Data Science

Pokok Bahasan : Data Understanding

Dosen Pengampu : Hary Sabita, ST., MTI

14 Juni 2022

Outline :

1. Apa itu telaah data?
2. Sumber, susunan, tipe dan model data

Next - Apa tujuan mempelajari ini ??

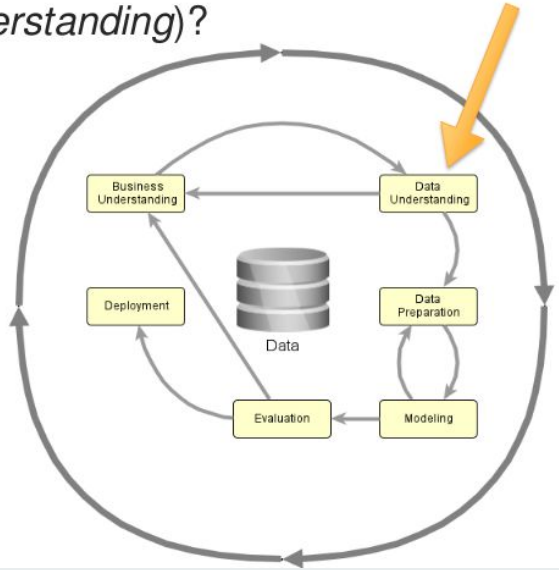
Tujuan yang akan dicapai :

1. Mampu melakukan pengambilan data untuk proses sains data dari sumber data terbuka

Next - Apa tujuan telaah data ??

Apa itu telaah data (*data understanding*)?

- Dilakukan setelah problem bisnis terdefinisikan sebagai hasil tahapan business understanding.
- Tujuan: mendapatkan gambaran utuh atas data.
- Dilanjutkan ke persiapan data (data preparation), jika pemahaman awal data cukup atau kembali ke business understanding jika definisi permasalahan bisnis harus direvisi.



2021

Next - Mengapa perlu data understanding?

Mengapa perlu data understanding?

- Data = bahan mentah solusi AI
- Data dari masing-masing sumber belum tentu dapat langsung dipakai karena:
  - maksud dan tujuan data berbeda-beda
  - keadaan asal terpisah-pisah atau justru terintegrasi secara ketat.
  - tingkat kekayaan (*richness*) berbeda-beda
  - tingkat keandalan (*reliability*) berbeda-beda
- Data understanding memberikan gambaran awal tentang:
  - kekuatan data
  - kekurangan dan batasan penggunaan data
  - tingkat kesesuaian data dengan masalah bisnis yang akan dipecahkan
  - ketersediaan data (terbuka/tertutup, biaya akses, dsb.)

Next - Bagaimana tahapannya?

Bagian-bagian proses telaah data

- Identifikasi "titik sentuh" data dengan proses bisnis
- Penentuan sumber utama data dan cara aksesnya
- Asesmen nilai tambah bisnis dari data
- Identifikasi sumber data tambahan untuk perbaikan

Next - Sumber, susunan, tipe dan model

Sumber data

Internal sources	Spreadsheets (Excel, CSV, JSON, etc.)
	Databases: can be queried via SQL, etc.
	Text documents
	Multimedia documents (audio, video)
External sources	Open data repositories
	Public domain web pages

Next - Sumber data daring

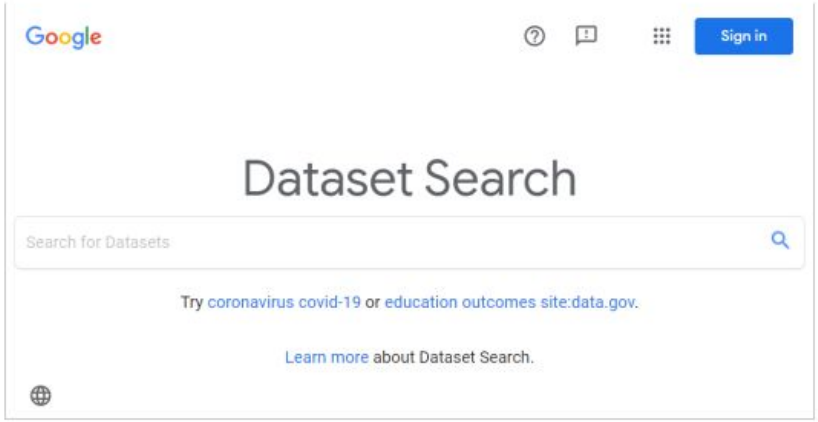
Sumber data daring

- Portal Satu Data Indonesia (https://data.go.id)
- Portal Data Jakarta (https://data.jakarta.go.id)
- Portal Data Bandung (http://data.bandung.go.id)
- Badan Pusat Statistik (https://www.bps.go.id)
- Badan Informasi Geospasial (https://tanahair.indonesia.go.id/)
- UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php)
- Kaggle (https://www.kaggle.com/datasets)
- World Bank Open Data (https://data.worldbank.org)
- UNICEF Data (https://data.unicef.org)
- WHO Open Data (https://www.who.int/data)
- IBM Data Asset eXchange (https://developer.ibm.com/exchanges/data/)
- DBPedia (https://www.dbpedia.org/resources/)
- Wikidata (https://www.wikidata.org/) .

Next - Sumber data daring

Sumber data daring

- Cari via Google Dataset Search: https://datasetsearch.research.google.com



Next - Susunan data

Susunan data

Butir data (*datum*): satuan terkecil data; satu nilai untuk satu variable tertentu

Data: kumpulan butir data yang membawa satu kesatuan makna (mendeskripsikan satu objek) tertentu.

Himpunan data (*dataset*): kumpulan data.

Metadata: data yang menjelaskan data yang lain.

symboling	normalized-losses	make	fuel-type
3	?	alfa-romero	gas
3	?	alfa-romero	gas
1	?	alfa-romero	gas
2	164	audi	gas
2	164	audi	gas

"make":
- tipe: string,
- deskripsi: nama pabrikan merek kendaraan

Next - Tipe data berdasar susunannya

Tipe data berdasarkan susunannya

	Data terstruktur (structured data)	Data takterstruktur (unstructured data)
Sifat	• Model data terdefinisikan sebelumnya • Format butir data (biasanya) teks. • Antar butir data terbedakan dengan jelas. • Ekstraksi/kueri langsung cukup mudah.	• Model data tidak terdefinisikan sebelumnya • Format butir data (biasanya) teks, citra, suara, video, dan format lainnya. • Antar butir data tidak cukup jelas terbedakan karena ketidakteraturan dan ambiguitas. • Ekstraksi/kueri langsung cukup sulit.
Contoh	Data tabular, data berorientasi objek, *time series*	Data teks dalam dokumen teks bebas, data audio, data video.

Data semi-terstruktur (*semi-structured data*): Data terstruktur yang tidak mengikuti model struktur tabular yang seperti pada basis data relasional, namun tetap mengandung *tags* atau penanda lainnya yang dapat memisahkan elemen-elemen semantik pada data serta mengatur hierarki antara butir-butir datanya.

Next - Tipe butir data

Tipe butir data (1)

	Nominal/kategorikal	Ordinal	Interval	Rasio
Sifat himpunan asal	Diskret, tidak terurut	Diskret, terurut	Kontinu/numerik, terurut, perbedaan menunjukkan selisih	Kontinu/numerik, terurut, nilai menunjukkan rasio terhadap kuantitas satuan/unit di jenis yang sama
Contoh	Warna (merah, hijau, biru)	Nilai huruf mahasiswa (A, B, C, D, E)	Suhu dalam Celcius, tanggal dalam kalender tertentu	Panjang jalan, suhu dalam Kelvin
Ukuran data menyatakan ...	Membership	Membership, comparison	Membership, comparison, difference	Membership, comparison, difference, magnitude
Operasi matematika	=, ≠	=, ≠, <, >	=, ≠, <, >, +, -	=, ≠, <, >, +, -, ×, ÷

Next - Tipe butir data

Tipe butir data (2)

	Nominal/kategorikal	Ordinal	Interval	Rasio
Representasi nilai tipikal	Modus	Modus, median	Modus, median, rerata aritmetis	Modus, median, rerata aritmetis, rerata geometris, rerata harmonis
Representasi sebaran	Grouping	Grouping, rentang (*range*), rentang antarkuartil	Grouping, rentang (*range*), rentang antarkuartil, varians, simpangan baku	Grouping, rentang (*range*), rentang antarkuartil, varians, simpangan baku, koefisien variasi
Memiliki nol sejati yang menyatakan nilai mutlak terbawah.	Tidak	Tidak	Tidak	Ya

Next - Contoh model data

Contoh model data: Tabular

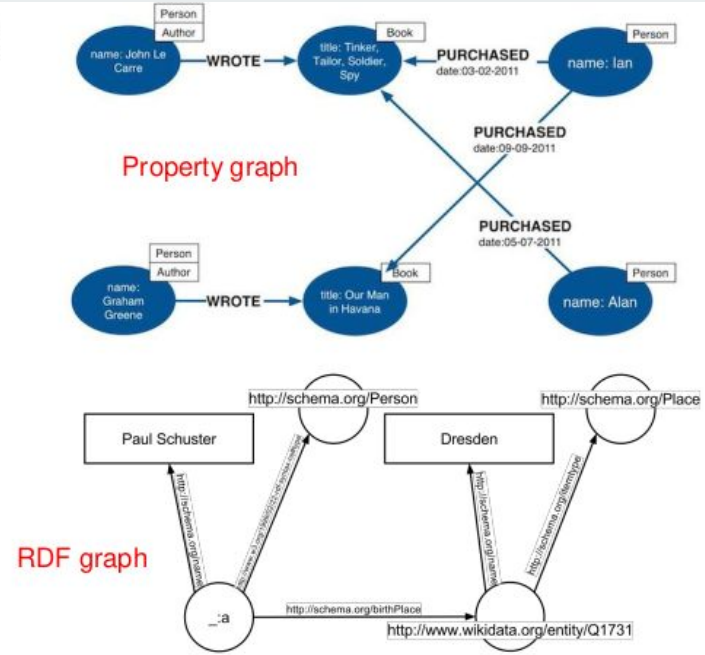
- Terdiri dari N buah rekord (*record*)
- Masing-masing rekord mengandung D buah atribut
- Rekord = baris, *data point*, instans, *example*, transaksi, tupel, entitas, objek, vector fitur.
- Atribut = kolom, *field*, dimensi, fitur.
- Atribut yang sama untuk setiap rekord biasanya diasumsikan memiliki tipe butir data yang sama.
- Struktur dapat bersifat ketat/strict (contoh: basis data relasional) atau longgar/loose (contoh: Excel *spreadsheet*).
- Tergantung keketatan strukturnya, bisa ada bahasa kueri formal untuk mengakses butir-butir data di dalamnya (contoh: SQL).

symboling	normalized-losses	make
3	?	alfa-romero
3	?	alfa-romero
1	?	alfa-romero
2	164	audi
2	164	audi

Next - Model data jejaring

Contoh model data: Graf/Jejaring

- Tersusun dari simpul-simpul (*nodes*) dan sisi/koneksi antar simpul (*edges*)
- Satu node (biasanya) mewakili satu rekord
- Dapat mengekspresikan relasi antar rekord secara eksplisit.
- Termasuk model data graf adalah model data hierarkis/pohon, model data berorientasi objek (*object-oriented data model*).
- Model data graf modern:
  - *Property graph*
  - *Resource description framework (RDF)*



Next - Model data sekuens

TERIMA KASIH