Mengenal Data Mining dan Fasilitas DM di Apache Spark

Dr. Bambang Purnomosidi D. P.
https://zimera-systems.com

Zimera
Software and Data Technologies

Agenda

zimera systems

1. Pengertian Data Mining
2. Arti Penting Data Mining
3. Data Mining dan Istilah-istilah Terkait
4. DM dan Proses KDD
5. Metodologi DM: CRISP-DM dan ASUM-D
6. *Tasks* DM:
    a. Deteksi Anomali
    b. Association Rule Learning
    c. Clustering
    d. Klasifikasi
    e. Regresi
    f. Peringkasan (*Summarization*)

Agenda

zimera systems

7. Apache Spark dan DM
8. DataFrame dan MLlib di Apache Spark
9. Contoh *task* DM: *Clustering* dengan K-Means

Pengertian Data Mining

zimera systems

- Pengertian sederhana: **proses untuk meng-ekstrak dan menemukan pola pada *dataset*.**
- Dataset: koleksi dari data. Bentuk dari dataset bisa bermacam-macam dengan berbagai macam format. Ada yang berupa tabel, ada yang berupa JSON atau CSV, dan lain-lain. Dataset bisa berasal dari DBMS maupun non DBMS (misal dari hasil observasi statistik, data realtime dari suatu devices, dll).
- DM merupakan interseksi dari *database systems*,statistika, dan *machine learning*.
- Istilah DM sebenarnya bisa memicu kesalahan pemahaman, seharusnya “Knowledge Mining” atau “Knowledge Mining from Data”
- DM awalnya disebut KDD (Knowledge DIscovery in Databases).

Arti Penting Data Mining

zimera systems

- Merupakan aktivitas untuk mengubah data mentah menjadi *insights* ataupun *knowledge*.
- *Insights* maupun *knowledge* tersebut digunakan untuk mengambil keputusan di domain tertentu (kedokteran, bisnis, pendidikan tinggi, dll).

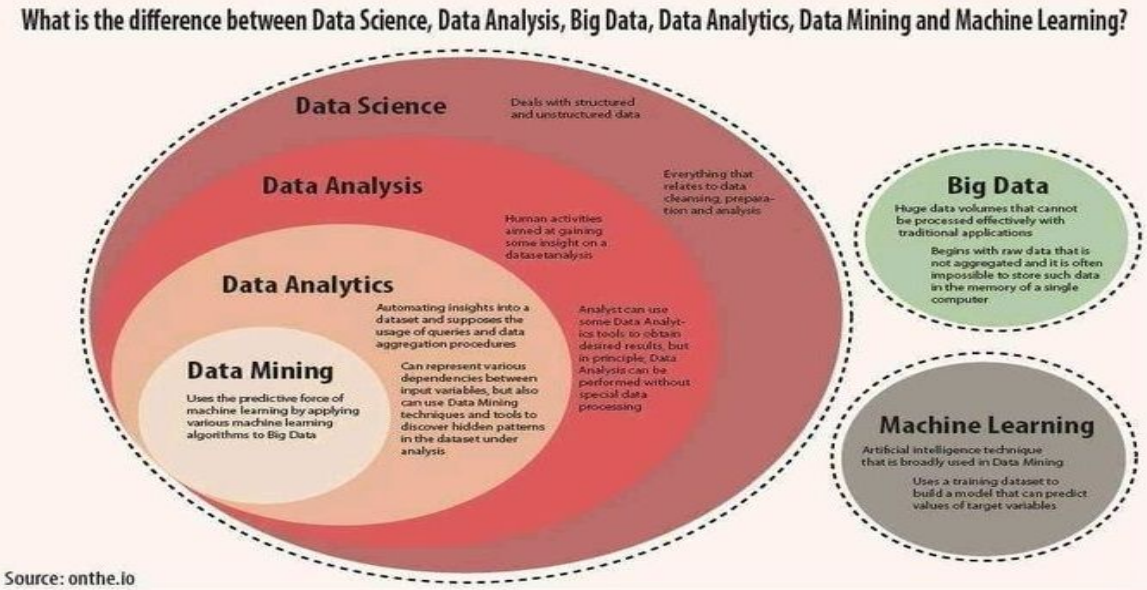
Data Mining dan Istilah-istilah Terkait

zimera systems

- **DM dan KDD**: KDD merupakan istilah awal dari DM. Saat ini, KDD juga disebut KD from Data. KDD juga berarti KD and DM (SIG - Special Interest Group dari ACM - http://www.kdd.org/). Inti dari DM dan KDD adalah *Knowledge Discovery*.
- **DM dan Machine Learning**: DM terkait dengan dataset dan pengambilan simpulan berupa *insights* atau *knowledge*. ML menggunakan berbagai algoritma untuk membuat mesin bisa “belajar”. DM dan ML keduanya menggunakan algoritma yang kompleks dan mungkin menggunakan algoritma yang sama.
- **DM, Data Analyis, Data Analytics, Data Science**: lihat diagram setelah ini. Data analysis => What happened?. Data Analytics => data analysis + what will happen?
- **DM dan Data Engineering**: DE / Rekayasa Data merupakan aktivitas dan proses untuk ETL (Data Warehouse) maupun ELT (Data Lake). Untuk ETL, DM akan menggunakan dataset di DW untuk KD. Untuk ELT, DM merupakan proses T.

Data Mining dan Istilah-istilah Terkait

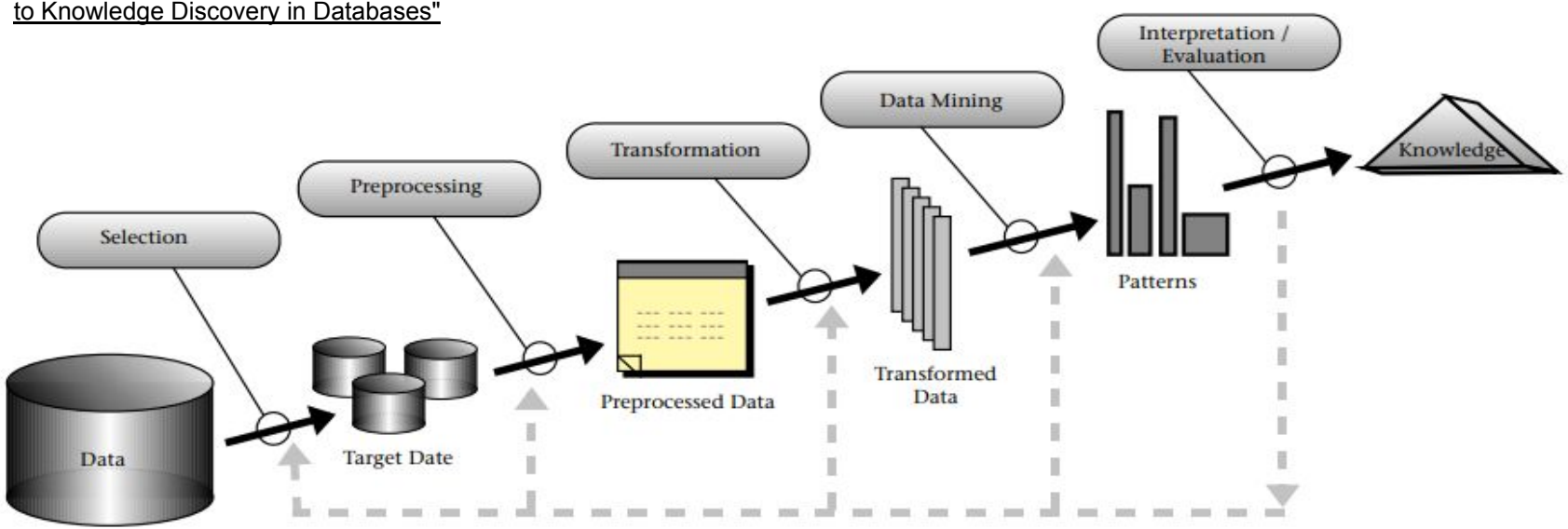
zimera systems



Data Mining dan Proses KDD

zimera systems

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases"



Metodologi Data Mining

zimera systems

- Metode: cara teratur yang digunakan untuk melaksanakan suatu pekerjaan agar tercapai sesuai dengan yang dikehendaki (KBBI)
- Metodologi: ilmu tentang metode; uraian tentang metode (KBBI).
- Metodologi berisi rerangka kontekstual untuk melakukan riset, suatu skema yang logis dan masuk akal berbasis pada berbagai pandangan, kepercayaan, dan nilai-nilai yang memandu peneliti atau siapapun itu.
- Metodologi untuk Data Mining membahas tentang uraian dan rerangka kontekstual serta skema yang logis dan masuk akal untuk melaksanakan *tasks* yang termasuk dalam kategori Data Mining

Metodologi Data Mining (2)

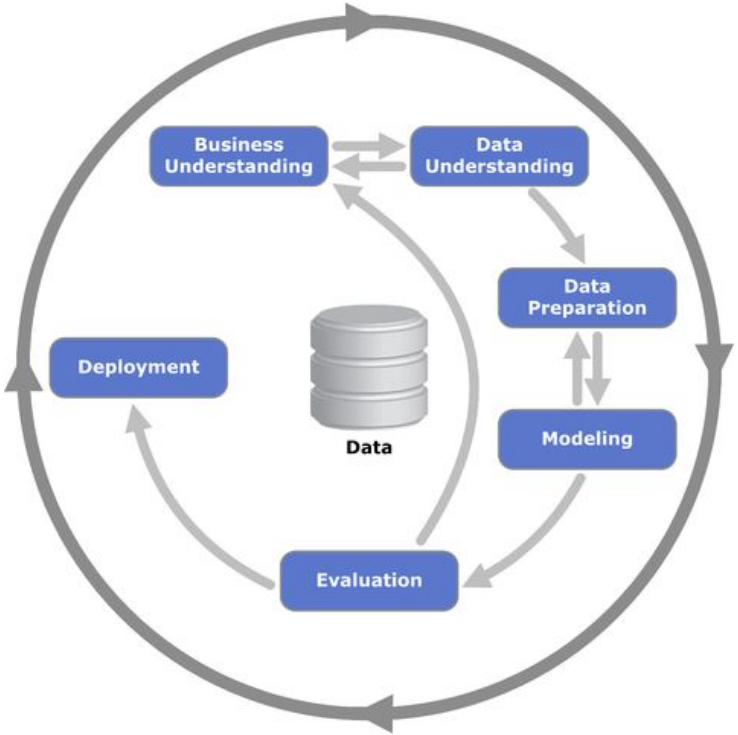
zimera systems

- Ada beberapa metodologi DM:
  - CRISP-DM (Cross-industry standard process for data mining).
  - ASUM-DM (Analytics Solutions Unified Method for Data Mining/Predictive Analytics).
  - SEMMA (Sample, Explore, Modify, Model, and Assess).

Metodologi Data Mining: CRISP-DM

zimera systems

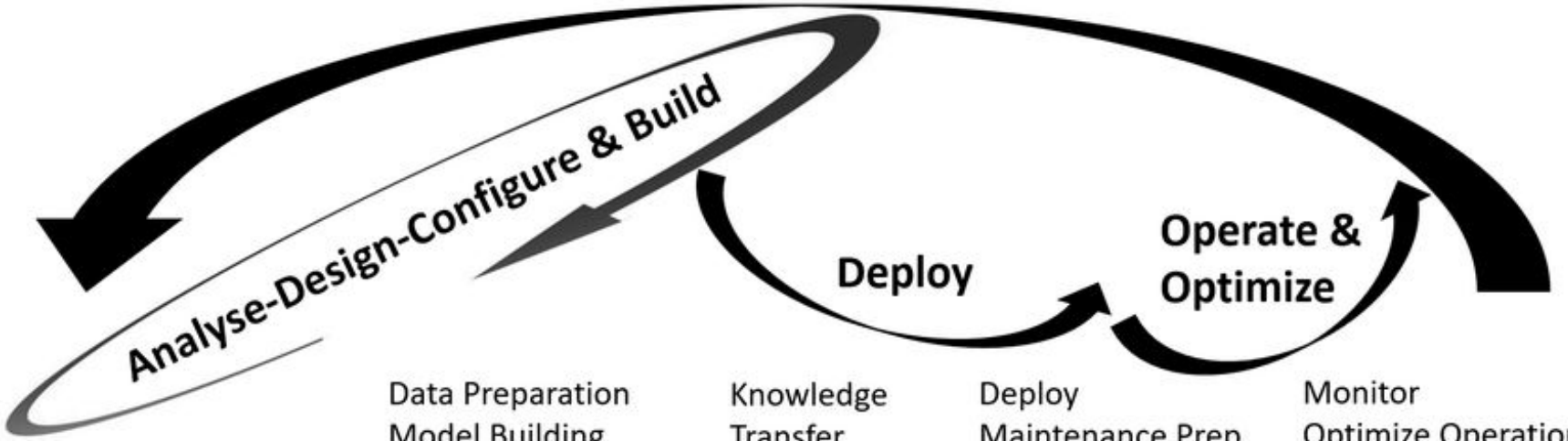
- Dirumuskan tahun 1996 dan mulai dikembangkan dan digunakan oleh konsorsium proyek Uni Eropa yang dipimpin oleh 5 perusahaan pada tahun 1997.
- Merupakan metodologi yang paling banyak digunakan.



Metodologi Data Mining: ASUM-DM

zimera systems

Dirumuskan oleh IBM sebagai perbaikan dari CRISP-DM.



Metodologi Data Mining: SEMMA

zimera systems

- Dibuat oleh SAS Institute - produsen SAS Enterprise Miner
- Berisi serangkaian langkah untuk menyelesaikan tugas inti dari DM:
  - Sample: menentukan sampel data.
  - Explore: memahami data - relasi antar variabel dalam data serta abnormalitas data. Fase ini juga mencakup visualisasi data.
  - Modify: modifikasi (select, create, transform) untuk persiapan pemodelan data.
  - Model: menerapkan teknik data mining terhadap dataset.
  - Assess: evaluasi hasil

Tasks pada Data Mining

zimera systems

- Deteksi Anomali (*Anomaly Detection, Coutlier/Change/Deviation Detection*).
- *Association Rule Learning* (*Dependency Modeling*)
- *Clustering*
- Klasifikasi
- Regresi
- Peringkasan (*Summarization*)

Task di Data Mining - Deteksi Anomali

zimera systems

Merupakan tugas / task DM untuk mengidentifikasi anomali yang biasanya dikaitkan dengan kecurigaan tertentu dari sekumpulan data.

Contoh:

- Error pada suatu teks
- Masalah medis
- Fraud / kecurangan

Task DM - Association Rule Learning

zimera systems

Merupakan task / tugas DM untuk mengidentifikasi keterkaitan antar variabel pada sekumpulan data.

Contoh:

- *Market Basket Analysis*: menentukan produk apa saja yang biasanya dibeli secara bersamaan. Ini bermanfaat untuk merekomendasikan suatu produk yang akan ditawarkan pada saat konsumen membeli suatu produk tertentu.

Task DM - Clustering

zimera systems

Merupakan task / tugas DM untuk menemukan grup dan struktur dalam data yang “dianggap” mirip.

Contoh:

- Penentuan cluster penyakit COVID 19.
- Penentuan cluster lulusan suatu perguruan tinggi dan lapangan kerja yang dimasuki.

Task DM - Classification

zimera systems

Merupakan tugas DM untuk menggeneralisir suatu struktur tertentu dan mengaplikasikan generalisasi tersebut untuk data baru:

Contoh:

- Klasifikasi e-mail ke dalam inbox dan spam.
- Klasifikasi karakter / huruf tertentu pada suatu tulisan tangan.
- Klasifikasi jenis tumbuhan tertentu
- Klasifikasi kopi jenis tertentu.

Task DM - Regression

zimera systems

Merupaan task / tugas DM untuk memprediksi huubngan antar variabel:

- *Outcome* / hasil / *dependant*
- *Independent variables / predictors / covariates / features*

Contoh:

- Apakah kondisi sosial ekonimi dan ras seseorang menentukan pencapaian akademik seseorang?
- Apakah minum kopi dan merokok mempengaruhi kualitas hasil software yang dikembangkan seorang programmer?

Task DM - Summarization

zimera systems

Merupakan *task* / tugas DM untuk menyajikan dan mempresentasikan ringkasan data baik dalam suatu laporan mautpun visualisasi.

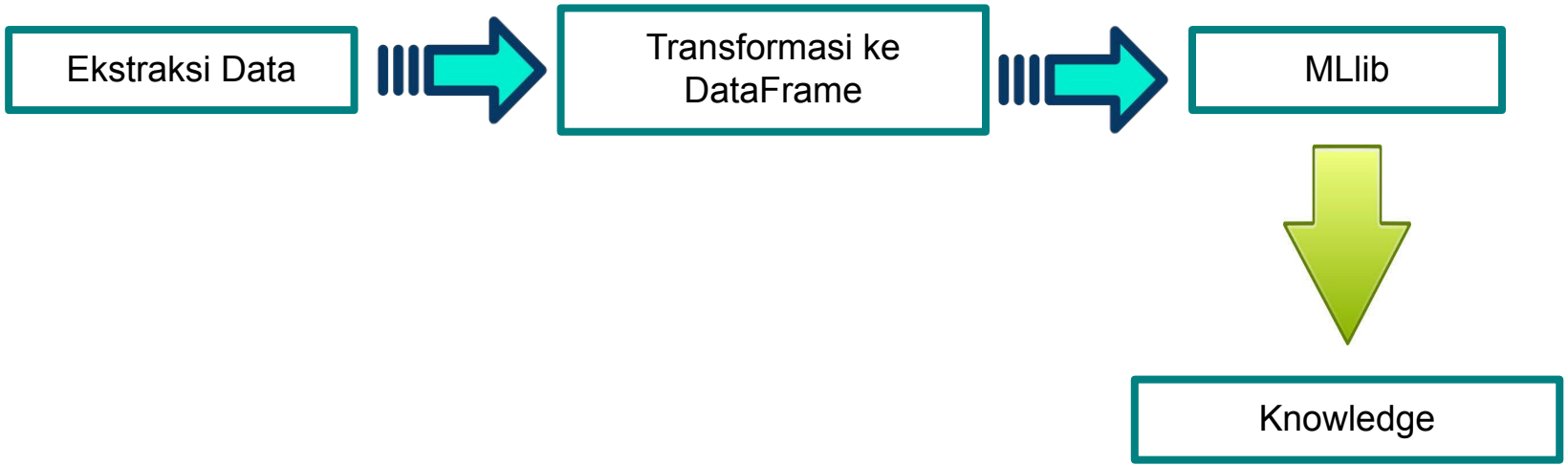
Apache Spark dan DM

zimera systems

Apache Spark bersama dengan framework lain (misal Apache Kafka) bisa digunakan untuk proses data engineering: ETL dan ELT. ETL Menghasilkan Data Warehouse ELT menghasilkan Data Lake. Keduanya ini merupakan obyek dari DM untuk melaksanakan berbagai tasks DM. Pustaka di Apache Spark utk ML (MLlib) dan Graph menyediakan fasilitas-fasilitas untuk DM.

DataFrame dan MLlib di Apache Spark

zimera systems



Demo DM: Clustering dengan K-Means

zimera
systems

Penutup

zimera
systems

Thanks!

Got question(s)?

zimera.systems@gmail.com
https://zimera-systems.com

Credits: This template includes Icons by **Flaticon** and images by **Freepik**