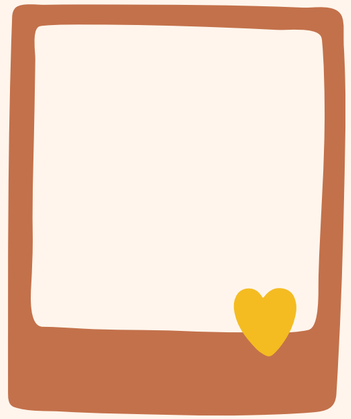
Getting to Know Your Data

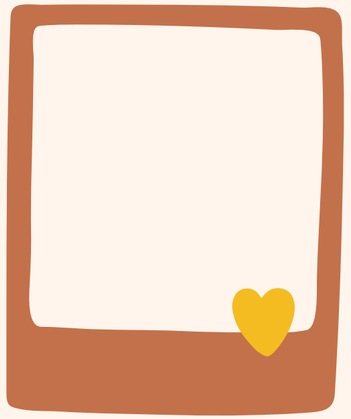
Presented by Asep Muhidin, S.Kom., M.Kom.,

START

Topic

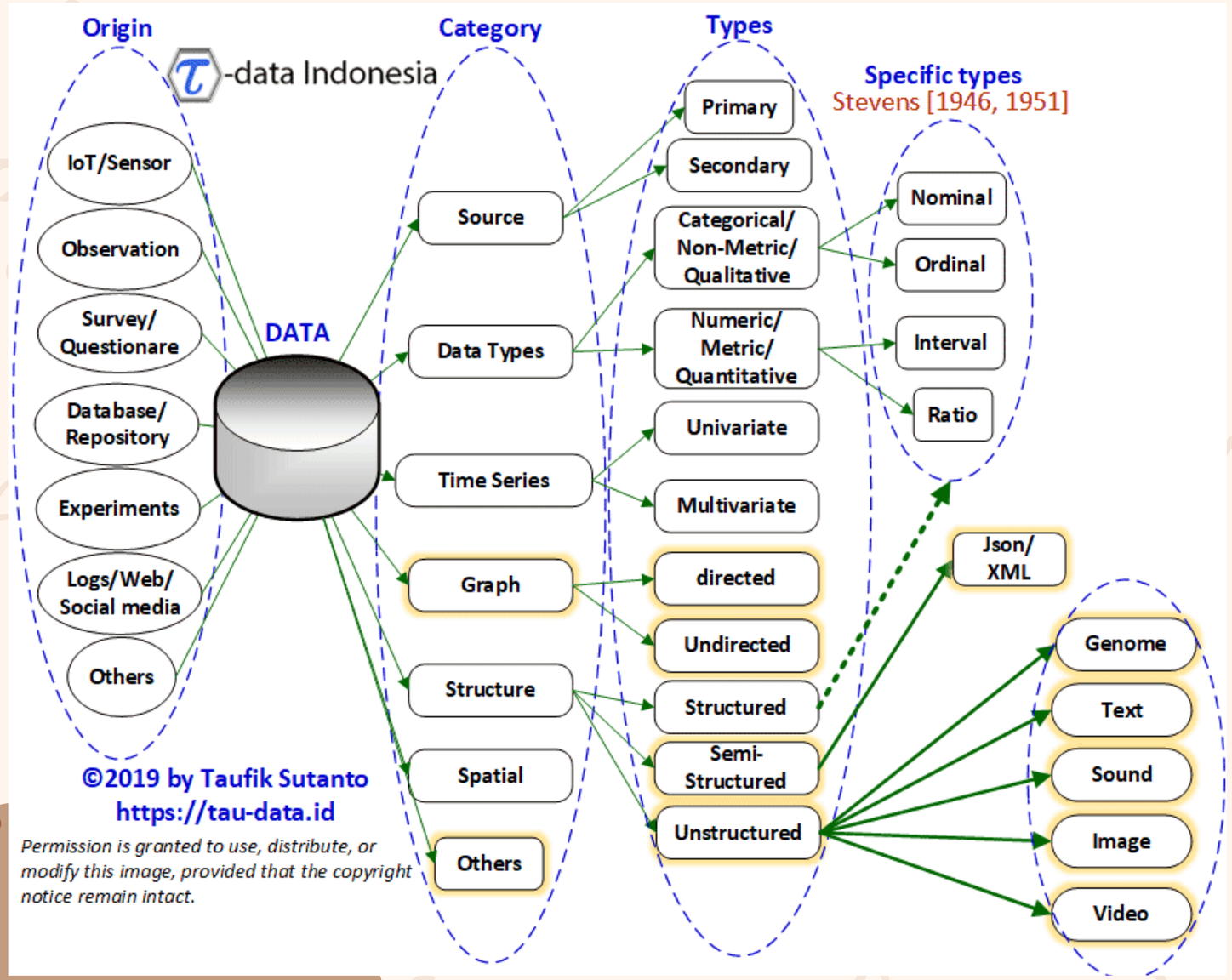


Data type



Data on Google Colaboratory

Asal, Jenis, dan Tipe Data.



Tipe Data Terstruktur

Kategorik/ Kualitatif/ Non-Metric

Numerik/ Kuantitatif/ Metric

Tipe Data Terstruktur

		Penjelasan	Contoh	Operasi	Visualisasi	Pengukuran
Kategorik/ Kualitatif/ Non-Metric	Nominal	Pemetaan/label, bukan pengukuran yang sesungguhnya dan tidak bermakna kuantitas, hanya pembeda	Jenis kelamin, agama, Negara, kode pos, warna	Modus, entropy	Pie, bar, bubble chart	Count/ Frekuensi
	Ordinal	Memiliki cukup informasi di data untuk mengurutkan, tapi tidak memiliki sifat selisih/interval dan perbandingan	Tingkat pendidikan, pangkat militer, rangking, nilai huruf	Rank correlation, run test, sign test	Bar, line chart dan mossaic plot	Frekuensi, Median, percintiles
Numerik/ Kuantitatif/ Metric	Interval	Numerik, tidak memiliki nol mutlak, memiliki sifat selisih, namun tidak memiliki sifat perbandingan	IQ/ EQ/ SQ, Nilai Toefl/ GRE/ TPA, Suhu (C&F)	Anova, regresi, uji T, korelasi pearson	Bar, line chart	Variance/ Std. Deviasi, median, percentiles, freq.distribution
	Rasio	Numerik, memiliki nol mutlak, dan dapat diperbandingkan	Berat, tinggi badan, gaji, umur	Uji F,T, anova, regresi, clustering	Scatter plot, line chart, histogram	Variance/ Std. Deviasi, median, percentiles, freq.distribution

Time Series Data (Runtun Waktu)

Beberapa data tertentu bergantung terhadap waktu, sebut saja pergerakan nilai mata uang (**kurs**)/harga **saham**, **suhu**/temperature udara di suatu daerah tertentu, atau data ***logs*** suatu website.

Saat nilai data di masa depan lebih banyak (dominan) hanya dipengaruhi dari nilai-nilainya di masa lalu, maka model-model runtun waktu **univariate** (satu peubah/variabel) seperti ARIMA (*Autoregressive Integrated Moving Average*) dapat digunakan.

Namun bila satu atau beberapa peubah yang bergantung waktu dipengaruhi juga oleh variable lain selain nilai-nilainya di masa lalu, maka model runtun waktu peubah ganda (**multivariate**) seperti VaR (*Vector autoRegression*) dapat digunakan.

Date	Time	CO(GT)	PT08.S1(CO)
3/10/2004	18:00:00	2.6	1360
3/10/2004	19:00:00	2	1292
3/10/2004	20:00:00	2.2	1402
3/10/2004	21:00:00	2.2	1376
3/10/2004	22:00:00	1.6	1272
3/10/2004	23:00:00	1.2	1197
3/11/2004	0:00:00	1.2	1185
3/11/2004	1:00:00	1	1136
3/11/2004	2:00:00	0.9	1094
3/11/2004	3:00:00	0.6	1010
3/11/2004	4:00:00	-200	1011
3/11/2004	5:00:00	0.7	1066
3/11/2004	6:00:00	0.7	1052
3/11/2004	7:00:00	1.1	1144
3/11/2004	8:00:00	2	1333

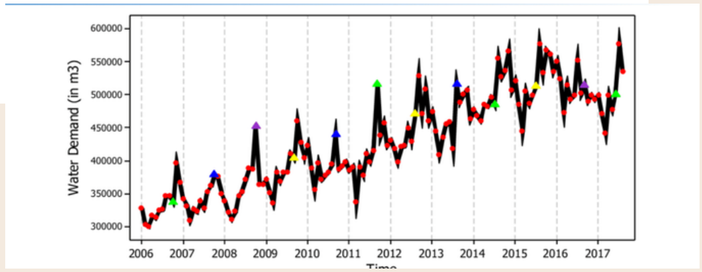
Time Series Data (Runtun Waktu)

Beberapa data tertentu bergantung terhadap waktu, sebut saja pergerakan nilai mata uang (**kurs**)/harga **saham**, **suhu**/temperature udara di suatu daerah tertentu, atau data ***logs*** suatu website.

Saat nilai data di masa depan lebih banyak (dominan) hanya dipengaruhi dari nilai-nilainya di masa lalu, maka model-model runtun waktu **univariate** (satu peubah/variabel) seperti ARIMA (*Autoregressive Integrated Moving Average*) dapat digunakan.

Namun bila satu atau beberapa peubah yang bergantung waktu dipengaruhi juga oleh variable lain selain nilai-nilainya di masa lalu, maka model runtun waktu peubah ganda (**multivariate**) seperti VaR (*Vector autoRegression*) dapat digunakan.

Date	Time	CO(GT)	PT08.S1(CO)
3/10/2004	18:00:00	2.6	1360
3/10/2004	19:00:00	2	1292
3/10/2004	20:00:00	2.2	1402
3/10/2004	21:00:00	2.2	1376
3/10/2004	22:00:00	1.6	1272
3/10/2004	23:00:00	1.2	1197
3/11/2004	0:00:00	1.2	1185
3/11/2004	1:00:00	1	1136
3/11/2004	2:00:00	0.9	1094
3/11/2004	3:00:00	0.6	1010
3/11/2004	4:00:00	-200	1011
3/11/2004	5:00:00	0.7	1066
3/11/2004	6:00:00	0.7	1052
3/11/2004	7:00:00	1.1	1144
3/11/2004	8:00:00	2	1333



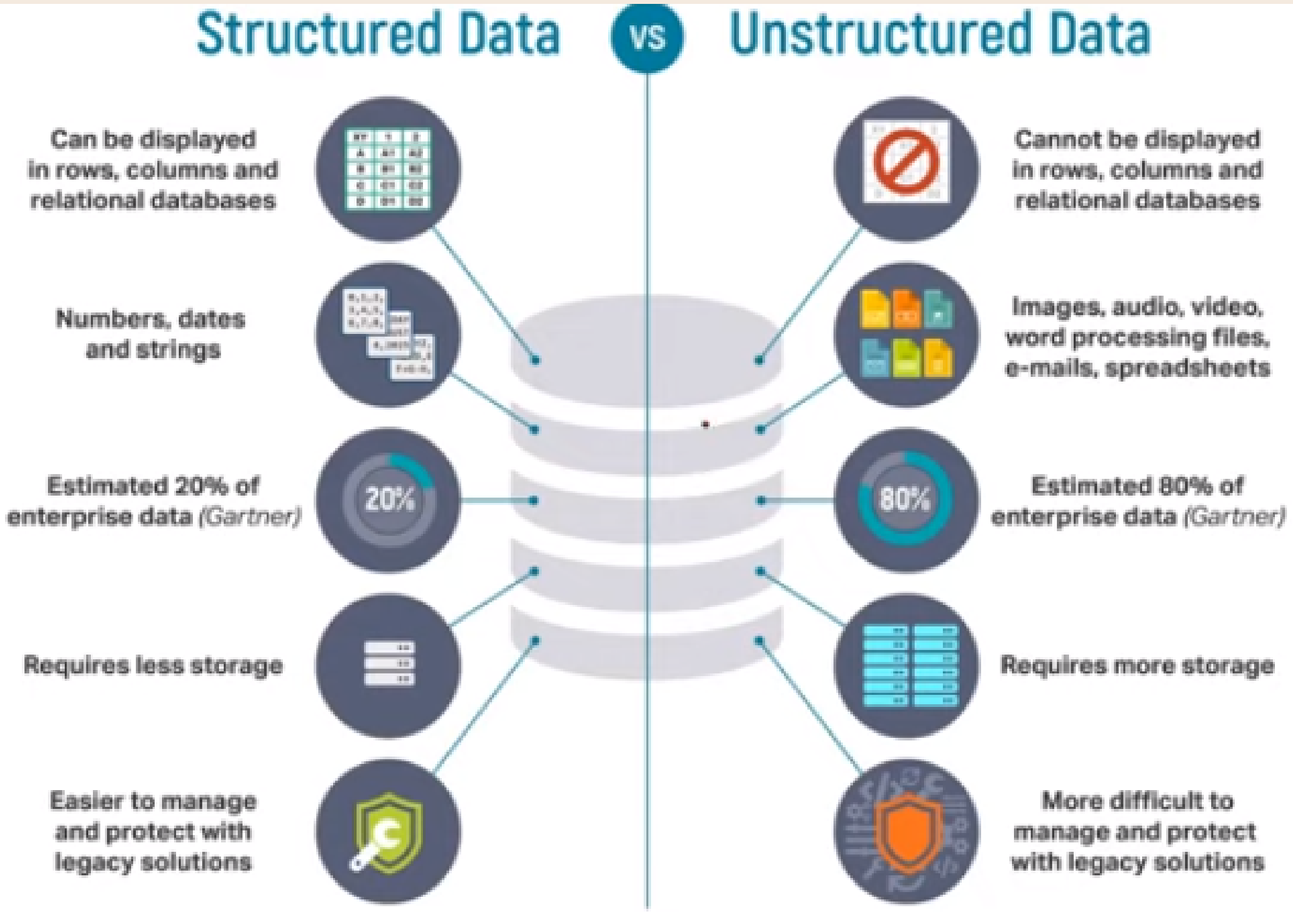
(Geo) Spatial Data

Ada kalanya penelitian yang dilakukan bergantung pada lokasi/tempat, sebut saja penelitian yang berkenaan dengan kadar **mineral/gas** di suatu daerah tertentu, penelitian tentang **penyebaran** suatu **penyakit** menular tertentu (misal: flu burung dan HIV/AIDS), **gempa** bumi, atau penelitan tentang dukungan **politik** di suatu daerah tertentu.

Saat datanya bergantung pada lokasi (**Spatial**) maka model-model statistik Spatial Data Analysis seperti *spatial autocorrelation, spatial interpolation, spatial regression, spatial interaction, dan multiple-point geostatistics* dapat digunakan. Terkait dengan data Spatial dan data mining (machine learning) akhir-akhir ini terdapat topik baru yang cukup menarik: "Geospatial Intelligence".

publicid	eventtype	origintime	modificat	longitude	latitude	magnitude	depth
2015p717507		2015-09-2	2015-09-2	176.0944492	-38.50245621	2.459332875	149.375
2015p717354		2015-09-2	2015-09-2	178.4734322	-38.25412784	1.987484953	28.90625
2015p717280		2015-09-2	2015-09-2	176.1695696	-38.46475897	2.456398653	153.125
2015p717262		2015-09-2	2015-09-2	177.4477559	-37.69434544	2.172615393	52.8125
2015p717174		2015-09-2	2015-09-2	172.4038845	-43.61736644	2.221402972	9.921875
2015p717142		2015-09-2	2015-09-2	176.5673243	-37.85364822	2.381214887	90.3125
2015p717	earthquak	2015-09-2	2015-09-2	175.6586602	-39.26602063	0.784917104	67.8125
2015p717090		2015-09-2	2015-09-2	174.8915884	-41.11647523	2.882207598	30.07813
2015p717068		2015-09-2	2015-09-2	176.0801814	-39.96313712	1.609327328	24.45313
2015p717018		2015-09-2	2015-09-2	175.7615368	-38.67061338	3.457217946	142.3438
2015p716785		2015-09-2	2015-09-2	174.6997467	-39.30605315	2.33739074	24.92188
2015p716768		2015-09-2	2015-09-2	174.6949536	-39.29756737	2.807979919	24.92188
2015p716752		2015-09-2	2015-09-2	174.773634	-39.17986966	2.502129961	27.73438
2015p716720		2015-09-2	2015-09-2	174.8950946	-41.11324348	2.818673726	29.60938
2015p716649		2015-09-2	2015-09-2	176.9622346	-39.76140542	1.781550321	8.515625
2015p716596		2015-09-2	2015-09-2	174.7270089	-41.17664948	1.834855653	27.26563
2015p716	earthquak	2015-09-2	2015-09-2	175.9976775	-39.27133327	1.024882054	43.90625
2015p716366		2015-09-2	2015-09-2	176.1486309	-39.05088715	1.315958386	48.125
2015p716	earthquak	2015-09-2	2015-09-2	176.0763517	-38.63151112	1.911692052	5.820313
2015p716257		2015-09-2	2015-09-2	179.3621839	-45.28313588	2.628057845	6.40625

Data Terstruktur vs Data Tidak Terstruktur



https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

- **Data Terstruktur** biasanya mengacu pada data dalam format Baris-Kolom dan memiliki dimensi yang tetap.
  - Data ini biasanya disimpan dalam format table-table yang saling berelasi/hubungan (biasa disebut relational database).
  - Berbentuk numerik, atau kategorik yang dapat di “encode” dengan mudah melalui pemasangan satu-satu. Misal Pria:1, dan Wanita:0.
- **Data Tidak Terstruktur** adalah negasi dari definisi diatas.

Colab Workshop

- Buat notebook baru di google colab
- Akses dataset yang ada di repo
  import pandas as pd
  dataset=”https://raw.githubusercontent.com/asepmuhidin/MK-Data-Mining-UPB/main/dataset/cereal.csv”
  df=pd_read_csv(dataset)
- Lihat 10 data pertama : df.head(10)
- Lihat 10 data terakhir :df.tail(10)
- lihat 10 data acak : df.sample(10)
- Sebutkan mana yang termasuk data categorical atau data numeric

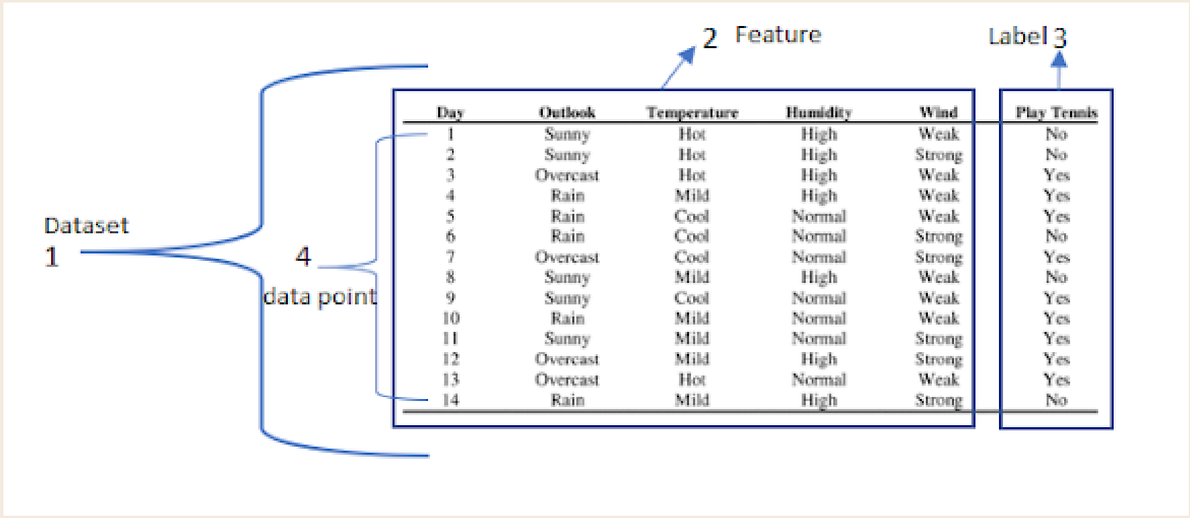
Field/Attribute/Fiture	Tipe Data

Colab Workshop

- Menampilkan nama kolom : df.columns
- Lihat nama kolom dan tipe data
  df.info()
- Melihat tipe data tiap kolom : df[col].dtype
- Lakukan kembali dengan menggunakan looping: for col in df.columns:print(df[col].dtype)
- Merubah tipe data object menjadi category : df[cols] = df[cols].astype('category')
- Lihat kembali tipe data yang sudah dirubah : df.info()
- Kelompokkan kolom menjadi features /attributes dan label/kelas/target: X,y

  X=df.drop(columns=’rating’)

  y=df.rating



Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- Dari kolom fitur Kelompokkan menjadi 2 kelompok data yaitu data numeric dan kategori

Q

A

Question Time

Thank You