Tipe dan Jenis Data

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

October 3, 2017

Selayang Pandang

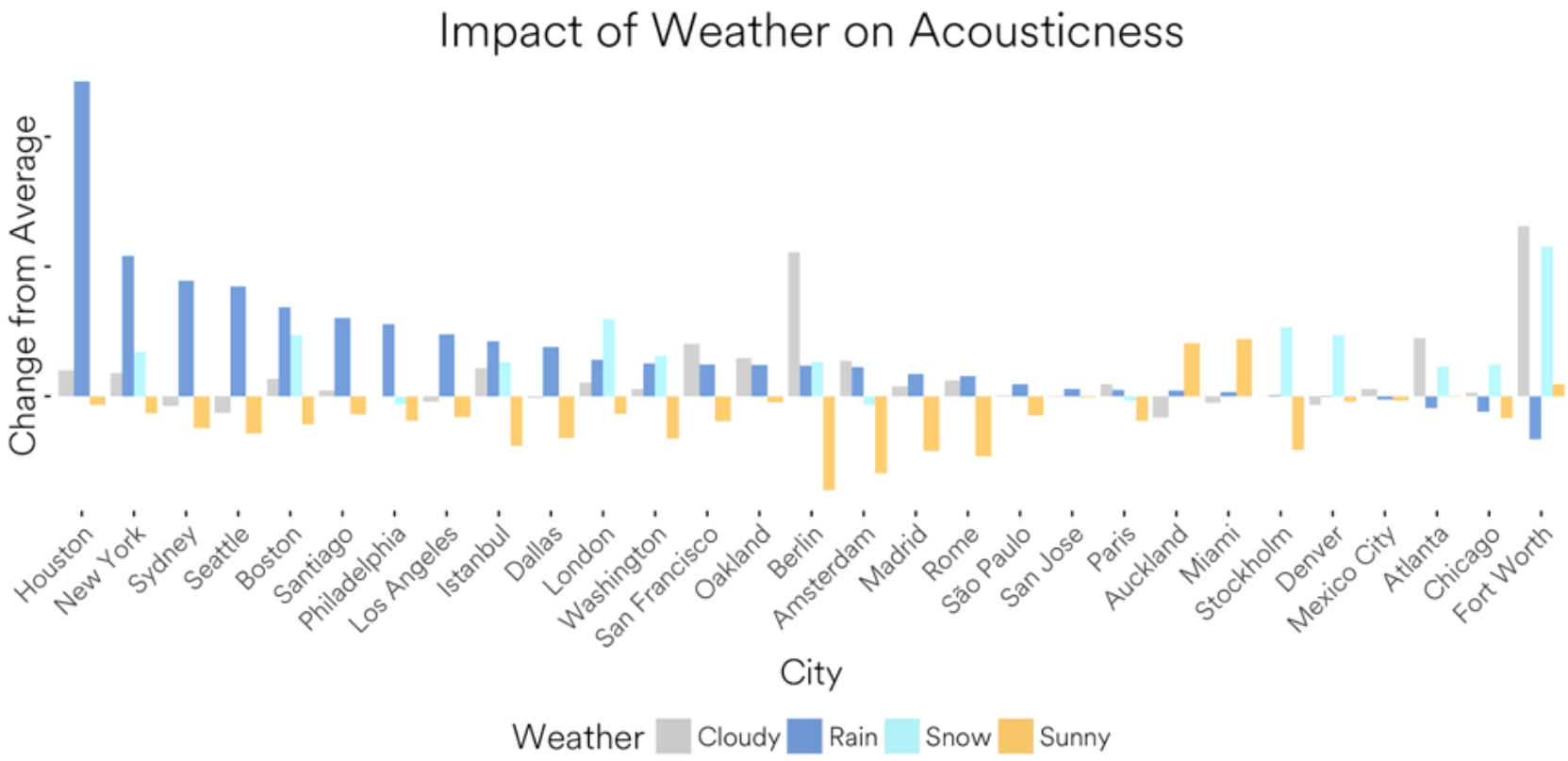
Pengumuman

1. Segera *enroll* untuk kuliah Data Mining di e-learning!
2. Pengumpulan tugas akan dilakukan melalui e-learning
3. Ada baiknya untuk mengecek `http://www.inf.ed.ac.uk/teaching/courses/it/cribsheet.2up.pdf` karena akan mulai dipakai dalam waktu dekat
4. Libur Nyepi di minggu ke-6, jadi materi minggu ke-4 akan dipadatkan ke minggu depan
5. Siapkan Python, Pandas, dan Jupyter Notebook!

Representasi Data

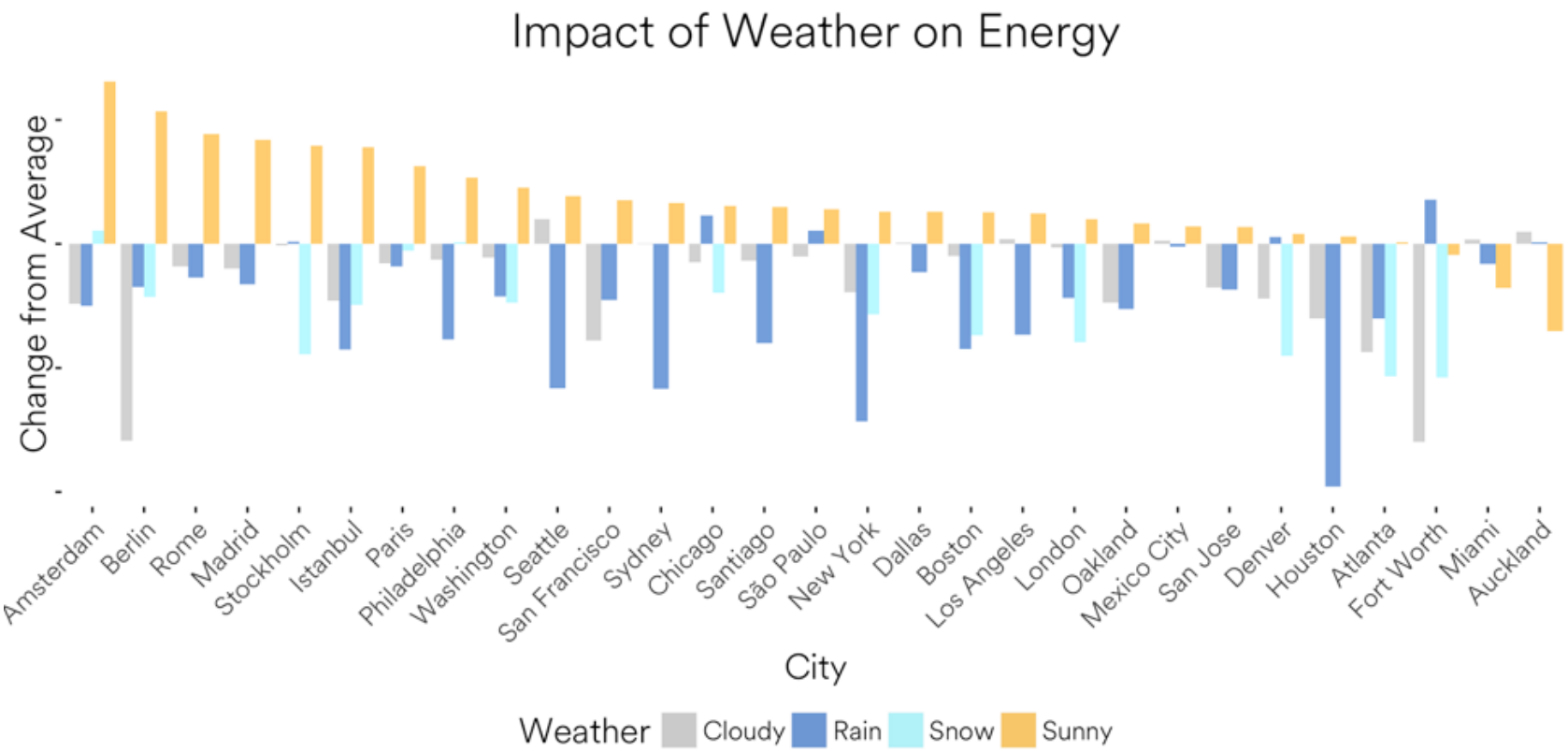
Variabel seperti apa yang dapat dipakai oleh
sistem rekomendasi berdasarkan konten dari aplikasi seperti Spotify?

Korelasi pada Spotify



Gambar: Hubungan cuaca dengan "keakustikan" musik [van Buskirk, 2017] yang dilihat dari bunyi-bunyi alat akustik, e.g. gitar dan tamborin, dibandingkan dengan bunyi-bunyi elektronik, e.g. *synthesizer*

Korelasi pada Spotify



Gambar: Hubungan cuaca dengan "energi" musik [van Buskirk, 2017] yang dilihat dari kecepatan, volume, dan kebisingan, misalnya perbandingan kontras antara musik *death metal* dan komposisi Bach

Data, Atribut, dan Objek

Data

Data merupakan kumpulan objek (*instances*) yang memiliki atribut-atribut tertentu

Data, Atribut, dan Objek

Data

Data merupakan kumpulan objek (*instances*) yang memiliki atribut-atribut tertentu

Atribut

Karakteristik dari suatu objek, dikenal juga dengan nama **variabel** atau **fitur**

Data, Atribut, dan Objek

Data

Data merupakan kumpulan objek (*instances*) yang memiliki atribut-atribut tertentu

Atribut

Karakteristik dari suatu objek, dikenal juga dengan nama **variabel** atau **fitur**

Objek

Dikenal juga dengan nama **record**, **poin**, **sampel**, **entitas**, atau **instance**

Dari contoh kasus Spotify tadi, mana yang merupakan atributnya dan mana yang merupakan objeknya?

Nilai dan Tipe dari Atribut

1. Nilai dari suatu atribut dapat berupa simbol maupun angka

Nilai dan Tipe dari Atribut

1. Nilai dari suatu atribut dapat berupa simbol maupun angka
2. Atribut yang sama dapat dipetakan ke beberapa nilai yang berbeda, misalnya karena beda satuan

Nilai dan Tipe dari Atribut

1. Nilai dari suatu atribut dapat berupa simbol maupun angka
2. Atribut yang sama dapat dipetakan ke beberapa nilai yang berbeda, misalnya karena beda satuan
3. Ada tiga tipe atribut secara umum: **categorical**/**nominal**, **ordinal**, **numeric**

Atribut Nominal

1. Atribut nominal bernilai saling lepas (*mutually exclusive*)
2. Perbandingan yang dapat dilakukan hanya menguji kesamaan ($=, \neq$)
3. Tidak dapat diurutkan maupun diukur jaraknya
4. Contoh: Warna mata, *genre* musik, pekerjaan

Atribut Ordinal

1. Terdapat urutan yang ada secara natural, e.g. $\{$kecil, sedang, besar$\}$ atau $\{$tidak suka, netral, suka$\}$
2. Dikodekan sebagai angka untuk mempertahankan urutan sehingga dapat dibandingkan $(<, =, >)$
3. Terkadang sulit untuk dibedakan dengan nominal, e.g. apakah ada urutan untuk $\{$belum menikah, menikah, bercerai$\}$?

Atribut Numerik

1. Dapat bernilai bulat atau riil sehingga bisa dijumlahkan atau dirata-rata
2. Sensitif terhadap nilai ekstrem, e.g. tinggi: {165,171,182,1850}
3. Terkadang dibedakan sebagai **ratio** dan **interval**

Ratio vs Interval

Ratio

Punya referensi nilai nol, e.g. berat, tinggi, jarak, suhu dalam Kelvin

Ratio vs Interval

Ratio

Punya referensi nilai nol, e.g. berat, tinggi, jarak, suhu dalam Kelvin

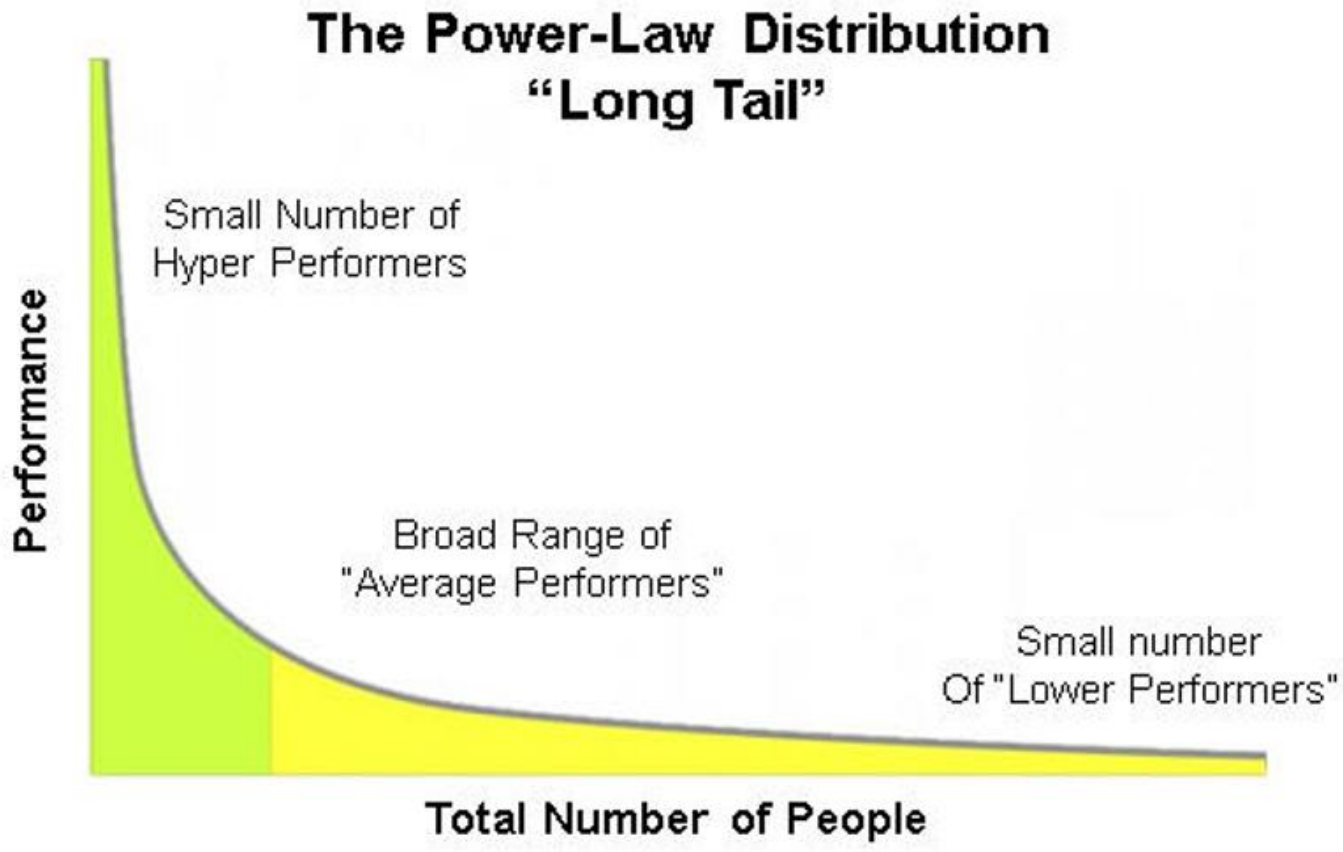
Interval

Tidak punya referensi nilai nol, e.g. suhu dalam Celsius atau Fahrenheit, tahun

Kasus dalam Atribut Numerik

1. Distribusi yang memiliki kecondongan, e.g. *power law distribution*
2. Efek non-monotonik dari atribut, e.g. usia dalam menentukan pemenang marathon
3. Terkadang perlu dilakukan normalisasi (berpusat di nol atau [0, 1])

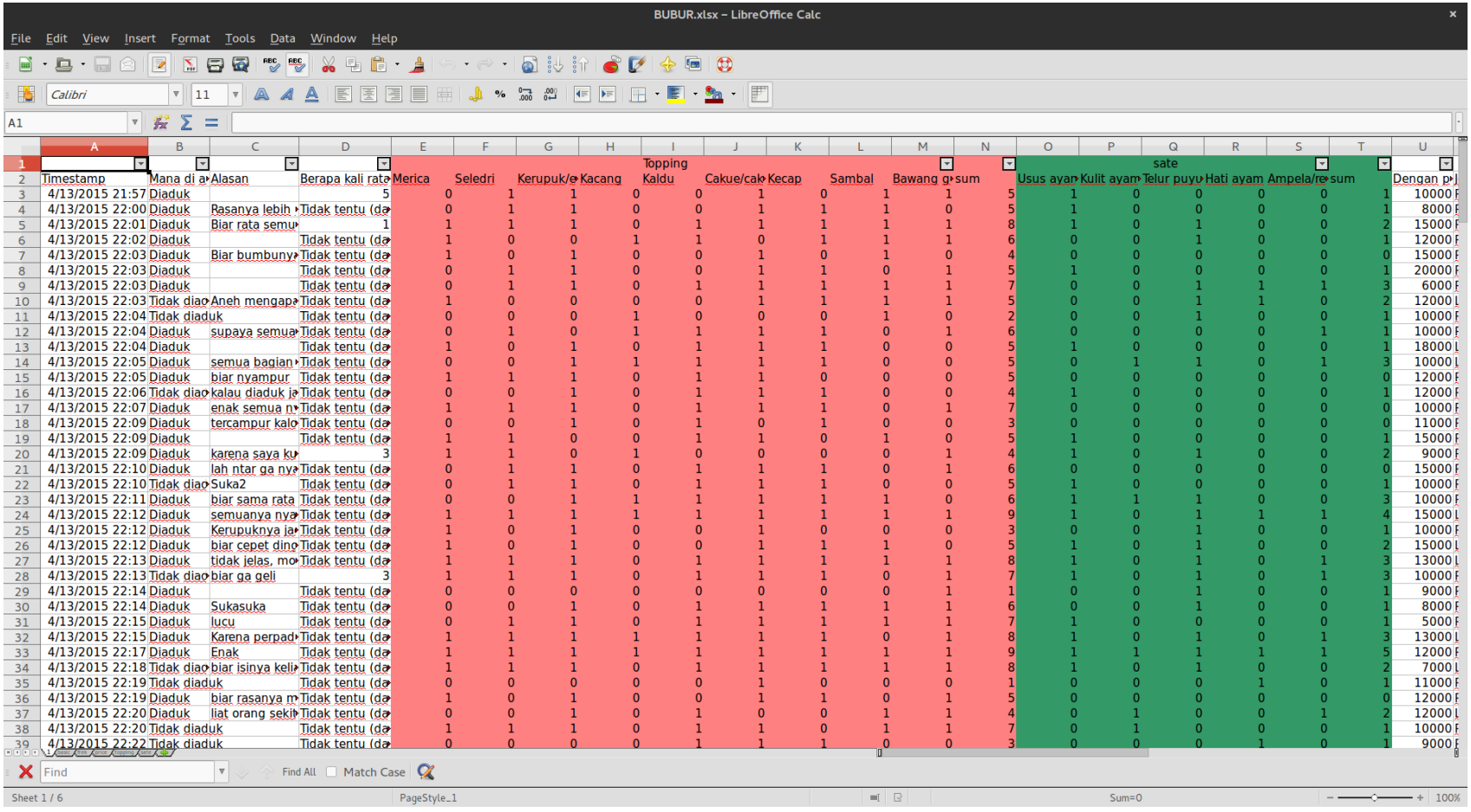
Distribusi yang Condong



Gambar: *Power law distribution* [Bersin, 2014]

Tipe Data

Data Matriks


BUBUR.xlsx – LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1									Topping								sate				
2	Timestamp	Mana di a	Alasan	Berapa kali rata	Merica	Seledri	Kerupuk/e	Kacang	Kaldu	Cakue/cak	Kecap	Sambal	Bawang g	sum	Usus ayan	Kulit ayam	Telur puyu	Hati ayam	Ampela/re	sum	Dengan p
3	4/13/2015 21:57	Diaduk		5	0	1	1	0	0	1	0	1	1	5	1	0	0	0	0	1	10000
4	4/13/2015 22:00	Diaduk	Rasanya lebih	Tidak tentu (da	0	1	1	0	0	1	1	1	0	5	1	0	0	0	0	1	8000
5	4/13/2015 22:01	Diaduk	Biar rata semu	1	1	1	1	0	1	1	1	1	1	8	1	0	1	0	0	2	15000
6	4/13/2015 22:02	Diaduk		Tidak tentu (da	1	0	0	1	1	0	1	1	1	6	0	0	1	0	0	1	12000
7	4/13/2015 22:03	Diaduk	Biar bumbuny	Tidak tentu (da	1	0	1	0	0	1	0	1	0	4	0	0	0	0	0	0	15000
8	4/13/2015 22:03	Diaduk		Tidak tentu (da	0	1	1	0	0	1	1	0	1	5	1	0	0	0	0	1	20000
9	4/13/2015 22:03	Diaduk		Tidak tentu (da	0	1	1	0	1	1	1	1	1	7	0	0	1	1	1	3	6000
10	4/13/2015 22:03	Tidak diaduk	Aneh mengapa	Tidak tentu (da	1	0	0	0	0	1	1	1	1	5	0	0	1	1	0	2	12000
11	4/13/2015 22:04	Tidak diaduk		Tidak tentu (da	0	0	0	1	0	0	0	1	0	2	0	0	1	0	0	1	10000
12	4/13/2015 22:04	Diaduk	supaya semua	Tidak tentu (da	0	1	0	1	1	1	1	0	1	6	0	0	0	0	1	1	10000
13	4/13/2015 22:04	Diaduk		Tidak tentu (da	1	0	1	0	1	1	1	0	0	5	1	0	0	0	0	1	18000
14	4/13/2015 22:05	Diaduk	semua bagian	Tidak tentu (da	0	0	1	1	1	1	1	0	0	5	0	1	1	0	1	3	10000
15	4/13/2015 22:05	Diaduk	biar nyampur	Tidak tentu (da	1	1	1	0	1	1	0	0	0	5	0	0	0	0	0	0	12000
16	4/13/2015 22:06	Tidak diaduk	kalau diaduk ja	Tidak tentu (da	0	0	1	0	1	1	1	0	0	4	1	0	0	0	0	1	12000
17	4/13/2015 22:07	Diaduk	enak semua n	Tidak tentu (da	1	1	1	0	1	1	1	0	1	7	0	0	0	0	0	0	10000
18	4/13/2015 22:09	Diaduk	tercampur kalo	Tidak tentu (da	0	0	1	0	1	0	1	0	0	3	0	0	0	0	0	0	11000
19	4/13/2015 22:09	Diaduk		Tidak tentu (da	1	1	0	0	1	1	0	1	0	5	1	0	0	0	0	1	15000
20	4/13/2015 22:09	Diaduk	karena saya ku	3	1	1	0	1	0	0	0	0	1	4	1	0	1	0	0	2	9000
21	4/13/2015 22:10	Diaduk	lah ntar ga nya	Tidak tentu (da	0	1	1	0	1	1	1	0	1	6	0	0	0	0	0	0	15000
22	4/13/2015 22:10	Tidak diaduk	Suka2	Tidak tentu (da	0	1	1	0	1	1	1	0	0	5	1	0	0	0	0	1	10000
23	4/13/2015 22:11	Diaduk	biar sama rata	Tidak tentu (da	0	0	1	1	1	1	1	1	0	6	1	1	1	0	0	3	10000
24	4/13/2015 22:12	Diaduk	semuanya nya	Tidak tentu (da	1	1	1	1	1	1	1	1	1	9	1	0	1	1	1	4	15000
25	4/13/2015 22:12	Diaduk	Kerupuknya ja	Tidak tentu (da	1	0	1	0	0	1	0	0	0	3	0	0	1	0	0	1	10000
26	4/13/2015 22:12	Diaduk	biar cepet ding	Tidak tentu (da	1	0	1	0	0	1	1	1	0	5	1	0	1	0	0	2	15000
27	4/13/2015 22:13	Diaduk	tidak jelas, mo	Tidak tentu (da	1	1	1	0	1	1	1	1	1	8	1	0	1	0	1	3	13000
28	4/13/2015 22:13	Tidak diaduk	biar ga geli	3	1	1	1	0	1	1	1	0	1	7	1	0	1	0	1	3	10000
29	4/13/2015 22:14	Diaduk		Tidak tentu (da	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	9000
30	4/13/2015 22:14	Diaduk	Sukasuka	Tidak tentu (da	0	0	1	0	1	1	1	1	1	6	0	0	1	0	0	1	8000
31	4/13/2015 22:15	Diaduk	lucu	Tidak tentu (da	0	1	1	0	1	1	1	1	1	7	1	0	0	0	0	1	5000
32	4/13/2015 22:15	Diaduk	Karena perpad	Tidak tentu (da	1	1	1	1	1	1	1	0	1	8	1	0	1	0	1	3	13000
33	4/13/2015 22:17	Diaduk	Enak	Tidak tentu (da	1	1	1	1	1	1	1	1	1	9	1	1	1	1	1	5	12000
34	4/13/2015 22:18	Tidak diaduk	biar isinya keli	Tidak tentu (da	1	1	1	0	1	1	1	1	1	8	1	0	1	0	0	2	7000
35	4/13/2015 22:19	Tidak diaduk		Tidak tentu (da	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	1	11000
36	4/13/2015 22:19	Diaduk	biar rasanya m	Tidak tentu (da	1	0	1	0	0	1	1	0	1	5	0	0	0	0	0	0	12000
37	4/13/2015 22:20	Diaduk	liat orang seki	Tidak tentu (da	0	0	1	0	1	0	0	1	1	4	0	1	0	0	1	2	12000
38	4/13/2015 22:20	Tidak diaduk		Tidak tentu (da	1	1	1	0	1	1	0	1	1	7	0	1	0	0	0	1	10000
39	4/13/2015 22:22	Tidak diaduk		Tidak tentu (da	0	0	0	0	1	1	1	0	0	3	0	0	0	1	0	1	9000

Find Find All Match Case

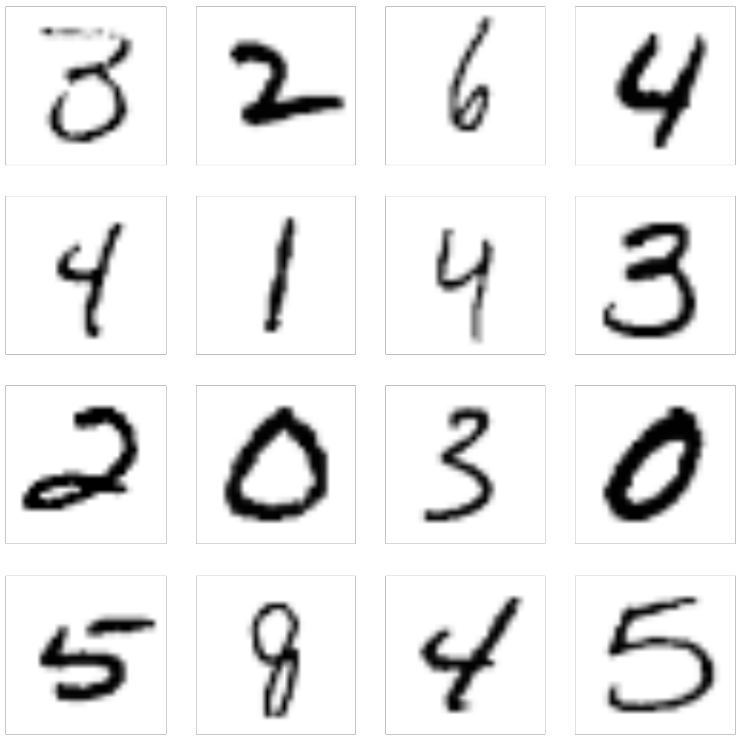
Sheet 1 / 6 PageStyle_1 Sum=0 100%

Gambar: Data preferensi bubur ayam

Data Matriks

1. Bentuk data paling sederhana
2. Sudah siap diolah
3. Dikenal juga sebagai **data terstruktur**
4. Contoh lain: data transaksi, hasil penapisan verbal

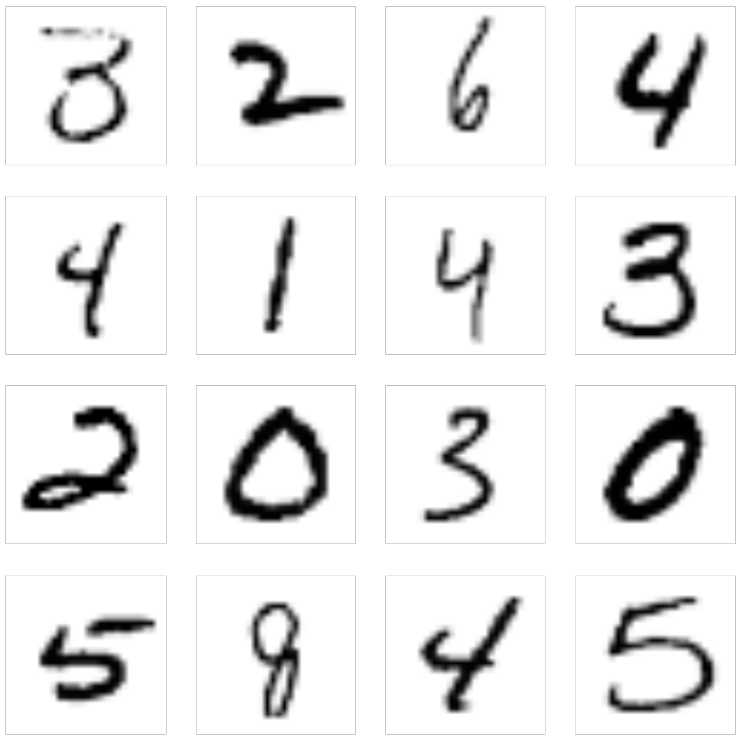
Gambar



Gambar: Contoh data MNIST [O'Shea, 2016]

1. Bagaimana cara merepresentasikan gambar?

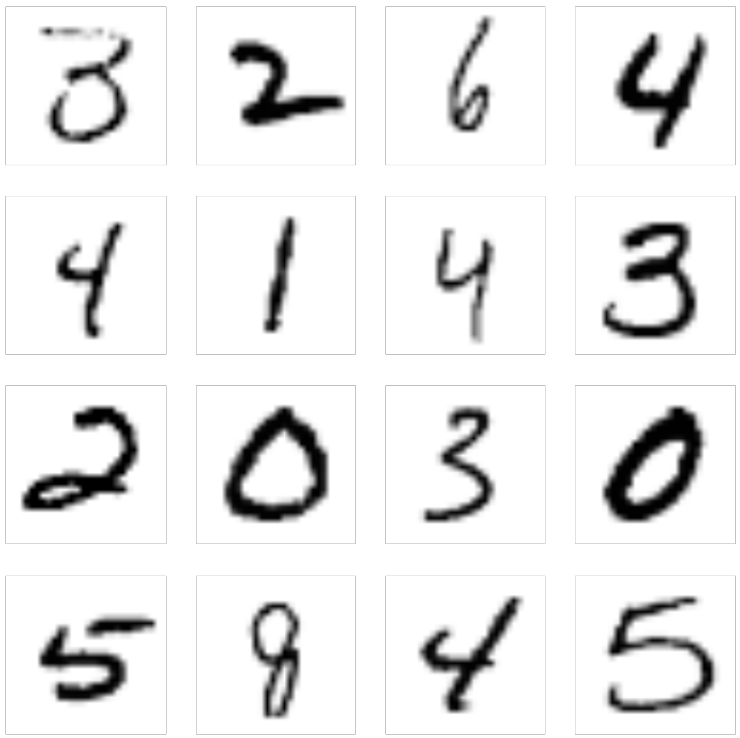
Gambar



Gambar: Contoh data MNIST [O'Shea, 2016]

1. Bagaimana cara merepresentasikan gambar?
2. Jika tiap pixel adalah atribut, berapa nilainya yang mungkin?

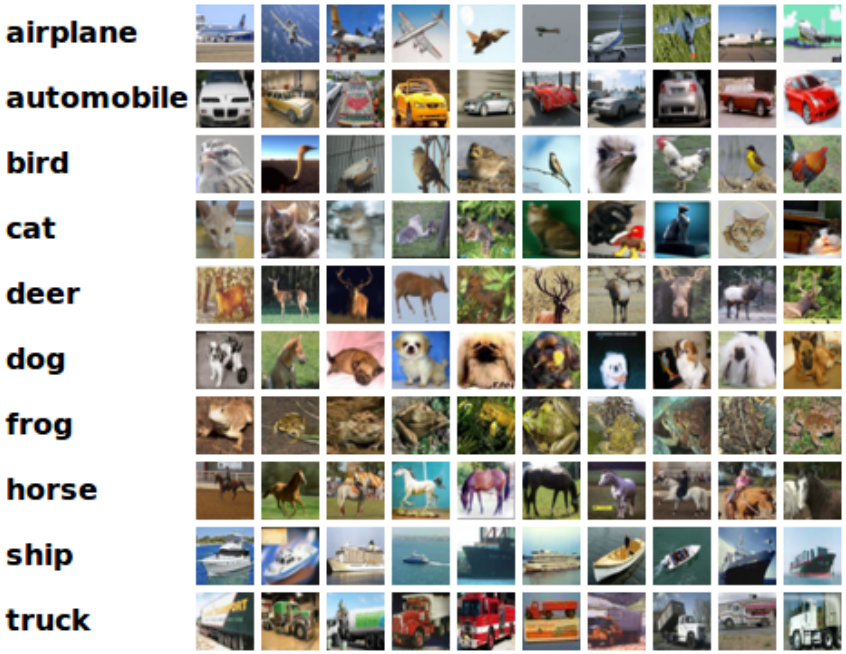
Gambar



Gambar: Contoh data MNIST [O'Shea, 2016]

1. Bagaimana cara merepresentasikan gambar?
2. Jika tiap pixel adalah atribut, berapa nilainya yang mungkin?
3. Apa kelebihan dan kekurangannya?

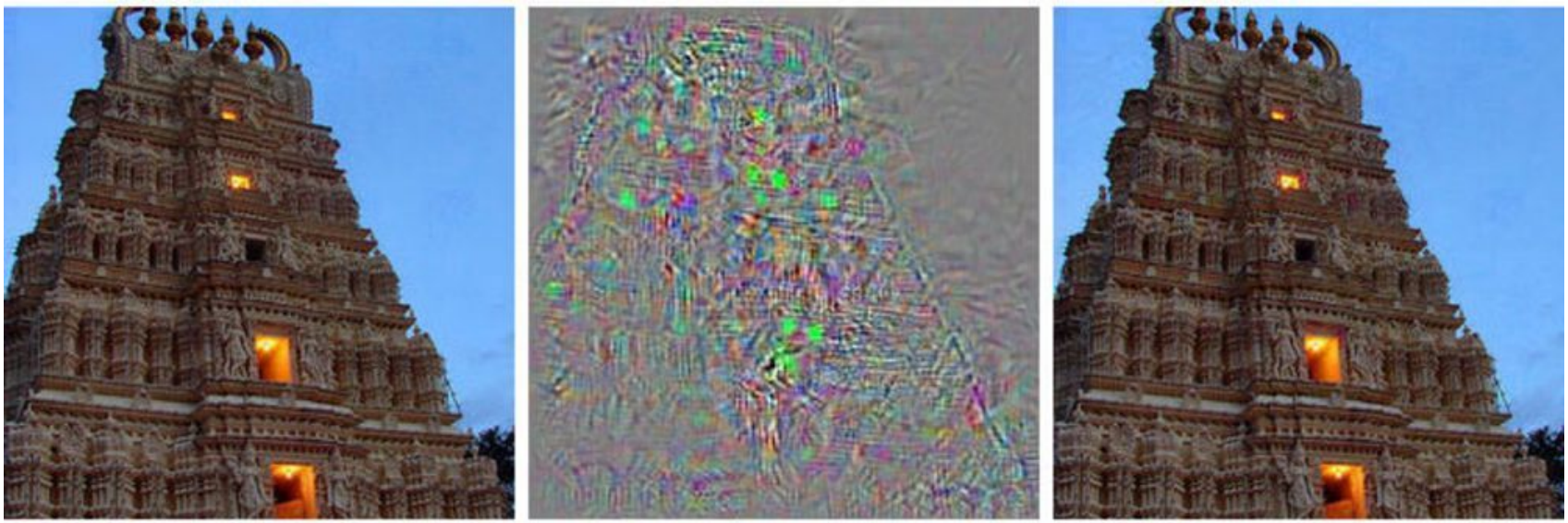
Gambar: Object Recognition



Gambar: Dataset CIFAR-10
[Krizhevsky, 2009]

1. Bagaimana dengan prediksi objek?
2. Tantangan: orientasi, skala, pencahayaan
3. Menggunakan pixels saja (mungkin) tidak cukup!
4. Bisa dibagi berdasarkan "region"

Misklasifikasi dalam Pengenalan Objek



Gambar: Gedung yang dianggap sebagai burung unta setelah diterapkan *noise*

Teks

Contoh tugas:

1. berita $\rightarrow$ topik
2. e-mail $\rightarrow$ spam
3. tweet $\rightarrow$ sentimen

Bagaimana merepresentasikannya?

Kata sebagai Atribut Numerik

1. Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**

Kata sebagai Atribut Numerik

1. Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**
2. Bernilai 1 jika terdapat di contoh teks, 0 jika tidak

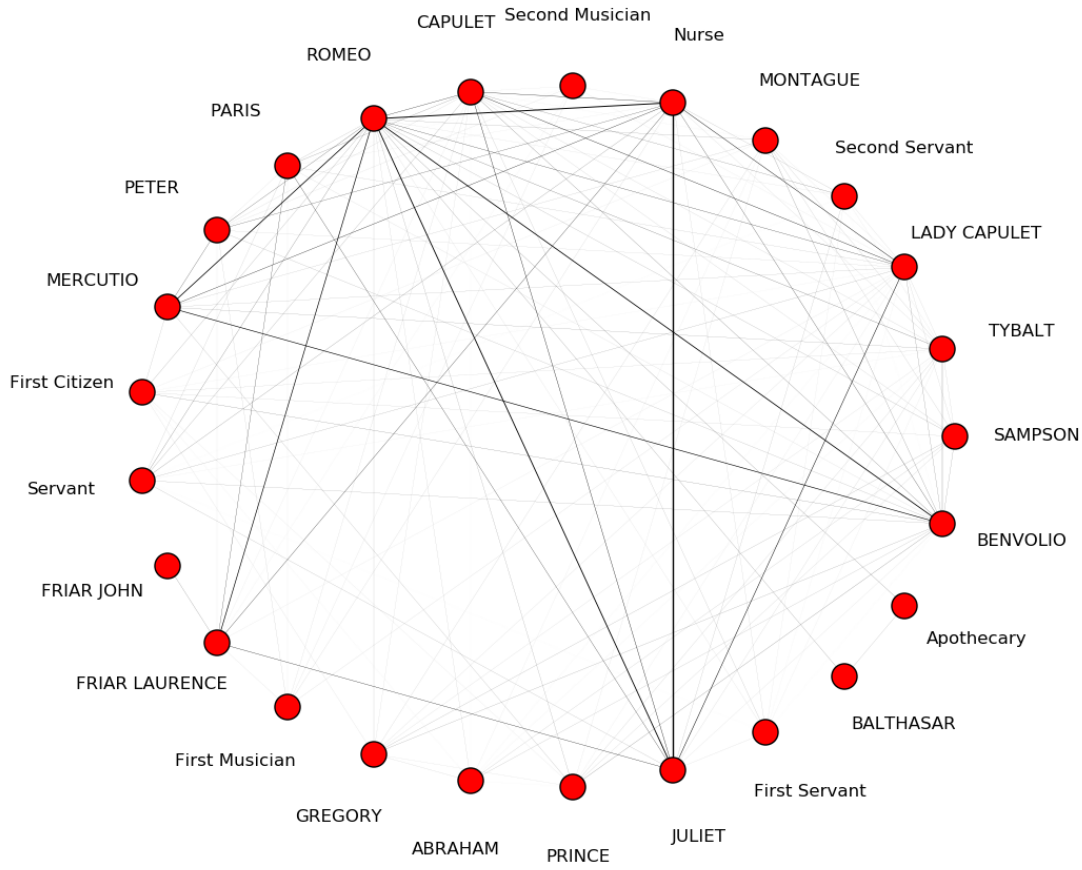
Kata sebagai Atribut Numerik

1. Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**
2. Bernilai 1 jika terdapat di contoh teks, 0 jika tidak
3. Dapat diubah menjadi frekuensi atau bobot (TF-IDF)

Kata sebagai Atribut Numerik

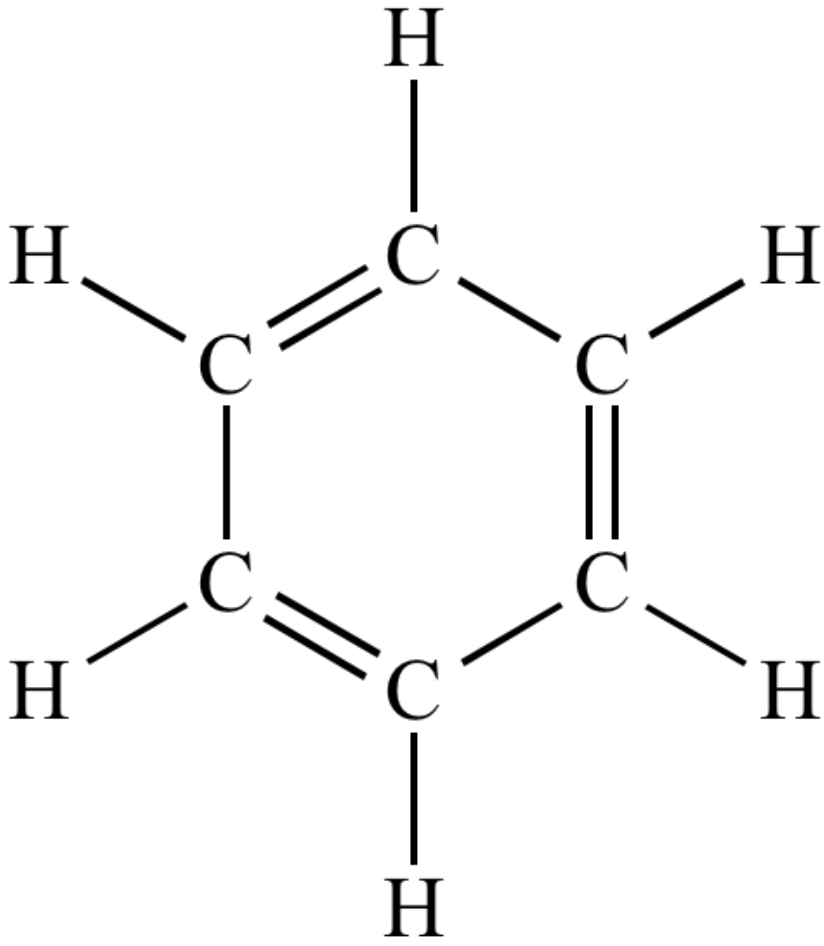
1. Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**
2. Bernilai 1 jika terdapat di contoh teks, 0 jika tidak
3. Dapat diubah menjadi frekuensi atau bobot (TF-IDF)
4. **Catatan:** Dimensinya bisa jadi sangat besar dan matriksnya akan menjadi *sparse*

Jejaring Sosial



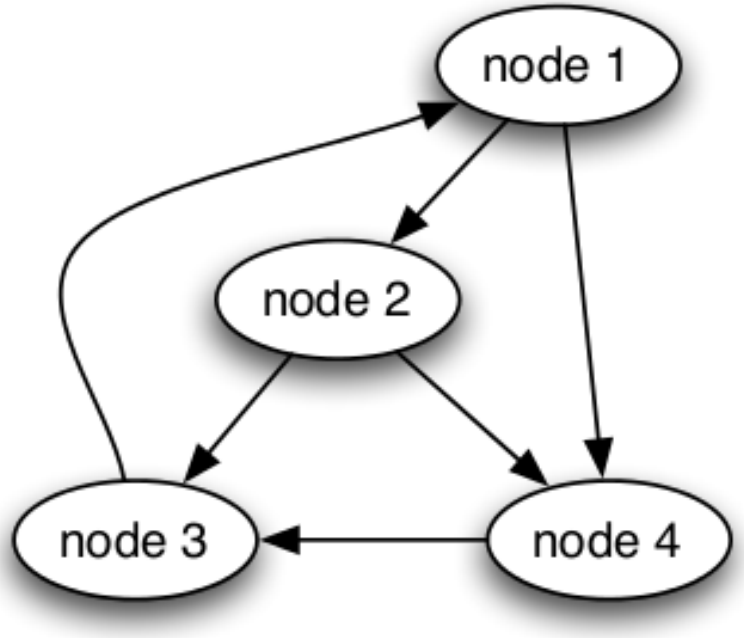
Gambar: Graf dari drama Romeo dan Juliet berdasarkan kemunculan karakter di satu babak yang sama

Struktur Kimia



Gambar: Struktur kimia benzena [Hardinger, 2017]

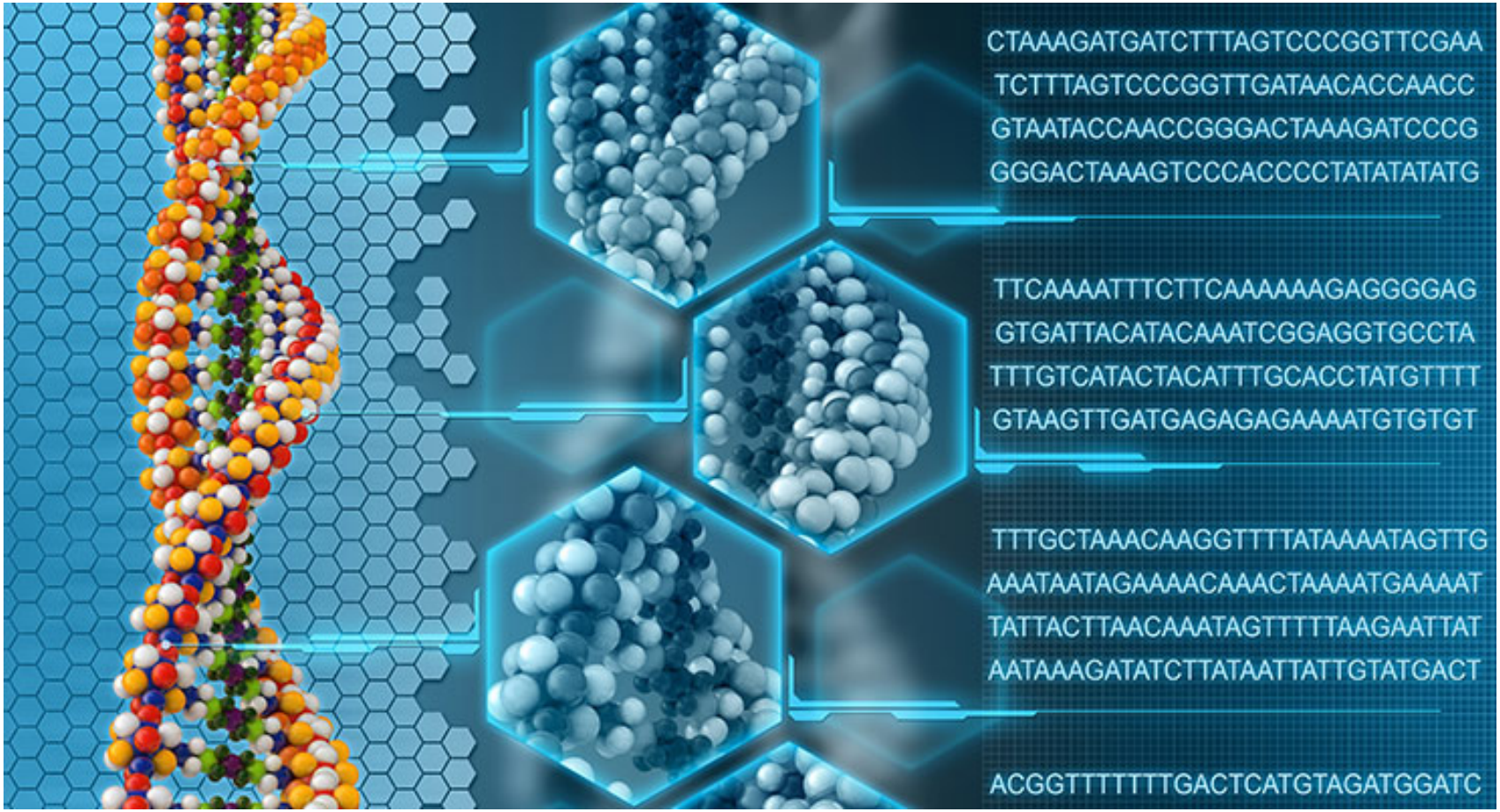
Adjacency Matrix



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

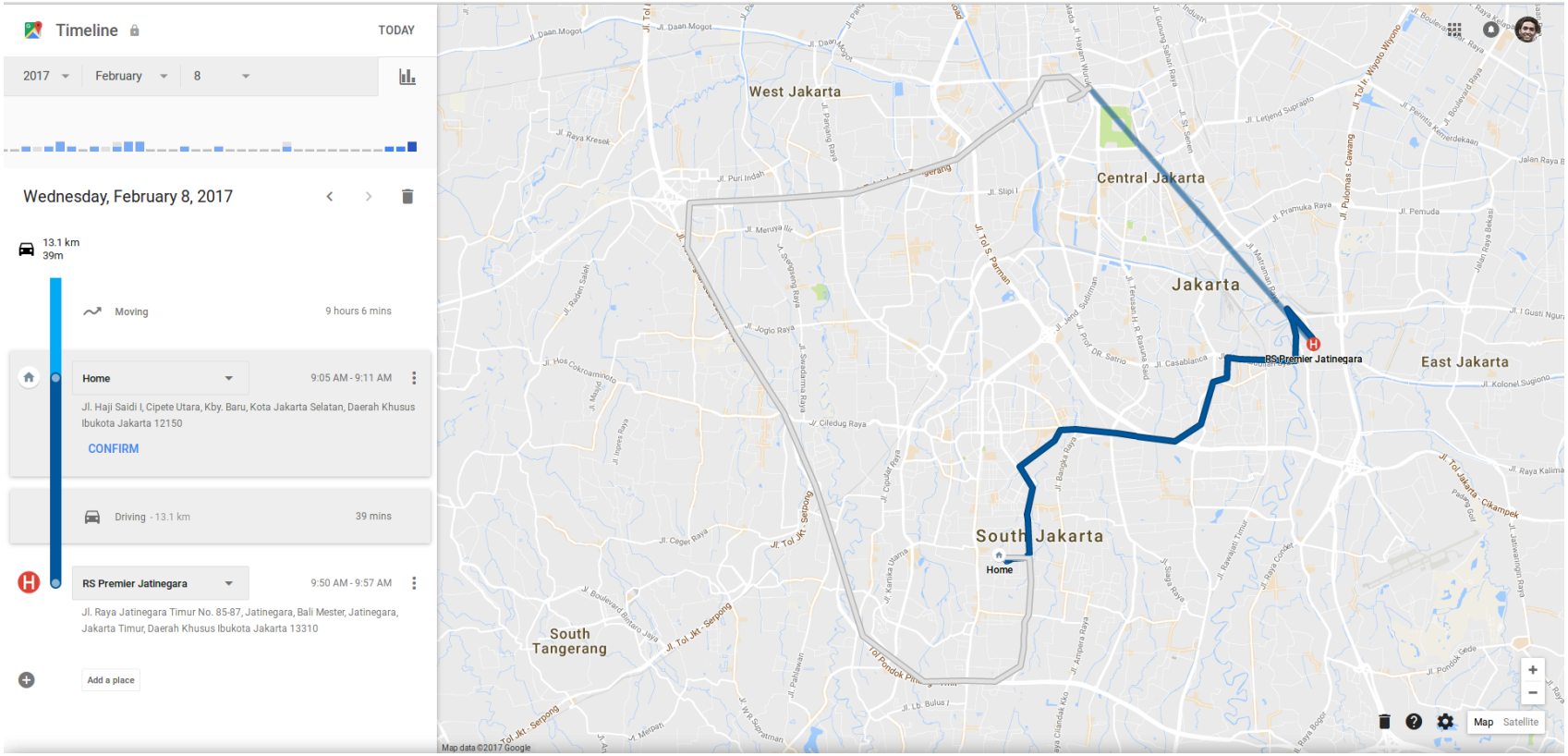
Gambar: Adjacency matrix dari graf [Easley dan Kleinberg, 2010]

Genomic Sequence



Gambar: Urutan genom [Global Biodefense, 2014]

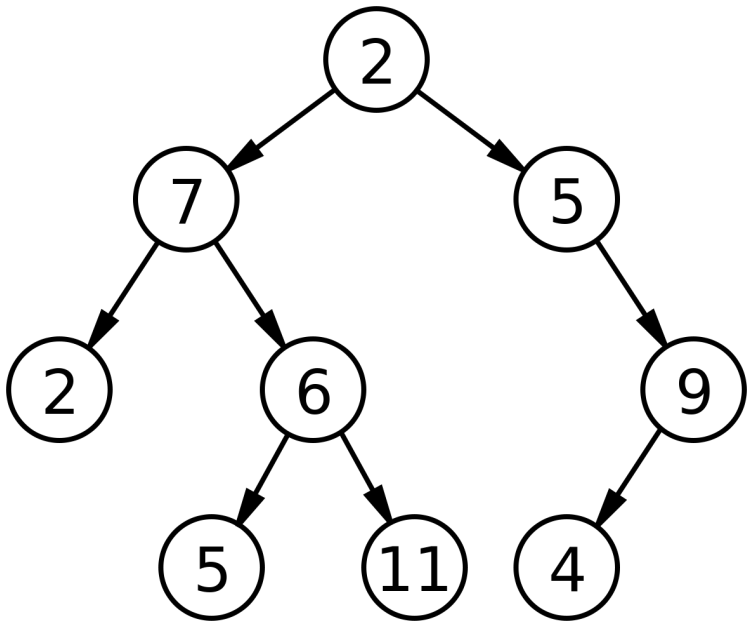
Spatio-Temporal



Gambar: Peta perjalanan seseorang yang direkam oleh Google Maps

Masalah pada Data

Dealing with Structures



Gambar: Data yang strukturnya berbentuk pohon

1. Atribut dapat berupa jalur dari akar ke daun
2. Contoh: {2-7-2-NA, 2-7-6-5, 2-7-6-11, ...}

Missing Values

1. Tipe: tidak diketahui, tidak tersimpan, tidak relevan
2. Penyebab: perubahan desain eksperimen, penggabungan dataset, dsb.
3. *Sangat mungkin terjadi!*

Missing Values - Solusi

1. Nominal: Gunakan label spesial, e.g. "NA"
2. Numerik: Diganti nilainya, e.g. rata-rata atau median atribut tersebut
3. Algoritma: Beberapa algoritma, e.g. Naïve Bayes dan *decision trees* dapat menyelesaikan kasus ini
4. Buang *instance*-nya

Inaccurate Values

1. Kasus-kasus pencilan, kesalahan pengukuran, duplikat
2. *Pahami datanya!*
3. Dapat dibuang dengan konsekuensi terhadap akurasi model

Unbalanced Data

1. Kasus umum pada klasifikasi, e.g. diagnosis pasien
2. Frekuensi salah satu kelas lebih banyak dibanding kelas lain
3. Mungkin perlu *metrics* selain akurasi
4. Ongkos kesalahan klasifikasi yang mungkin perlu dibuat tidak seimbang
5. Lihat [Kotsiantis, et al., 2006]!

Referensi

Eliot Van Buskirk (7 Februari 2017)

Spotify, Accuweather Reveal How Weather Affects Music Listening

https://insights.spotify.com/us/2017/02/07/spotify-accuweather-music-and-weather/

Josh Bersin (19 Februari 2014)

The Myth Of The Bell Curve: Look For The Hyper-Performers

https://www.forbes.com/sites/joshbersin/2014/02/19/the-myth-of-the-bell-curve-look-for-the-hyper-performers/

Tim O'Shea (Juli 2016)

MNIST Generative Adversarial Model in Keras

http://www.kdnuggets.com/2016/07/mnist-generative-adversarial-model-keras.html

Alex Krizhevsky (2009)

Learning Multiple Layers of Features from Tiny Images

https://www.cs.toronto.edu/~kriz/cifar.html

Referensi

Steve Hardinger (diakses 27 Februari 2017)

Illustrated Glossary of Organic Chemistry

http://web.chem.ucla.edu/~harding/IGOC/B/benzene_ring.html

David Easley & Jon Kleinberg (2010)

Networks, crowds, and markets: Reasoning about a highly connected world

Cambridge University Press

Global Biodefense (25 Juni 2014)

USAMRIID Leads Effort on Viral Genome Sequencing Standards

https://globalbiodefense.com/2014/06/25/usamriid-leads-effort-viral-genome-sequencing-standards/

Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas (2006)

Handling imbalanced datasets: A review

GESTS International Transactions on Computer Science and Engineering, 30(1), 25-36.

Terima kasih