*knowledge*) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada di dalam *database*, data *werehouse*, atau media penyimpanan informasi yang lain. Hal penting yang terkait di dalam *data mining* adalah:

a. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
b. Data yang akan diproses berupa data yang sangat besar.

c. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Cara pandang dan pengetahuan yang berbeda membuat para ahli memberikan definisi berbeda tentang DM. sebagian ahli menyatakan bahwa DM adalah langkah analisis terhadap proses penemuan pengetahuan di dalam basis data *knowledge discovery in databases* yang disingkat KDD. Pengetahuan bisa berupa pola data atau relasi antar data yang *valid* (yang tidak diketahui sebelumnya).



Gambar 1. *Basis Data Knowledge*

a. Pre-*processing* / *cleaning*
   Proses membersihkan data dari data noise dan tidak konsisten. Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

b. Data Integration
   Proses untuk menggabungkan data dari beberapa sumber yang berbeda.

c. Data *selection*
   Proses untuk memilih data dari database yang sesuai dengan tujuan analisis. Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

d. *Transformation*
   Proses mengubah bentuk data menjadi data yang sesuai untuk proses Mining. *Coding* adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

e. *Data mining*
   Proses penting yang menggunakan sebuah metode tertentu untuk memperoleh sebuah pola dari data. *Data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keselurahan.

f. *Interpretation* / *evalution*
   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagaian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang d*item*ukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

g. *Knowledge Presentation* adalah yang dapat merepresentasikan informasi yang dibutuhkan, proses dimana informasi yang telah didapatkan kemudian digunakan oleh pemilik data.

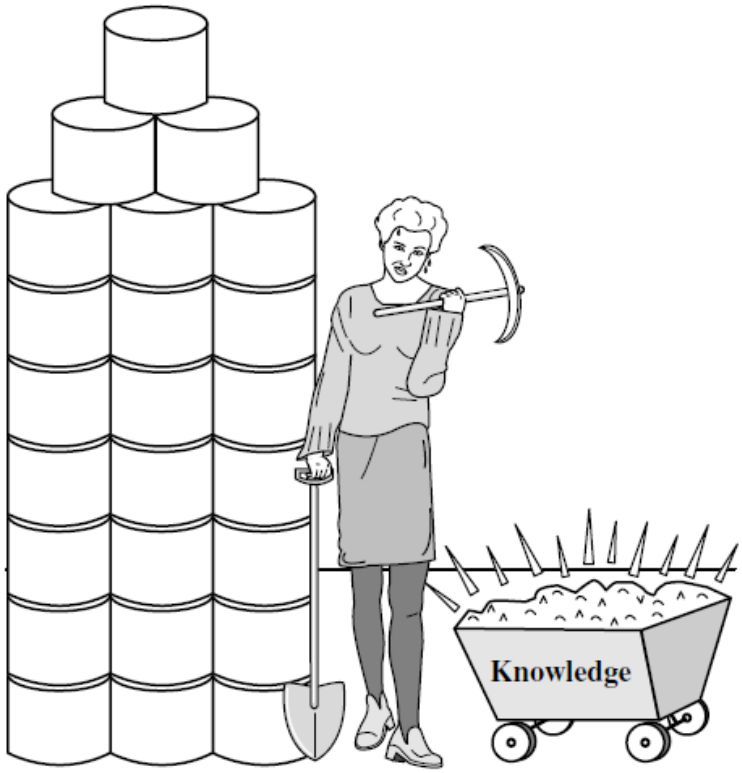
DM merupakan gabungan sejumlah disiplin ilmu komputer, yang didefinisikan sebagai proses penemuan pola-pola baru dari kumpulan-kumpulan data yang sangat besar, meliputi metode-metode yang merupa irisan dari *artificial intelligence, machine learning, statistics* dan *database systems.* DM ditunjukkan untuk mengekstrak

(mengambil intisari) pengetahuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basis data dan managemen data, prapemrosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, pascapemrosesan terhadap struktur yang ditemukan, visualisasi dan *online updating.*

Data mining dilakukan dengan *tool* khusus, yang mengeksekusi operasi *data mining* yang telah didefinisikan berdasarkan model analisis. *Data mining* merupakan proses analisis terhadap data dengan penekanan menemukan informasi yang tersembunyi pada sejumlah data besar yang disimpan ketika menjalankan bisnis perusahaan. Kemajuan luar biasa yang terus berlanjut dalam bidang *data mining* didorong oleh beberapa faktor antara lain:

a. Pertumbuhan yang cepat dalam kumpulan data.
b. Penyimpanan data dalam data *warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam *database* yang andal.
c. Adanya peningkataan akses data melalui navigasi web dan internet.
d. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
e. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi.
f. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

3. Mengapa Perlu *Data Minig?*



Data mining—searching for knowledge (interesting patterns) in data.

Gambar 2. Konsep *Data Mining*

Necessity, who is the mother of invention. – Plato

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. It is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. Even the term *data mining* does not really present all the major components in the picture. To refer to the mining of gold fromrocks or sand, we say *gold mining* instead of rock or sand mining. Analogously, data mining should have been more appropriately named “knowledge mining from data,” which is unfortunately somewhat long. However, the shorter term, *knowledge mining* may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

Beberapa tahun terakhir, data semakin heterogen dan kompleks dengan volume yang meningkay cepat secara eksponensial. Volume data pada tahun 2011 mencapai

1,8 zettabyte atau 1,8 trilyun gigabyte, pada tahun 2012 meningkat lebih dari 50% menjadi 2,8 zettabytes. Pada tahun 2013 volume data sudah menjadi 4,4 zettabytes, dan akan terus meningkat dengan cepat sehingga diperkirakan mencapai 44 zettabytes di tahun 2020. Oleh karena itu, saat ini dikenal istilah *big data;* yang menggambarkan volume data sangat besar, terstruktur maupun tidak, yang membanjiri dunia bisnis. *Big data* dapat dianalisis sehingga perusahan/instansi dapat mengambil keputusan-keputusan strategis bisnis dengan lebih baik. Beberapa kegiatan manusia yang memproduksi data:

a. Mengakses *world wide web / Log kunjungan Web.*
b. Riset science dan engineering / Akuisisi data dalam penelitian-penelitian. seperti; astronomi, kesehatan (menghasilkan rekam medis), dll.
c. Transaksi penjualan, baik transaksi penjualan online maupun transaksi penjualan di supermarket.
d. Transaksi perbankan dan kartu kredit, Dll

Dalam dunia kesehatan data mining dapat dimanfaatkan untuk mendapatkan informasi seperti penentuan kriteria suatu penyakit, misalnya tingkat kebutuhan transfusi darah dari penderita Thalassaemia. Kebutuhan waktu transfusi yang berbeda-beda dari setiap penderita menjadi sebuah masalah dalam mempersiapkan pemberian jumlah obat terafi kelasi besi dan kesiapan pendonor darah.

Contoh penerapan data mining dalam dunia nyata antara lain:

a. Midwest grocery chain menggunakan DM untuk menganalisisi 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.
b. Jika Anda mempunyai kartu kredit, sudah pasti Anda bakal sering menerima surat berisi brosur penawaran barang atau jasa. Jika Bank pemberi kartu kredit Anda mempunyai 1.000.000 nasabah, dan mengirimkan sebuah (hanya satu) penawaran dengan biaya pengiriman sebesar Rp. 1.000 per buah maka biaya yang dihabiskan adalah Rp. 1 Milyar!! Jika Bank tersebut mengirimkan penawaran sekali sebulan yang berarti 12x dalam setahun maka anggaran yang dikeluarkan per tahunnya adalah Rp. 12 Milyar!! Dari dana Rp. 12 Milyar yang dikeluarkan, berapa persenkah konsumen yang benar-benar membeli? Mungkin hanya 10 %-nya saja. Secara harfiah, berarti 90% dari dana tersebut terbuang sia-sia.

Contoh kasus di atas merupakan salah satu persoalan yang dapat diatasi oleh data mining dari sekian banyak potensi permasalahan yang ada. Data mining dapat menambang data transaksi belanja kartu kredit untuk melihat manakah pembeli-pembeli yang memang potensial untuk membeli produk tertentu. Mungkin tidak sampai

presisi 10%, tapi bayangkan jika kita dapat menyaring 20% saja, tentunya 80% dana dapat digunakan untuk hal lainnya.

4. Kegunaan *Data Minig?*

Secara umum, kegunaa DM dapat dibagi menjadi dua: deskriptif dan prediktif. Deskriptif berarti DM digunakan untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data. Sedangkan prediktif berarti DM digunakan untuk membentuk sebuah model pengetahuan yang akan digunakkan melakukan prediksi. Berdasarkan fungsionalitasnya, tugas-tigas DM bias dikelompokkan ke dalam enam kelompok berikut ini:

a. Klasifikasi *(classification)*
   Mengeneralisasi struktur yang diketahui untuk diaplikasikan pada data-data baru. Misalkan, klasifikasi penyakit ke dalam sejumlah jenis, klasifikasi email ke dalam spam atau bukan.
b. Klasterisasi *(clustering)*
   Mengelompokkan data, yang tidak diketahui label kelasnya, ke dalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripannya.
c. Regresi *(regression)*
   Menemukan suatu fungsi

5. Pengelompokan *Data Mining*

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

a. Deskripsi
   Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data. Sebagai contoh, petugas pengumpul suara mungkin tidak menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecendrungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecendrungan.
b. Estimasi
   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun dengan *record* lengkap menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.
c. Prediksi
   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

d. Klasifikasi
   Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.
e. Pengklusteran
   Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.
f. Asosiasi
   Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (*market basket analysis*).

6. Data Mining dari berbagai sudut pandang

Berikut ini menjelaskan mengenai DM dari beberapa sudut pandang, yaitu

a. Dari sudut **Data**
   Relational, datawarehouse,web, transaksional, stream, OO, spasial, text, multimedia
b. Dari sudur **Pengetahuan yang akan ditambang**
   Karakterisitik, diskriminasi, asosiasi, klasifikasi, clustering, trend, outlier
c. Dari sudut **Teknik**
   Database, OLAP, machine learning, statistik, visualiasi
d. Dari sudut **Penerapan**
   Retail, telekomunikasi, banking, analisis kejahatan, bio-data mining, saham, text mining, web mining

7. Data pada Data Mining

a. **Database Tradisional**
   Relational database, data warehouse, transactional database
b. **Advanced Database**
   a) Data streams dan data sensor
   b) Time-series data, temporal data, sequence data (incl. bio-sequences)
   c) Structure data, graphs, social networks and multi-linked data
   d) Object-relational databases

e) Heterogeneous databases dan legacy databases
f) Spatial data dan spatiotemporal data
g) Multimedia database
h) Text databases
i) World-Wide Web

8. Fungsi dari Data Mining

Fungsi atau sub kegiatan yang ada dalam data mining dalam rangka menemukan, menggali, atau menambang pengetahuan, terdapat enam fungsi dalam data mining, yaitu:

a. Fungsi deskripsi (description)
b. Fungsi estimasi (estimation)
c. Fungsi prediksi (prediction)
d. Fungsi klasifikasi (classification)
e. Fungsi pengelompokan (classification),
f. Fungsi asosiasi (association).

Keenam fungsi data mining tersebut dapat dipilah menjadi:

a. Fungsi minor atau fungsi tambahan, yang meliputi ketiga fungsi pertama, yaitu *deskripsi*, estimasi, dan prediksi
b. Fungsi mayor atau fungsi utama, yang meliputi ketiga fuingsi berikutnya, yaitu klasifikasi, pengelompokkan, dan asosiasi.

9. Model dalam data mining

a. **Verification Model**
Model ini menggunakan (hypothesis) dari pengguna, dan melakukan test terhadap perkiraan yang diambil sebelumnya dengan menggunakan data-data yang ada. Model *verifikasi* menggunakan pendekatan *top down* dengan mengambil hipotesa dari user dan memeriksa validitasnya dengan data sehingga bisa dibuktikan kebenaran hipotesa tersebut.

b. **Discovery Model**
Sistem secara langsung menemukan informasi-informasi penting yang tersembunyi dalam suatu data yang besar. Data-data yang ada kemudian dipilah-pilah untuk menemukan suatu pola, trend yang ada, dan keadaan umum pada saat itu tanpa adanya campur tangan dan tuntutan dari pengguna.

Model *knowledge discovery* menggunakan pendekatan *bottom up* untuk mendapatkan informasi yang sebelumnya tidak diketahui. Model ini terbagi menjadi dua *directed knowledge discovery* dan *undirected knowledge discovery*.

Pada *directed knowledge discovery*, data mining akan mencoba mencari penjelasan nilai target field tertentu (seperti pengahasilan, respons, usia, dan lain-lain) terhadap field-field yang lain.

Pada *undirected knowledge discovery* tidak ada target field karena komputer akan mecari pola yang ada pada data. Jadi *undirected knowledge discovery* digunakan untuk mengenali hubungan/relasi yang ada pada data sedangkan directed discovery akan menjelaskan hubungan/relasi tersebut.

10. Permasalahan dalam data mining

a. Metodologi
    a) Mining beragam pengetahuan dari beragam sumber data
    b) Kinerja: efesiensi, efektivitas dan skalabilitas
    c) Evaluasi pola
    d) Background knowledge
    e) Noise (gangguan) dan data yang tidak lengkap
    f) Distributed dan paralel method.
    g) knowledge fusion (penggabungan)
b. Interaksi pengguna
    a) Data mining query languages dan ad-hoc mining
    b) Visualisasi
    c) Interactive mining
c. Aplikasi
    a) Domain spesifiik
    b) Perlindungan data

11. Aplikasi pada data mining

a. Pemasaran/ Penyewaan
    a) Identifikasi pola pembayaran pelanggan
    b) Menemukan asosiasi diantara karakteristik demografik pelanggan
    c) Analisis keranjang pemasaran

b. Perbankan
    a) Mendeteksi pola penyalahgunaan kartu kredit
    b) Identifikasi pelanggan yang loyal
    c) Mendeteksi kartu kredit yang dihabiskan oleh kelompok pelanggan

c. Asuransi & Pelayanan Kesehatan
    a) Analisis dari klaim
    b) Memprediksi pelanggan yang akan membeli polis baru
    c) Identifikasi pola perilaku pelanggan yang berbahaya

d. Analisa Perusahaan dan Manajemen Resiko
    a) Perencanaan Keuangan dan Evaluasi Aset

b) Perencanaan Sumber Daya (Resource Planning)
c) Persaingan (competition) → Competitive Intelligence

e. Telecomunication
a) menerapkan data mining untuk melihat dari jutaan transaksi yang masuk, transaksi mana saja yang masih harus ditangani secara manual (dilayani oleh orang).

12. Kaitan Data Mining Dengan Beberapa Disiplin Ilmu

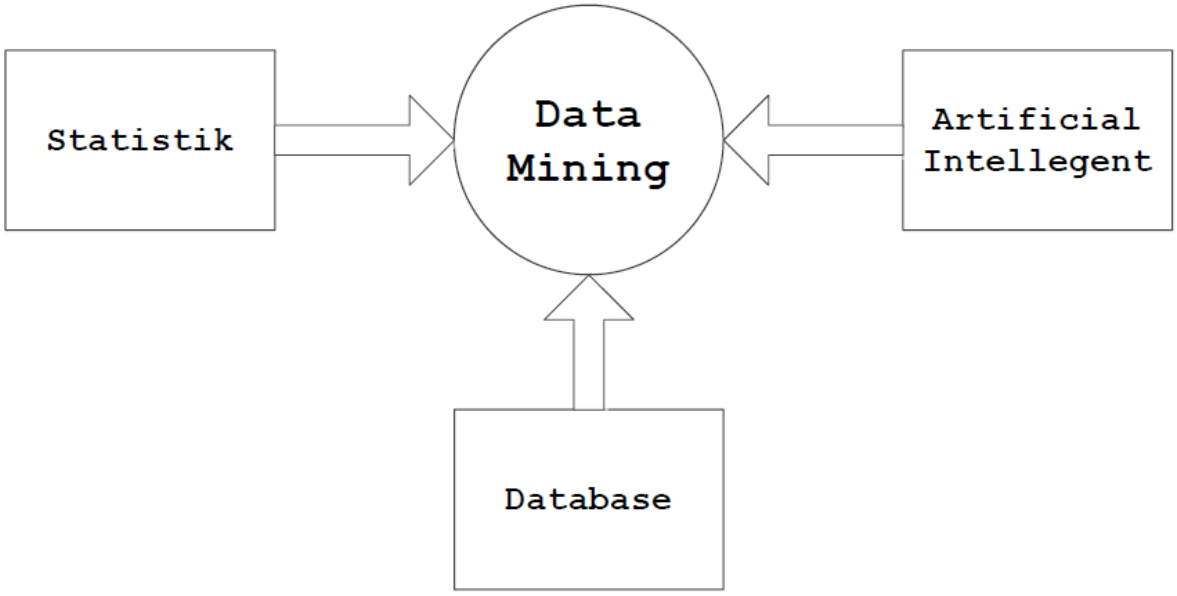
Setiap perusahaan atau organisasi lainnya yang sudah menerapkan sistem informasi, tentunya akan melibatkan penyimpanan data dalam proses bisnisnya. Hal tersebut mengakibatkan data-data tersebut tersimpan dalam sebuah basis data yang kapasitasnya semakin bertambah atau membesar. Kondisi data perusahaan yang semakin membesar menyebabkan biaya perawatan semakin meningkat. Yang menjadi pertanyaan, apakah data-data tersebut hanya kita gunakan untuk pelaporan saja kemudian dibuang atau dikubur dan dibiarkan? Tentu sangat disayangkan apabila data tersebut tidak bisa dimanfaatkan oleh perusahaan.

Kondisi di atas merupakan landasan munculnya data mining. Dimana data mining ini dapat kita It is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. Even the term *data mining* does not really present all the major components in the picture. To refer to the mining of gold fromrocks or sand, we say *gold mining* instead of rock or sand mining. Analogously, data mining should have been moregunakan untuk menghasilkan manfaat dari kumpulan data perusahaan yang sangat besar. Manfaat tersebut berupa informasi atau pengetahuan untuk membantu perusahaan dalam mengambil keputusan.

Sebagai contoh pengiriman surat penawaran barang dan jasa pada nasabah yang memiliki kartu kredit. Jika bank yang bersangkutan memiliki 1.000.000 nasabah dan biaya pengiriman surat per nasabah adalah 500 rupiah, maka biaya yang diperlukan adalah 500 juta rupiah padahal nasabah yang mungkin benar-benar membeli hanya sekitar 15% sehingga ada pembuangan/kerugian biaya sekitar 85% dari 500 juta atau sekitar 425 juta rupiah. Dengan data mining maka perusahaan dapat memanfaatkan data-data yang ada sehingga hanya mengirim surat kepada nasabah yang berpotensi untuk membeli, sehingga biaya pengiriman tersebut dapat ditekan atau diturunkan.

Hubungan yang dicari dalam data mining berupa keterkaitan pembelian suatu produk dengan produk lainnya. Sedangkan contoh pola adalah sebuah perusahaan yang akan meningkatkan fasilitas kartu kredit dari pelanggan, maka perusahaan akan mencari pola dari pelanggan-pelanggan yang ada untuk mengetahui pelanggan yang potensial dan pelanggan yang tidak potensial.

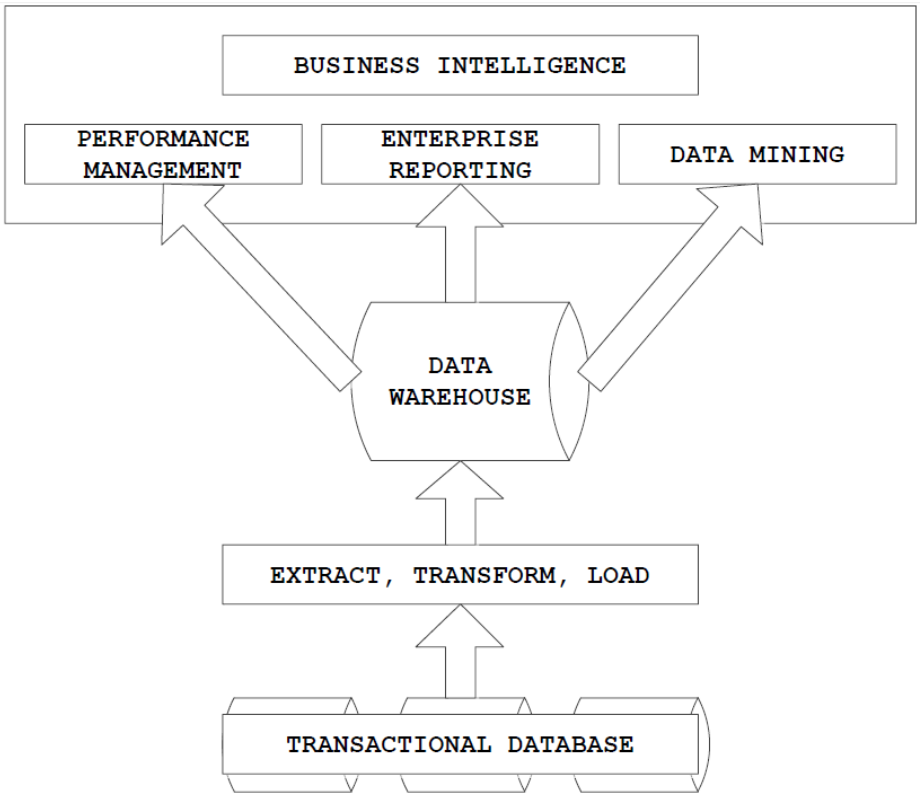
Data mining bukanlah bidang yang sama sekali baru. Data mining memiliki beberapa kesamaan karakteristik atau aspek dengan disiplin ilmu lainnya, yang diantaranya: statistik, basis data dan kecerdasan buatan. Kaitan data mining dengan disiplin ilmu lainnya digambarkan pada gambar 3 di bawah ini.



Gambar 3. Kaitan Data Mining dengan Bidang Ilmu Lain

Kesamaan bidang data mining dengan statistic adalah dalam hal penyampelan, estimasi dan pengujian hipotesis. Kesamaan dengan kecerdasan buatan (*artificial intelligence*), pengenalan pola (*pattern recognition*), dan pembelajaran mesin (*machine learing*) adalah algoritma pencarian, teknik pemodelan, dan teori pembelajaran. Bidang lain yang jung mempengaruhi data mining adalah teknologi basis data, yang mendukung penyediaan penyimpanan yang efisien (normalisasi), pengindekan dan pemrosesan *query*.

Bidang ilmu lainnya yang berkaitan dengan data mining adalah *information science, high performance computing*, visualisasi, *neural networks*, pemodelan matematika, *information retrieval and extraction*, serta pengenalan pola. Sudah diketahui pada bahasan sebelumnya bahwa. Data mining merupakan salah satu dari tahapan KDD. Selanjutnya, apa perbedaan antara data mining dengan data warehouse? Untuk menjawabnya pada gambar 4 di bawah ini dijelaskan tentang posisi antara data mining dan data warehouse.



Gambar 4. Posisi Data Mining dalam Business Intelligence

Dari gambar di atas dapat dilihat bahwa data mining menggunakan data yang dihasilkan oleh data warehouse, bersama dengan bidang yang menangani masalah pelaporan dan manajemen data. Sementara data warehouse sendiri bertugas untuk menarik data dari basis data untuk menghasilkan data yang nantinya digunakan oleh bidang lainnya. Data mining dan data warehouse dapat digunakan untuk menunjang sistem informasi pendukung keputusan dan sistem informasi manajemen. Data mining juga dapat digunakan untuk bidang knowledge management.

Operations	Data mining techniques
Predictive modeling	Classification
	Value prediction
Database segmentation	Demographic clustering
	Neural clustering
Link analysis	Association discovery
	Sequential pattern discovery
	Similar time sequence discovery
Deviation detection	Statistics
	Visualization

Gambar 5. Operasi Data Mining dan Teknik Asosiasi

13. Clustering

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Perlu diingat bahwa kata *mining* sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basisdata. Beberapa teknik yang sering disebut-sebut dalam literatur *data mining* antara lain yaitu *association rule mining*,*clustering*, *klasifikasi*, *neural network,* dan lain-lain.

C*lustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik *(similarity)* antara satu data dengan data yang lain. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data atau obyek ke dalam *cluster* (*group*) sehingga dalam setiap *cluster* dapat berisi data yang semirip mungkin. Dalam *clustering* metode ini berusaha untuk menempatkan obyek yang mirip (jaraknya dekat) dalam satu *cluster* dan membuat jarak antar *cluster* sejauh mungkin. Ini berarti obyek dalam satu *cluster* sangat mirip satu dengan lain dan berbeda dengan obyek dalam *cluster-cluster* yang lain. Dalam *data mining* ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering.*

Metode *non-hierarchical (partitioning) clustering* dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (dua *cluster*, tiga *cluster,* atau lain sebagainya). Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan. Metode ini biasa disebut dengan *K-Means Clustering*

14. Preprocess

Pada proses ini data dipersiapkan agar dapat digunakan untuk digali informasinya. Hal ini dilakukan untuk mendapatkan hasil analisis yang lebih akurat dalam pemakaian teknik-teknik *mechine learning* maupun *data mining*. Komponen yang terdapat dalam data itu sendiri terdiri dari : Obyek (record, point, case, sampel, entitas, instan) dan Atribut / variabel / field yaitu karakteristik dari obyek (status pernikahan, umur, dll)

Data cleaning terdapat proses pembersihan data yang kosong atau data-data lain yang dapat mengakibatkan *noise*/*error* dengan memperkecil adanya data *outlier*. *Data Cleaning* menjadi proses yang sangat penting dikarenakan data perlu dibersihkan agar analisa menjadi lebih akurat. Ada beberapa cara transformasi data yang dilakukan sebelum kita menerapkan suatu metode tergantung pada metode apa yang kita lakukan untuk proses *data mining.* Sebelum kita mempelajari teknik/metode yang digunakan dalam *data mining,* ada baiknya kita bedakan dulu metode belajar (*learning*) secara garis besar ke dalam dua pendekatan : *supervised* dan *unsupervised.* Dalam pendekatan pertama, *unsupervised learning,* metode kita terapkan tanpa adanya latihan (training) dan tanpa ada label dari data. Misalkan kita punya sekelompok pengamatan atau data tanpa ada label (output) tertentu yang menandai kemana data dikelompokkan.

15. Classification

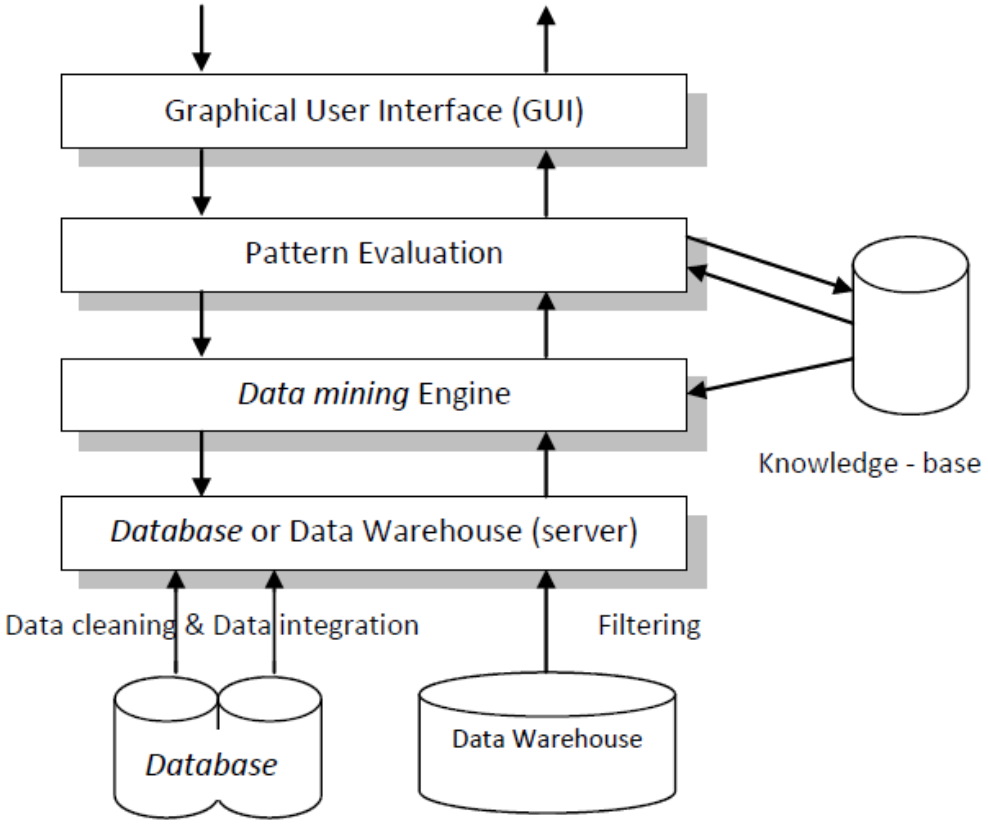
Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah di definisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik ini menggunkan *supervised induction*, yang memanfaatkan kumpulan pengujian dari record yang terklasifikasi untuk menentukan kelas-kelas tambahan. Salah satu contoh yang mudah dan populer adalah dengan Decision tree yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi.

16. Arsitektur Dari Sistem *Data Mining*

Arsitektur utama dari sistem *data mining*, pada umumnya terdiri dari beberapa komponen sebagai berikut:

a. *Database*, data *warehouse*, atau media penyimpanan informasi, terdiri dari satu atau beberapa *database*, data *warehouse*, atau data dalam bentuk lain. Pembersihan data dan integrasi data dilakukan terhadap data tersebut.
b. *Database*, data *warehose*, bertanggung jawab terhadap pencarian data yang relevan sesuai dengan yang diinginkan pengguna atau *user*.
c. Basis pengetahuan (*Knowledge Base*), merupakan basis pengetahuan yang digunakan sebagai panduan dalam pencarian pola.

d. *Data mining engine*, merupakan bagaian penting dari sistem dan idealnya terdiri dari kumpulan modul-modul fungsi yang digunakan dalam proses karakteristik (*characterization*), klasifikasi (*clasiffication*), dan analisis kluster (*cluster analysis*). Dan merupakan bagian dari *software* yang menjalankan program berdasarkan algoritma yang ada.
e. Evaluasi pola (*pattern evaluation*), komponen ini pada umumnya berinteraksi dengan modul-modul *data mining*. Dan bagian dari *software* yang berfungsi untuk menemukan *pattern* atau pola-pola yang terdapat dalam *database* yang 12 diolah sehingga nantinya proses *data mining* dapat menemukan *knowledge* yang sesuai.
f. Antar muka (*Graphical user interface*), merupakan modul komunikasi antara pengguna atau user dengan sistem yang memungkinkan pengguna berinteraksi dengan sistem untuk menentukan proses *data mining* itu sendiri.



Gambar 6. Arsitektur *Data Mining*

DAFTAR PUSTAKA

Gorunescu, F. (2011). *Data Mining : Concepts, Models and Techniques.* New York: Springer-Verlag.

Han, J. Kamber, M & Jian, Pei. *Data Mining : Concepts and techniques, Third Edition.* America: Morgan Kauffman, San Francisco, 2011.

Hofmann M, Klinkenberg R. 2014. Rapid Miner Data Mining Use Case and Business Analytics Application. Taylor and Francis Group, LLC

Prasetyo Eko. 2014. Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. Andi Offset

Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*.Yogyakarta: Graha Ilmu.

Santoso, S. (2010). *Statistik Multivariat.* Jakarta: Elex Media Komputindo.