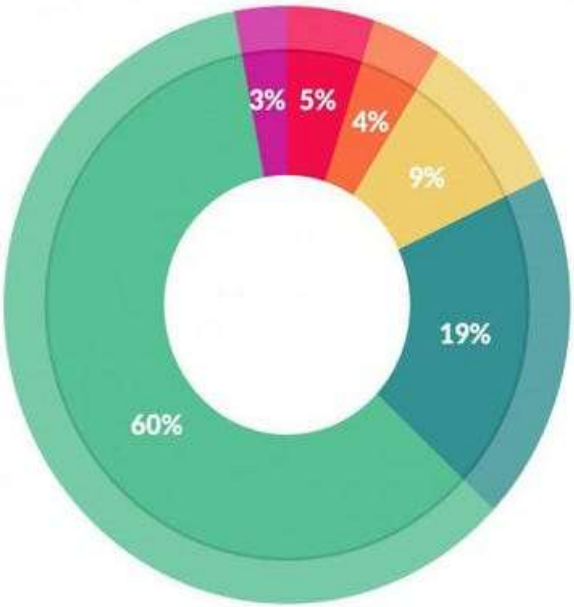
PERTEMUAN 6

Teknik Pemrosesan Data

Data Processing

Data preprocessing merupakan sekumpulan teknik yang diterapkan pada database untuk menghapus noise, missing value, dan data yang tidak konsisten atau dalam istilah lain adalah *Data Pre-processing, Data Manipulation, Data Cleansing/ Normalization.*

Fakta Terkait Data Preparation



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=690c8f796f63

Fakta Terkait Data Preparation

60-80% porsi kegiatan data saintis berfokus pada data preparation (forbes, crowdflower 2016) data yang ada saat ini dari banyak sumber data dan format yang beragam (terstruktur, semi, dan tidak terstruktur) kualitas model prediktif bergantung pada kualitas data (GIGO)

Pentingnya Data Preparation

Data perlu diformat sesuai dengan software yang digunakan

data perlu disesuaikan dengan metode data sains yang digunakan

data real-world cenderung ‘kotor’:

tidak komplit: kurangnya nilai attribute, kurangnya atribut tertentu/penting, hanya berisi data agregate. misal: pekerjaan=”” (tidak ada isian)

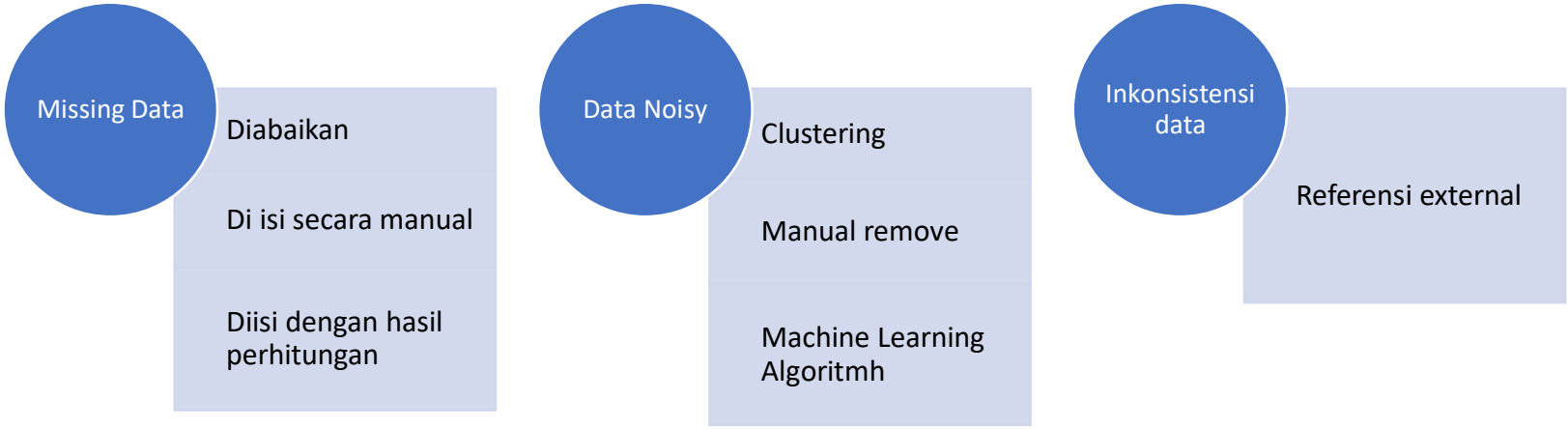
noisy: memiliki error atau outlier. misal: Gaji=“-10”, Usia=“222

Pentingnya Data Preparation (lanjut)

Data real-world cenderung ‘kotor’:

tidak konsisten: memliki perbedaan dalam kode dan nama. misal : Usia= “32” TglLahir=”03/07/2000”; rating “1,2,3” -- > rating “A, B, C”

kolom dan baris yang saling bertukar banyak variabel dalam satu kolom yang sama



Manfaat Data Preparation

Kompilasi Data menjadi Efisien danEfektif (menghindari duplikasi)

Identifikasi dan Memperbaiki Error

Mudah Perubahan Secara Global

Menghasilkan Informasi yang Akurat utk Pengambilan Keputusan

Nilai Bisnis dan ROI (Return on Ivestment) akan Meningkat

https://searchbusinessanalytics.techtarget.com/definition/data-preparation

Tantangan Data Preparation

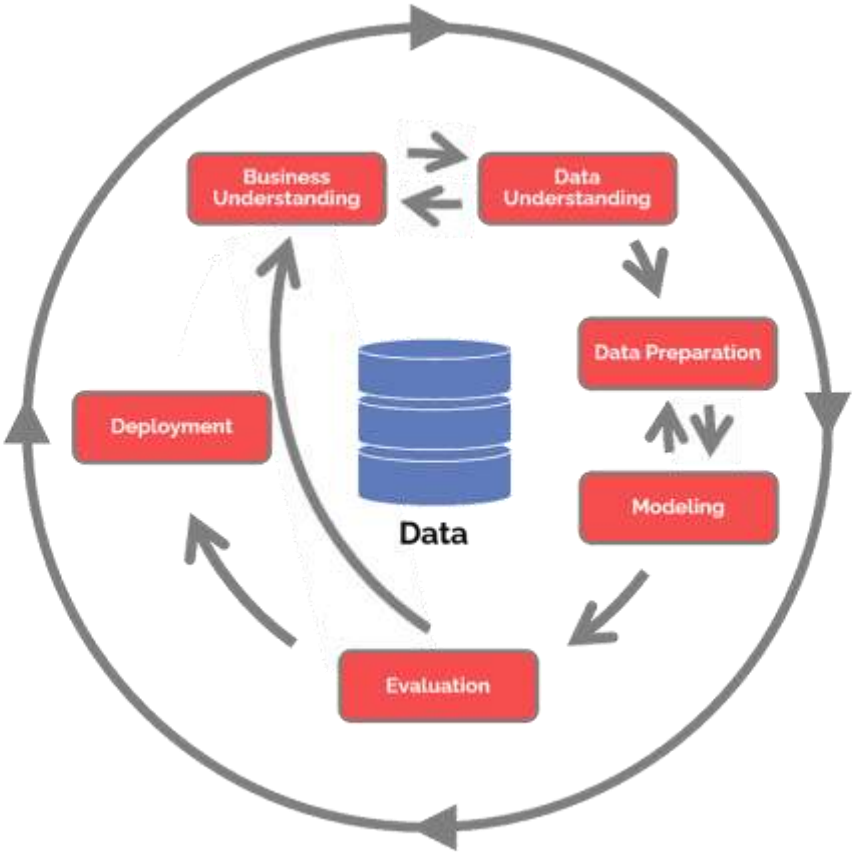
- Memakan Waktu Lama
- Porsi Teknis yang Dominan
- Data yang Tersedia Tidak Akurat atau Jelas/Tidak Langsung Pakai
- Data tidak Balance Saat Pengambilan Sampel
- Rentan akan Error

Data Preparation dalam CRISP-DM

Akronim dari: CRoss Industry Standard Process Data Mining

Metodologi umum untuk data mining, analitik, dan proyek data sains, berfungsi menstandarkan proses data mining lintas industry

Digunakan untuk semua level dari pemula hingga pakar



https://www.datascience-pm.com/crisp-dm-2/

Data Preparation Versi Sederhana

Data Collection

- Articulating the problem
- Defining data required
- Gathering and combining data from difernet source

Data Preprocessing

- Formating
- Cleaning
- Sampling

Data Transformation

- Sclaing/Normalizing
- Decomposition
- Aggregation

Pemilihan, Pembersihan & Validasi

Pilih/ Select Data

- Pertimbangkan apemilihan data
- Tentukan dataset yang akan digunakan
- Kumpulkan data tambahan yang sesuai (internal atau eksternal)
- Pertimbangkan penggunaan teknik pengambilan sampel
- Jelaskan mengapa data tertentu dimasukkan atau dikecualikan

Pemilihan, Pembersihan & Validasi

Bersihkan/ Clean Data

- Perbaiki, hapus atau abaikan noise
- Putuskan bagaimana menangani nilai-nilai khusus dan maknanya
- Tingkat agregasi, nilai yang hilang (missing value), dll
- Bersihkan atau manipulasi outlier

Pemilihan, Pembersihan & Validasi

Validasi Data

- Periksa/Nilai Kualitas Data
- Periksa/Nilai Tingkat Kecukupan Data

Rincian Tahapan Data Preparation

Bangun/ Construct Data

1. Atribut turunan.
2. Latar belakang pengetahuan.
3. Bagaimana atribut yang hilang dapat dibangun atau diperhitungkan

Integrasi/ Integrate Data

Mengintegrasikan sumber dan menyimpan hasil (tabel dan catatan baru)

Bentuk/ Format Data