Introductory Applied Machine Learning

Ali Akbar Septiandri

November 10, 2017

untuk Astra Graphia IT

Daftar Isi

1. Pendahuluan

2. Konsep Data Mining

3. Tugas-tugas dalam Data Mining

4. Representasi Data

5. Tipe Data

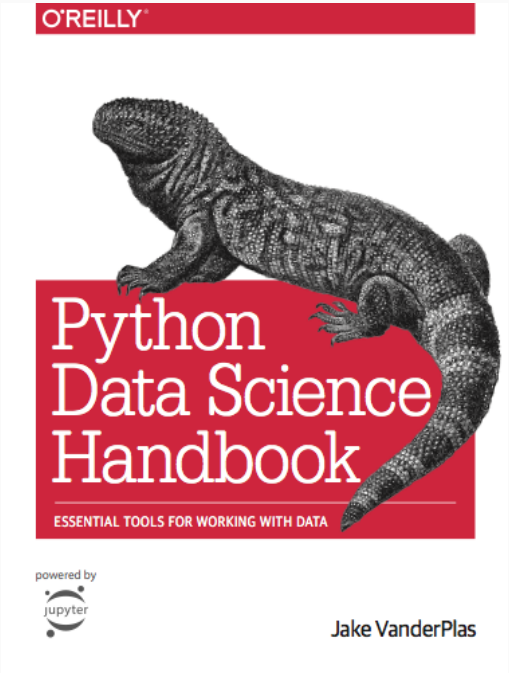
6. Masalah pada Data

Pendahuluan

Materi

1. Konsep *Data Mining*
2. Tipe Data
3. Konsep Jarak Antardata
4. Eksplorasi Data
5. Klasifikasi
6. Regresi
7. *Clustering*
8. Dimensionality Reduction
9. Asosiasi / Sistem Rekomendasi

Referensi



VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.

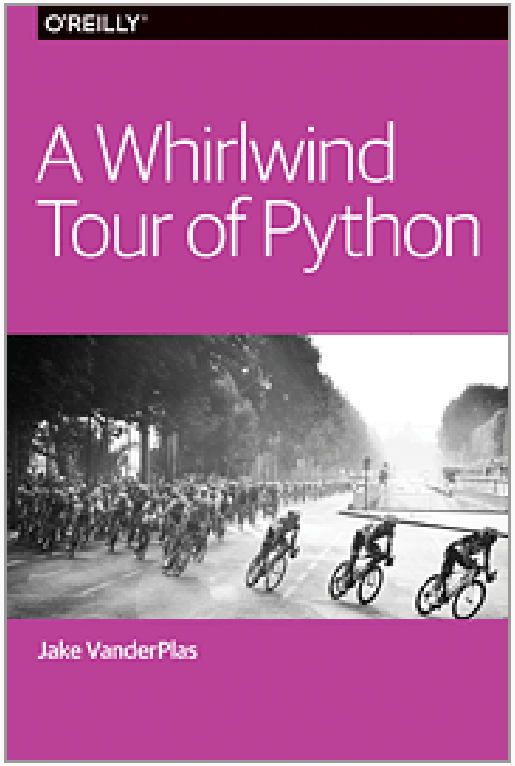
Referensi

1. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. (*slides* tersedia online)
2. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press. (tersedia online)
3. Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media, Inc.
4. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87. (tersedia online)

Administrasi

1. Materi bisa dilihat di `https://github.com/aliakbars/iaml`
2. Pertemuan setiap hari Sabtu, pukul 09.00-16.00
3. Bahasa/teknologi pengantar: Python, pandas, scikit-learn, Jupyter Notebook
4. Terdapat 3 tugas
5. Praktikum di tiap pertemuan

Belajar Python



A Whirlwind Tour of Python by Jake VanderPlas (OReilly).
Copyright 2016 OReilly Media, Inc., 978-1-491-96465-1

Konsep Data Mining

Apa itu *Data Mining*?

Data Mining

- *Generic*: "the discovery of **'models'** for data" [Leskovec, et al. 2014]

Data Mining

- *Generic*: "the discovery of **'models'** for data" [Leskovec, et al. 2014]
- *Statisticians*: "the construction of **statistical model**, that is, an **underlying distribution** from which the visible data is drawn" [Leskovec, et al. 2014]

Data Mining

- *Generic*: "the discovery of **'models'** for data" [Leskovec, et al. 2014]
- *Statisticians*: "the construction of **statistical model**, that is, an **underlying distribution** from which the visible data is drawn" [Leskovec, et al. 2014]
- **Menemukan pola** dalam data yang dapat memberikan **wawasan** atau memungkinkan **pengambilan keputusan** yang cepat dan akurat [Witten, et al. 2016]

Keterkaitan dengan Machine Learning

- Dalam prosesnya, algoritma *machine learning* sering digunakan untuk mempermudah proses *data mining*

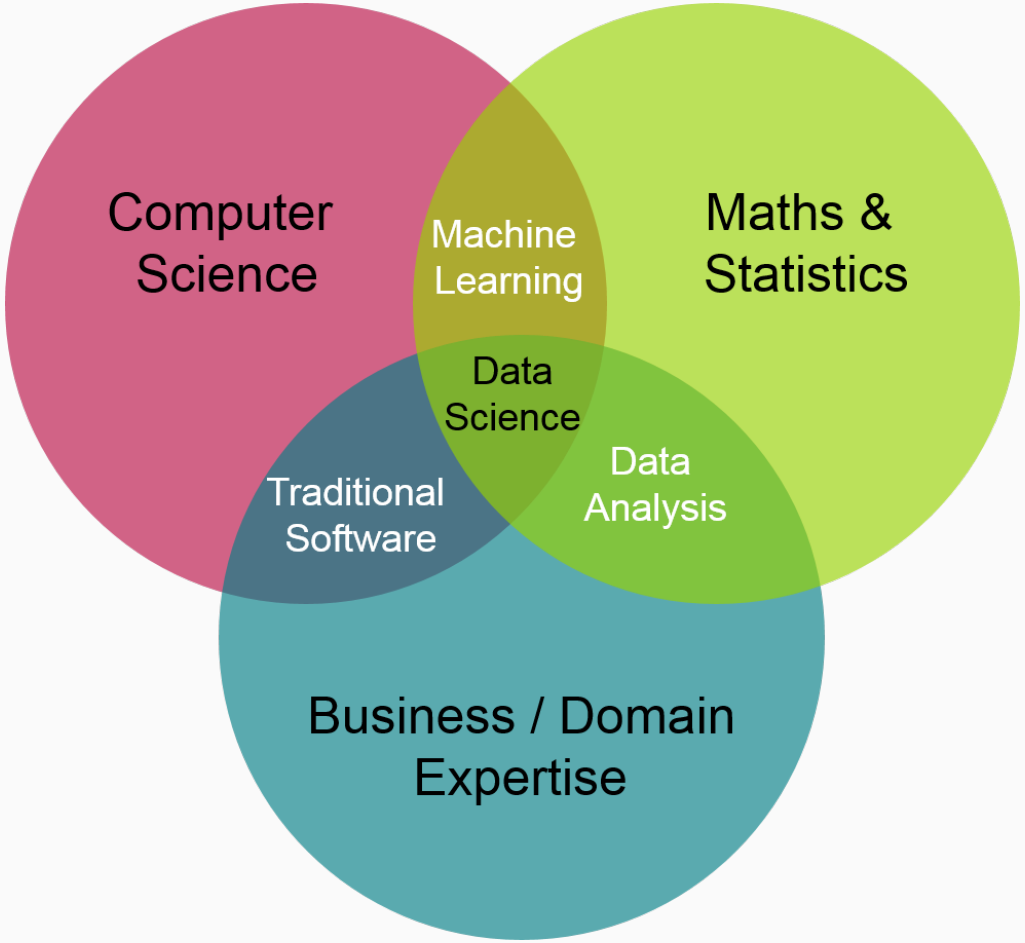
Keterkaitan dengan Machine Learning

- Dalam prosesnya, algoritma *machine learning* sering digunakan untuk mempermudah proses *data mining*
- *Machine learning* dapat bekerja dengan baik jika pengetahuan yang kita miliki terbatas

Keterkaitan dengan Machine Learning

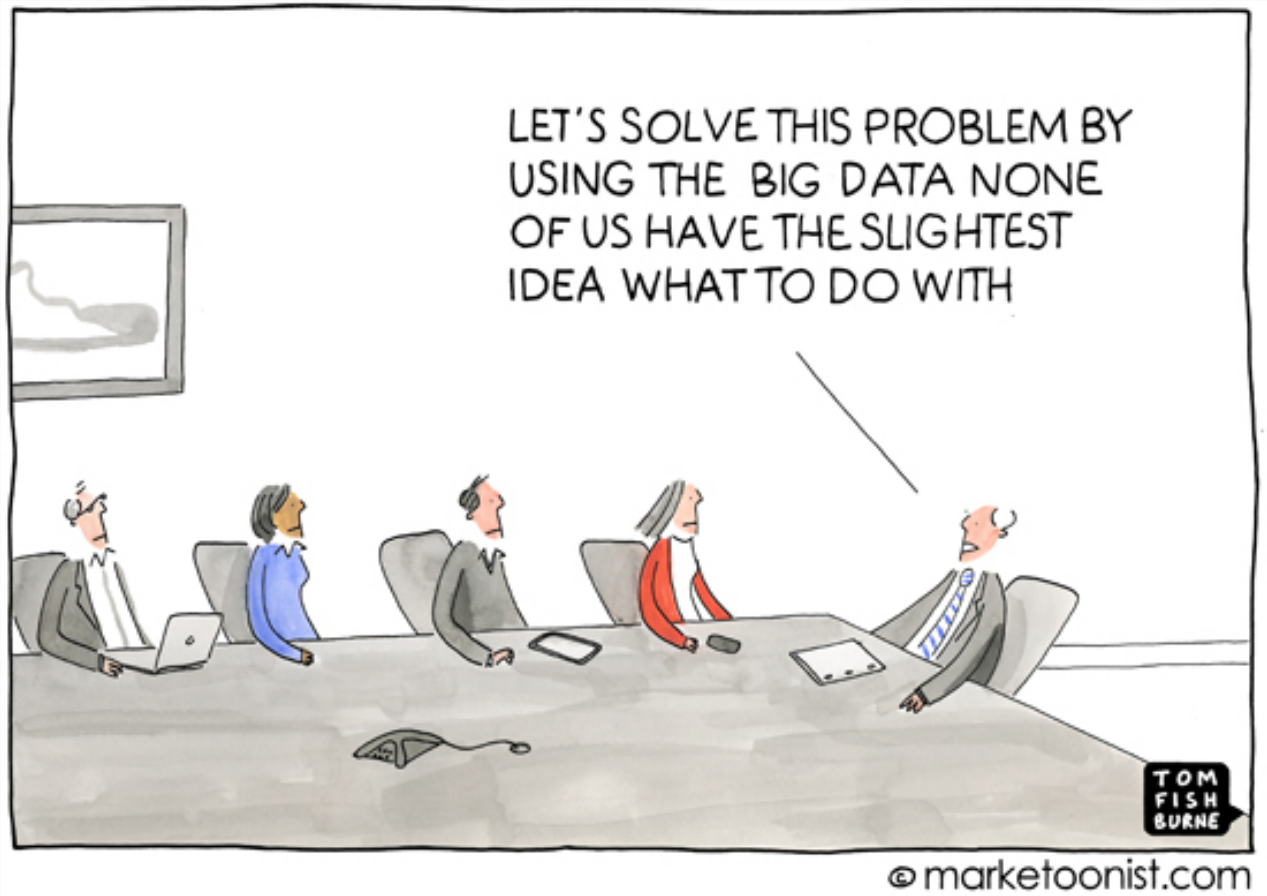
- Dalam prosesnya, algoritma *machine learning* sering digunakan untuk mempermudah proses *data mining*
- *Machine learning* dapat bekerja dengan baik jika pengetahuan yang kita miliki terbatas
- Jika polanya sudah *straightforward*, gunakan saja *if-then-else*!

Data Science Venn Diagram



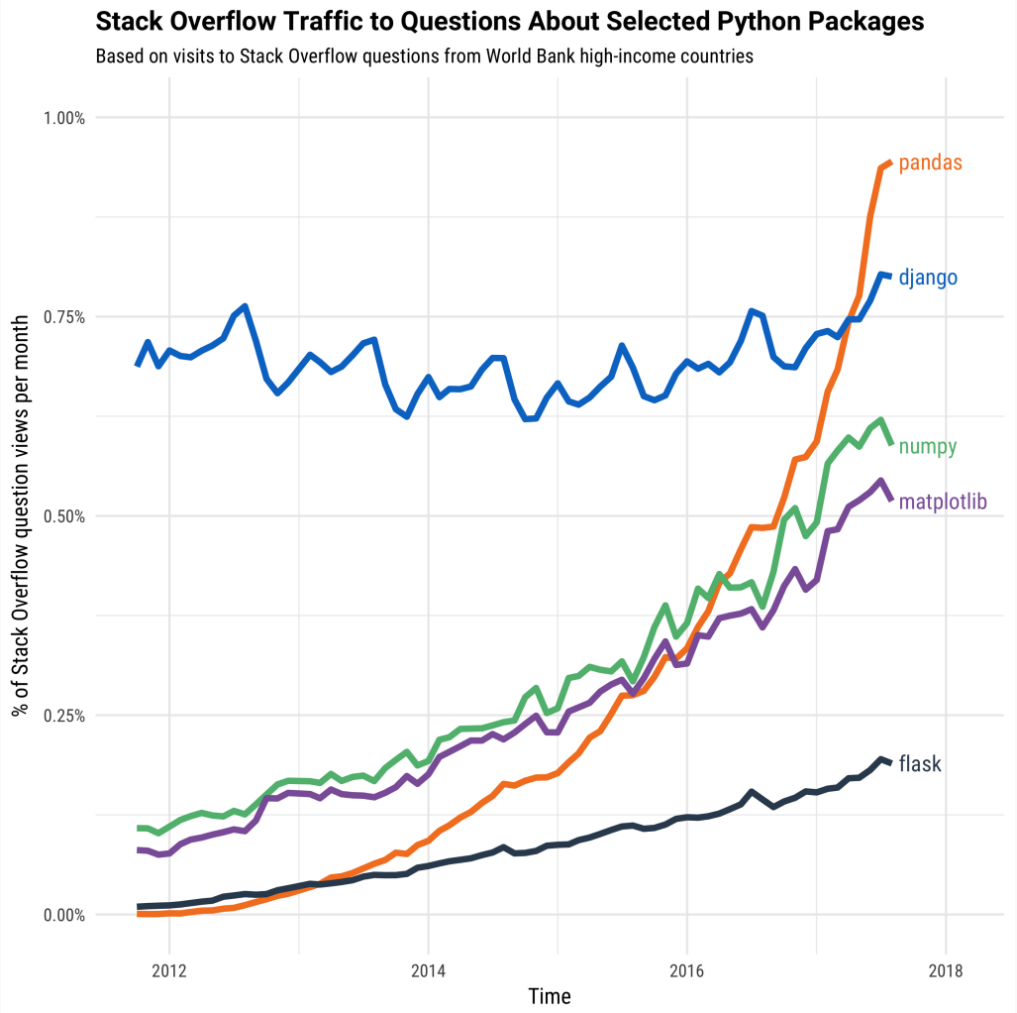
Gambar 1: Pelatihan ini akan difokuskan pada *machine learning*

Data Mining & Big Data



Gambar 2: Dari
`https://marketoonist.com/2014/01/big-data.html`

Tren Data Mining



Gambar 3: Peningkatan minat data mining dilihat dari pustaka Python populer [Robinson, 2017]

Data Mining Deskriptif

- Tidak semua tugas dalam *data mining* memerlukan model yang melakukan prediksi
- Terdapat tugas yang sifatnya hanya deskriptif
- Salah satu contoh yang terkenal adalah algoritma PageRank (Page, et al. 1999)

PageRank

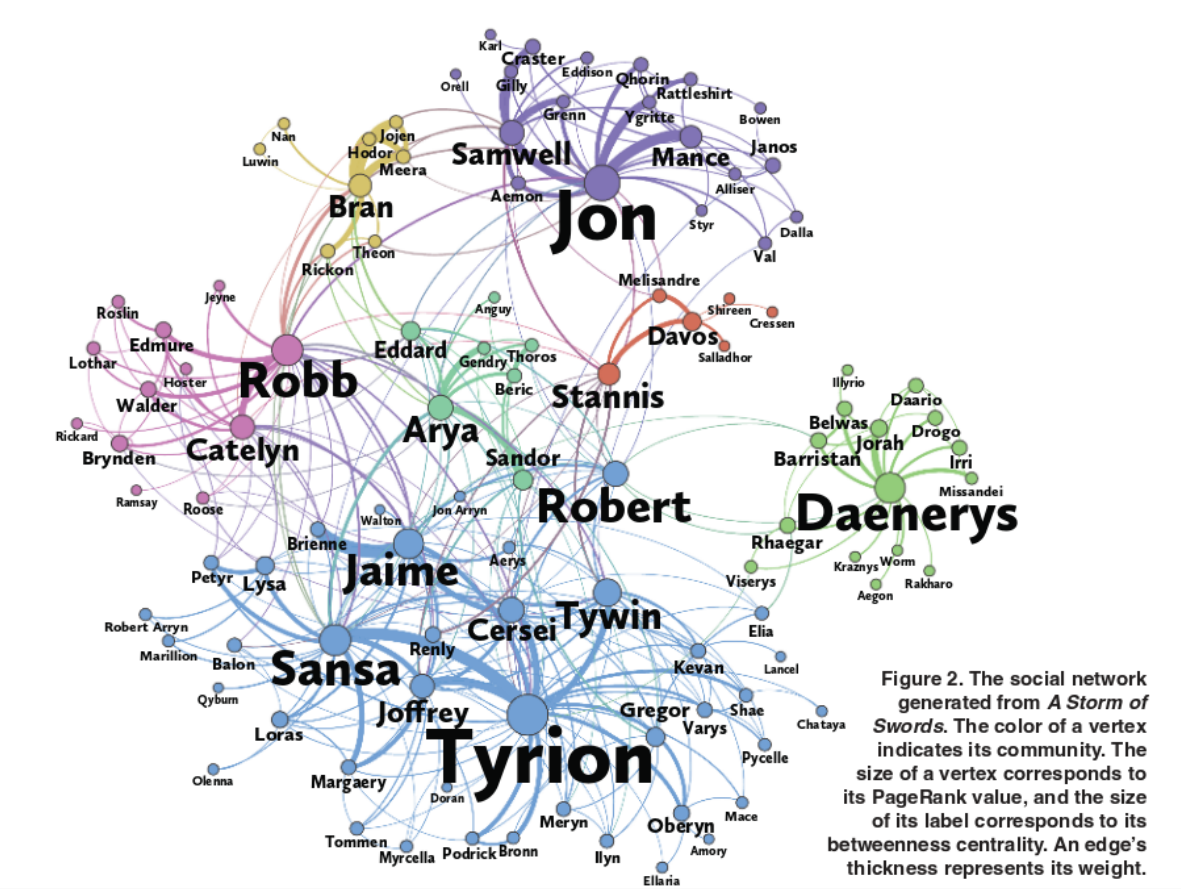
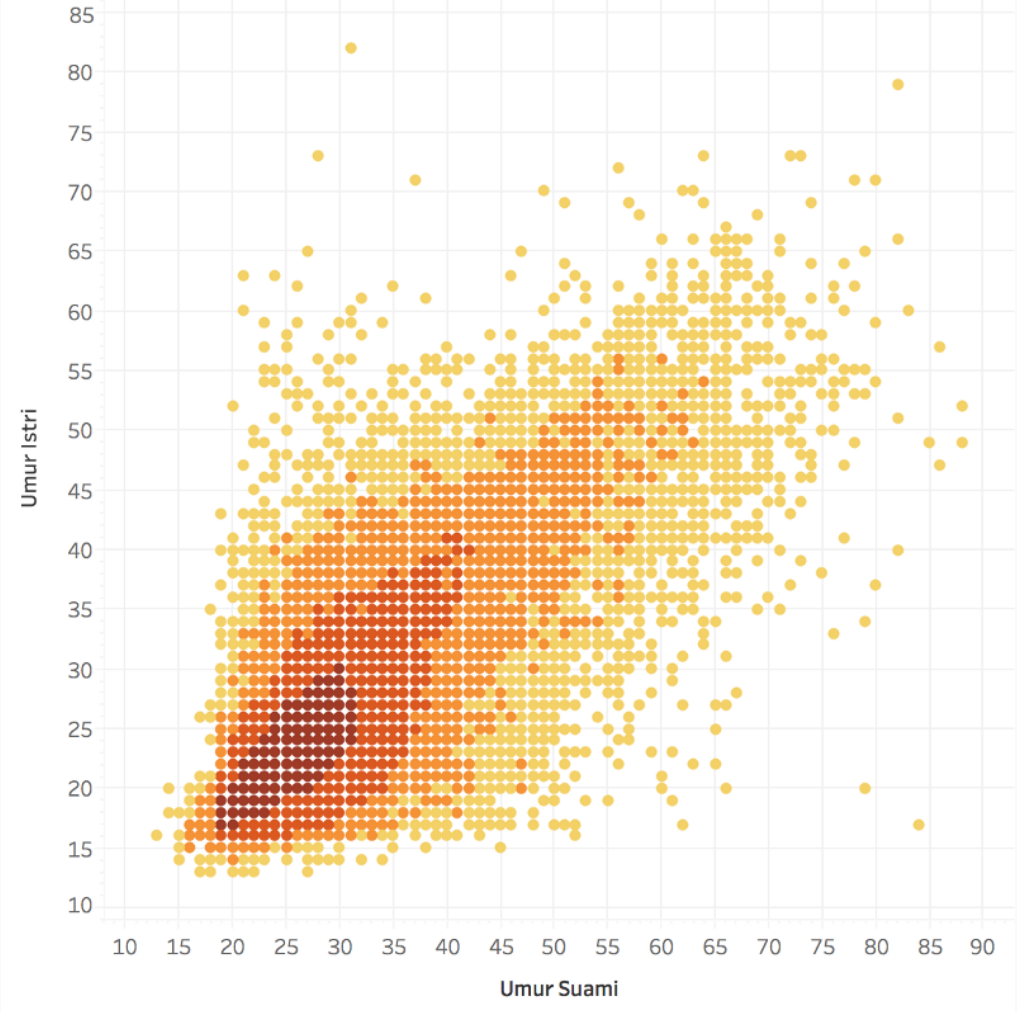


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

Gambar 4: Penerapan PageRank pada karakter serial Game of Thrones [Beveridge and Shan, 2016]

Heatmap



Gambar 5: Usia pernikahan [Yanurzha, 2017]

"First-timers are often surprised by how little time in a machine learning project is spent actually doing machine learning."

Sumber Data

Beberapa situs yang menyediakan data yang sudah siap diolah:

1. Kaggle (`https://www.kaggle.com/datasets`)
2. UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets.html`)
3. Portal Data Indonesia (`http://data.go.id/`)
4. SNAP (`http://snap.stanford.edu/`)

Sumber Data

Beberapa situs tidak menyediakan API untuk memberikan data karena:

1. tidak dikembangkan sejak awal;
2. tidak ingin datanya disebarkan, e.g. Instagram; atau
3. hanya bisa diakses terbatas, e.g. Microdata BPS

sehingga **mungkin** perlu dilakukan *scraping*.

“visible $\neq$ accessible $\neq$ storable $\neq$ presentable”

[Lavrenko, 2010]

Tugas-tugas dalam Data Mining

Klasifikasi

1. Memprediksi nilai yang sudah pasti

Klasifikasi

1. Memprediksi nilai yang sudah pasti
2. *Biasanya* direpresentasikan sebagai kelas biner $\{0, 1\}$ atau $\{-1, 1\}$

Klasifikasi

1. Memprediksi nilai yang sudah pasti
2. *Biasanya* direpresentasikan sebagai kelas biner $\{0, 1\}$ atau $\{-1, 1\}$
3. Membutuhkan label

Klasifikasi

1. Memprediksi nilai yang sudah pasti
2. *Biasanya* direpresentasikan sebagai kelas biner $\{0, 1\}$ atau $\{\text{-}1, 1\}$
3. Membutuhkan label
4. Mempunyai *evaluation metrics* yang jelas, e.g. akurasi

Klasifikasi

1. Memprediksi nilai yang sudah pasti
2. *Biasanya* direpresentasikan sebagai kelas biner $\{0, 1\}$ atau $\{\text{-}1, 1\}$
3. Membutuhkan label
4. Mempunyai *evaluation metrics* yang jelas, e.g. akurasi
5. Contoh: identifikasi spam, MNIST digit recognition

Regresi

1. Membutuhkan label

Regresi

1. Membutuhkan label
2. Memprediksi nilai kontinu

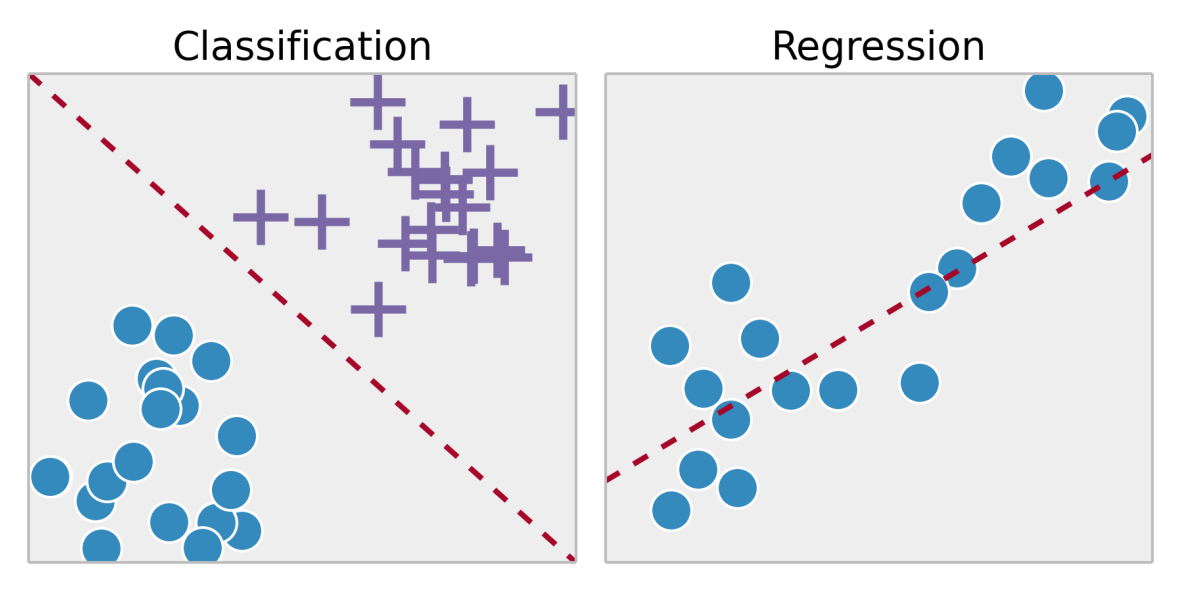
Regresi

1. Membutuhkan label
2. Memprediksi nilai kontinu
3. *Evaluation metrics* berupa *error*, e.g. Mean Squared Error (MSE), Mean Absolute Error (MAE)

Regresi

1. Membutuhkan label
2. Memprediksi nilai kontinu
3. *Evaluation metrics* berupa *error*, e.g. Mean Squared Error (MSE), Mean Absolute Error (MAE)
4. Contoh: prediksi nilai saham, jumlah RT dari suatu *tweet*

Klasifikasi vs Regresi



Gambar 6: Perbedaan klasifikasi dan regresi [Rossant, 2014]

Klasifikasi dan Regresi

Fungsi

Kedua tugas ini dapat dilihat sebagai fungsi f yang memetakan atribut x ke label y.

Clustering

1. Mencoba memberikan deskripsi terhadap data

Clustering

1. Mencoba memberikan deskripsi terhadap data
2. Tidak berhubungan dengan label

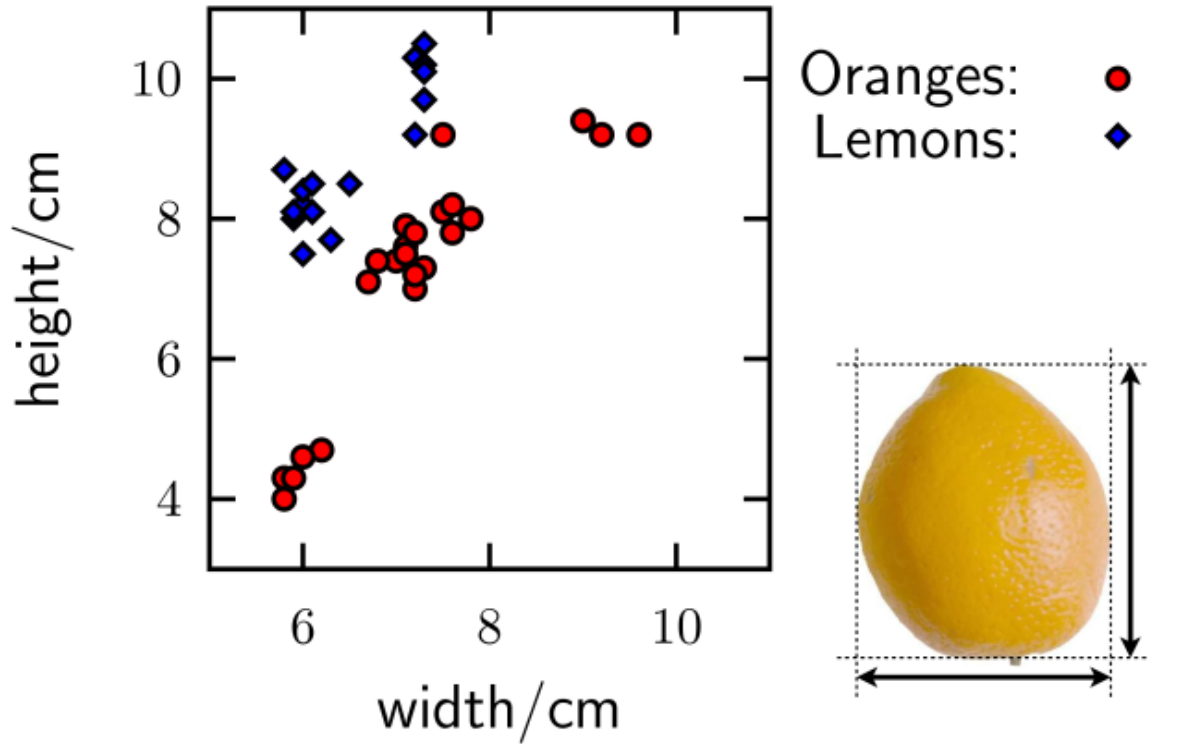
Clustering

1. Mencoba memberikan deskripsi terhadap data
2. Tidak berhubungan dengan label
3. Menemukan pola yang “menarik” dalam data

Clustering

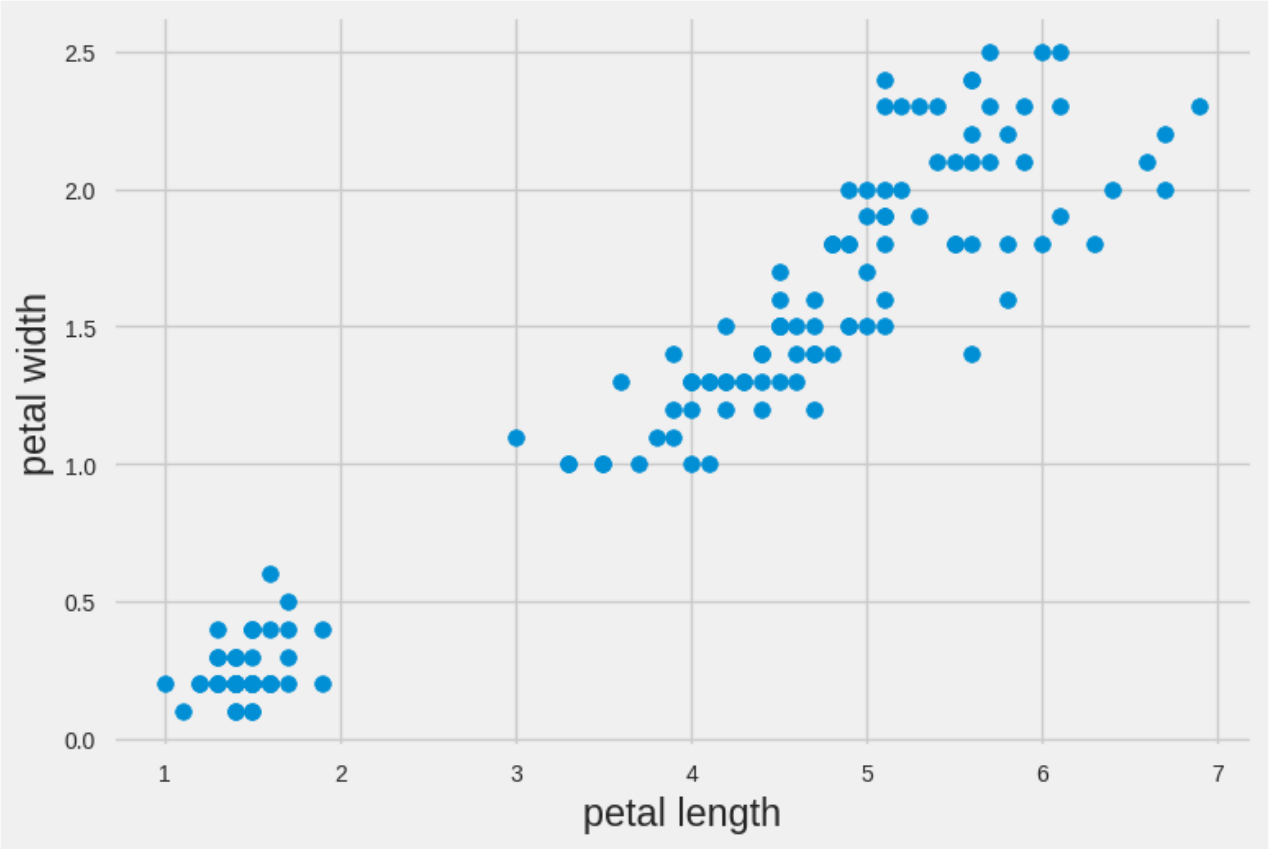
1. Mencoba memberikan deskripsi terhadap data
2. Tidak berhubungan dengan label
3. Menemukan pola yang “menarik” dalam data
4. Tidak mempunyai *evaluation metrics* yang pasti

Contoh Clustering



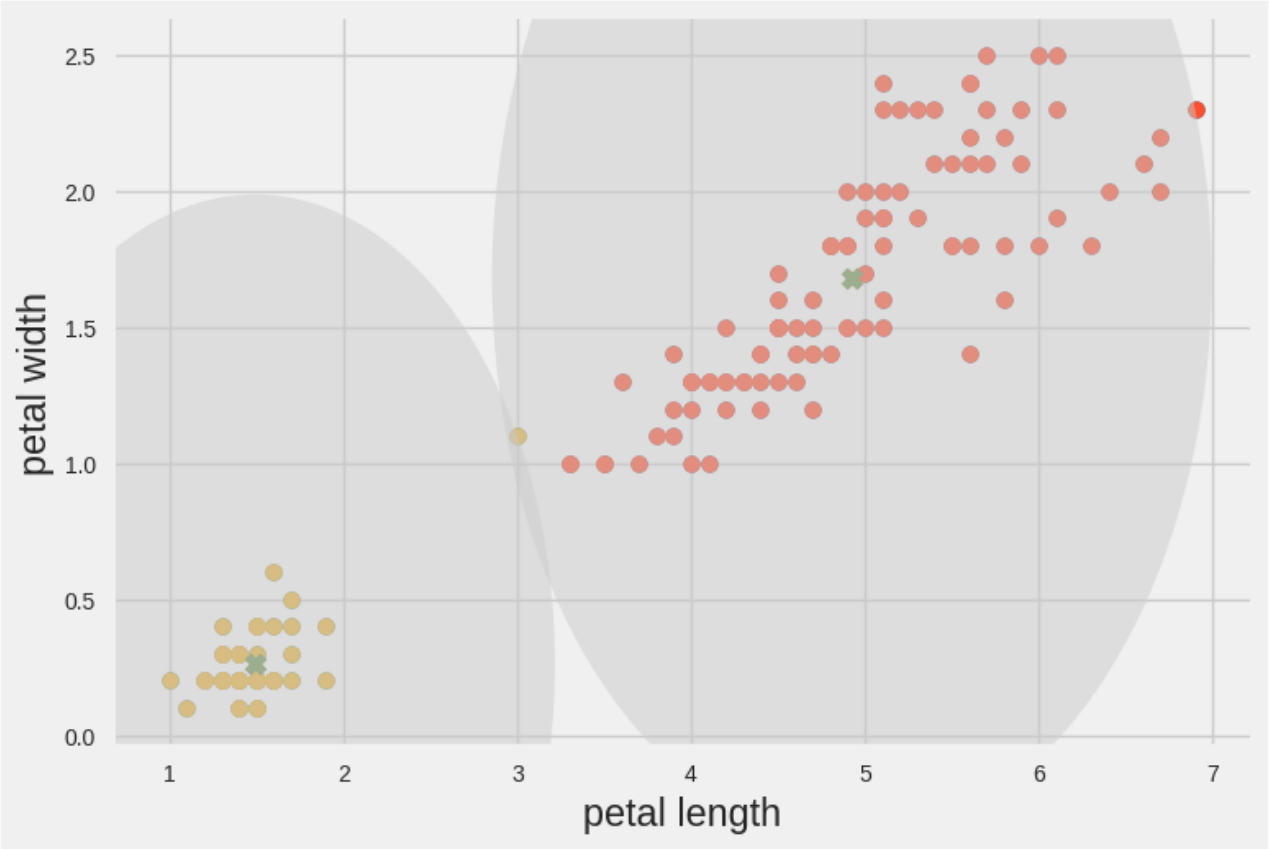
Gambar 7: *Clustering* buah lemon dan jeruk [Murray, 2011]

Contoh Clustering



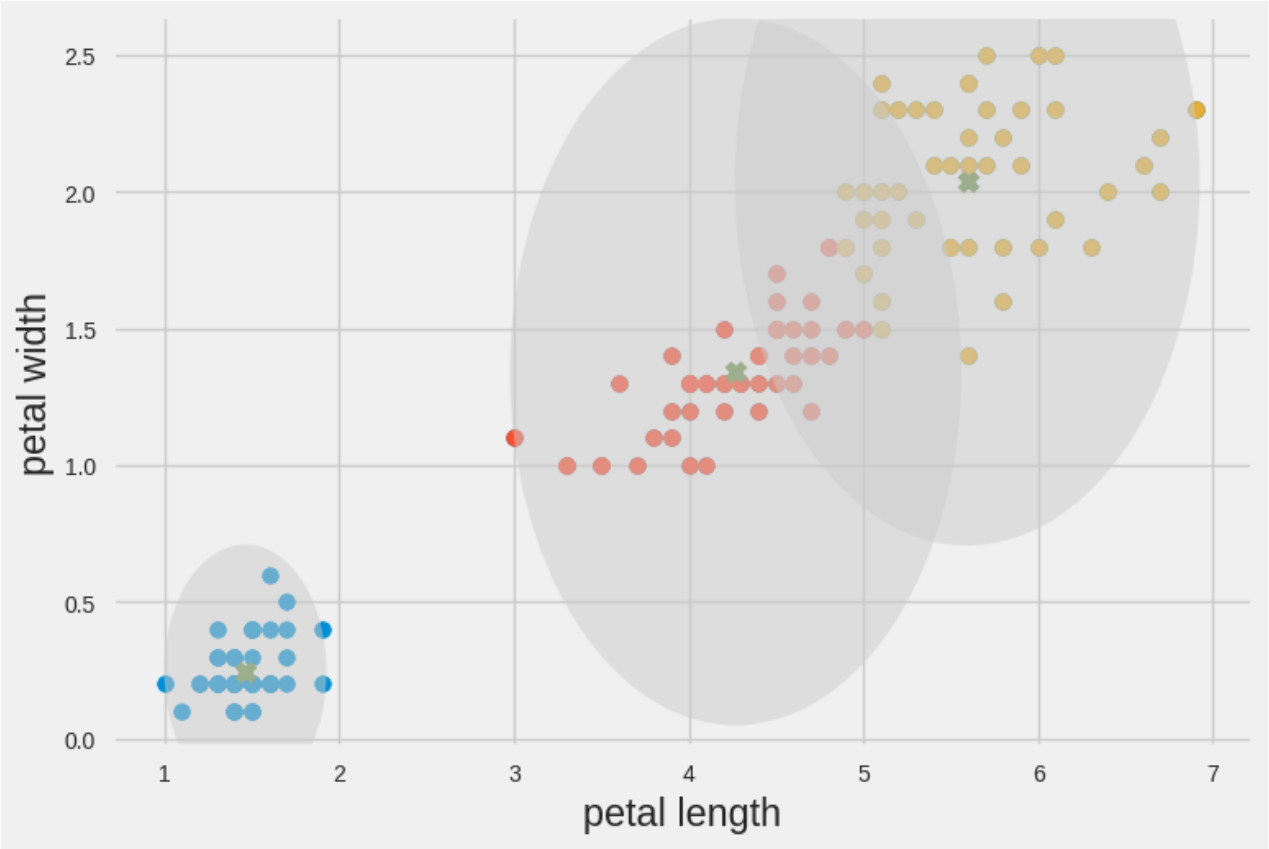
Gambar 8: Klaster dari dataset Iris

Contoh Clustering



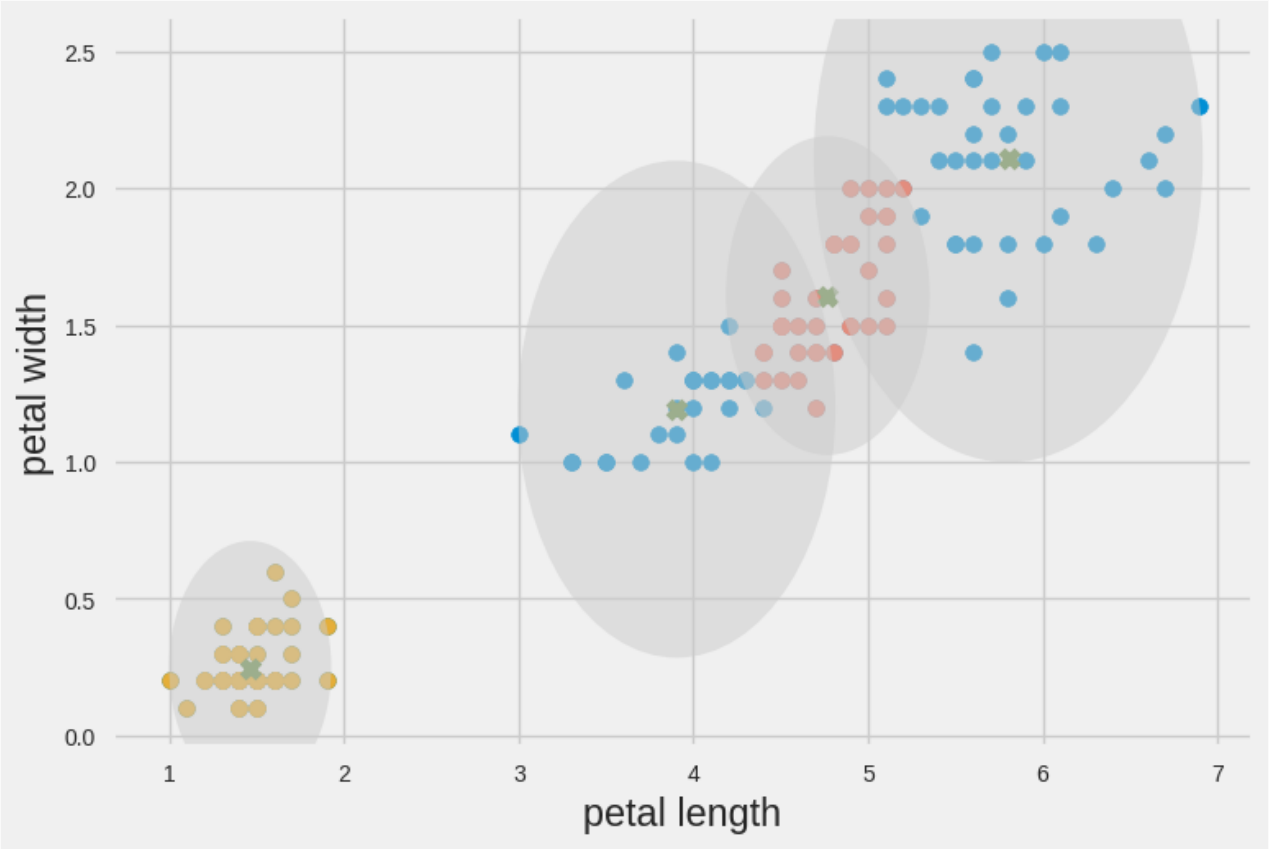
Gambar 8: Klaster dari dataset Iris

Contoh Clustering



Gambar 8: Klaster dari dataset Iris

Contoh Clustering



Gambar 8: Klaster dari dataset Iris

Perhitungan Jarak

1. Untuk mengetahui kedekatan, perlu diukur jarak antarcontoh (*instances*)

Perhitungan Jarak

1. Untuk mengetahui kedekatan, perlu diukur jarak antarcontoh (*instances*)
2. Jarak bernilai non-negatif

Perhitungan Jarak

1. Untuk mengetahui kedekatan, perlu diukur jarak antarcontoh (*instances*)
2. Jarak bernilai non-negatif
3. Contoh perhitungan jarak: *Jaccard distance*, *cosine similarity*, *Euclidean distance*

Asosiasi dengan Aturan

Jika diberikan sejumlah barang dalam beberapa keranjang belanja, tentukan aturan yang dapat menjelaskan adanya benda lain dalam keranjang tersebut!

Barang-barang

1. Roti, soda, susu
2. Bir, roti
3. Bir, soda, popok, susu
4. Bir, roti, popok, susu
5. Soda, popok, susu

Asosiasi dengan Aturan

Jika diberikan sejumlah barang dalam beberapa keranjang belanja, tentukan aturan yang dapat menjelaskan adanya benda lain dalam keranjang tersebut!

Barang-barang

1. Roti, soda, susu
2. Bir, roti
3. Bir, soda, popok, susu
4. Bir, roti, popok, susu
5. Soda, popok, susu

Aturan yang ditemukan

1. $\{\text{Susu}\} \rightarrow \{\text{Soda}\}$
2. $\{\text{Popok, susu}\} \rightarrow \{\text{Bir}\}$

Sistem Rekomendasi

Gambar 9: Rekomendasi pada situs Amazon

Berikan rekomendasi sejumlah K konten kepada pengguna u, dari pilihan M konten yang tersedia!

Jenis-jenis Sistem Rekomendasi

1. **Rekomendasi berdasarkan konten**
   "Pilih K konten yang variabelnya paling sesuai dengan variabel preferensi pengguna u"

Jenis-jenis Sistem Rekomendasi

1. **Rekomendasi berdasarkan konten**
   "Pilih K konten yang variabelnya paling sesuai dengan variabel preferensi pengguna u"
2. **Collaborative filtering**
   "Pilih K konten yang rating-nya paling sesuai dengan preferensi (rating) pengguna u"

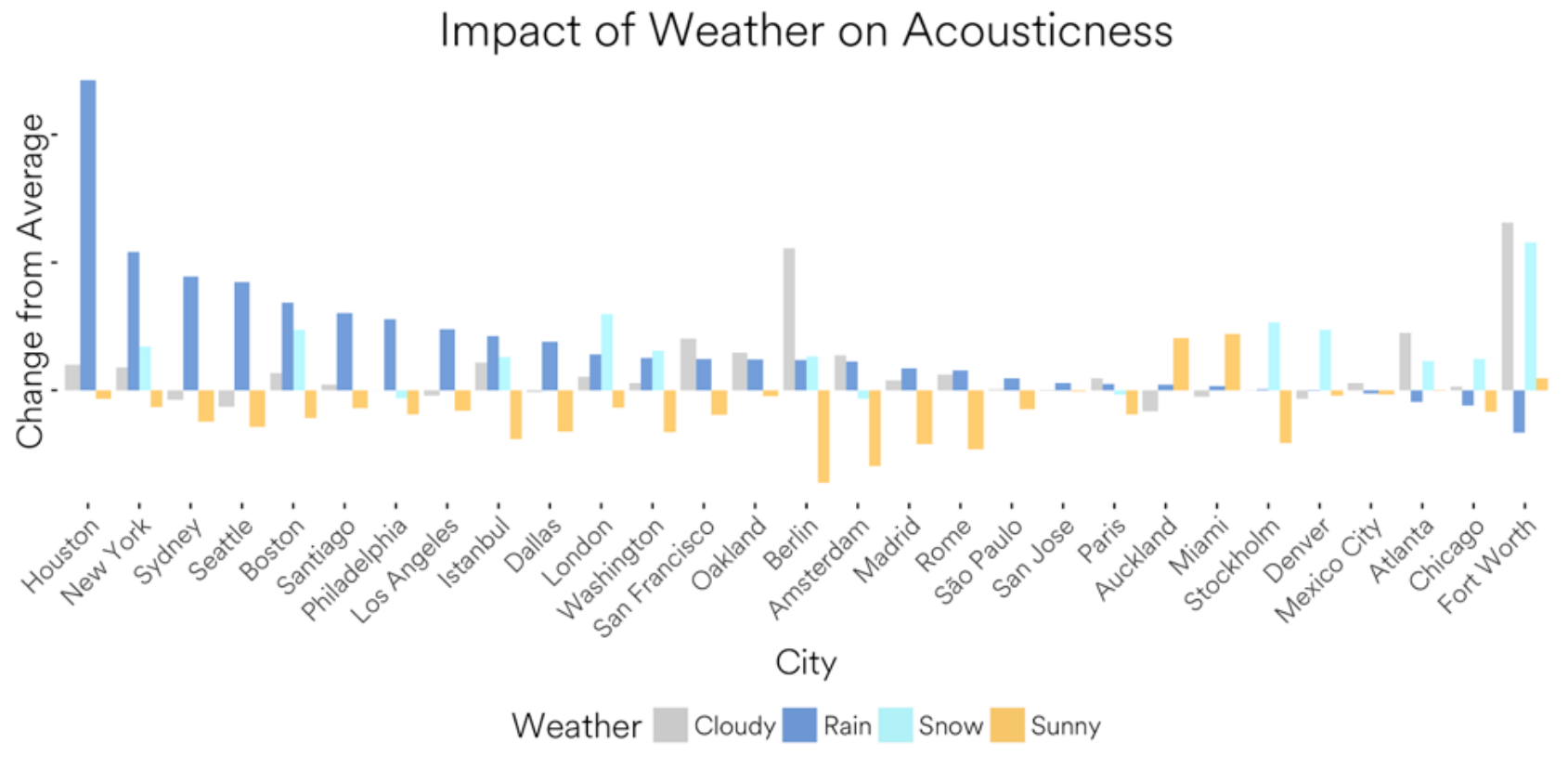
Jenis-jenis Sistem Rekomendasi

1. **Rekomendasi berdasarkan konten**
   "Pilih K konten yang variabelnya paling sesuai dengan variabel preferensi pengguna u"
2. **Collaborative filtering**
   "Pilih K konten yang rating-nya paling sesuai dengan preferensi (rating) pengguna u"
3. **Rekomendasi melalui klasifikasi**
   "Pilih K konten yang diklasifikasikan sebagai kelas positif untuk pengguna u"

Representasi Data

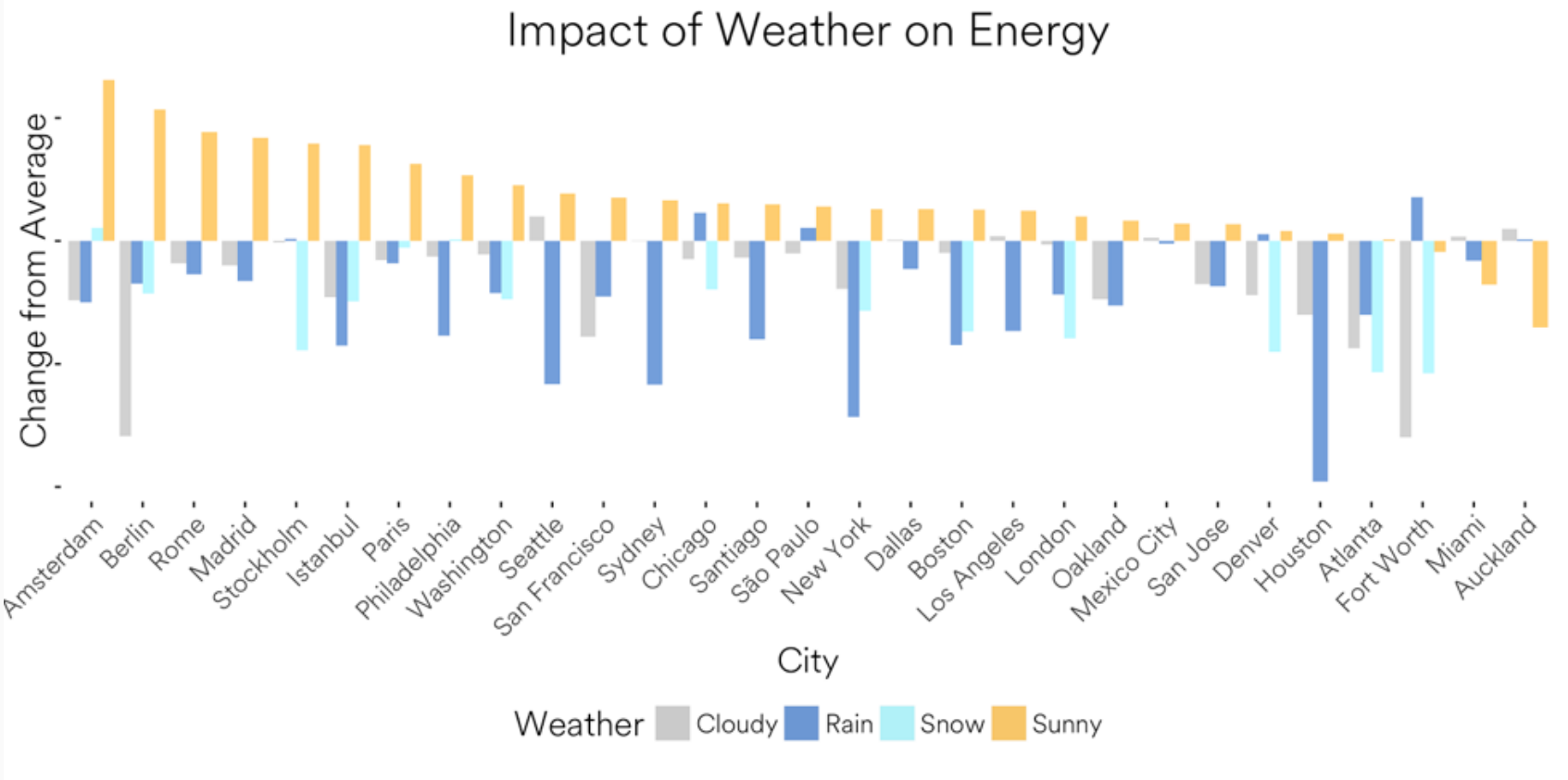
Variabel seperti apa yang dapat dipakai oleh sistem rekomendasi berdasarkan konten dari aplikasi seperti Spotify?

Korelasi pada Spotify



Gambar 10: 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 11: 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

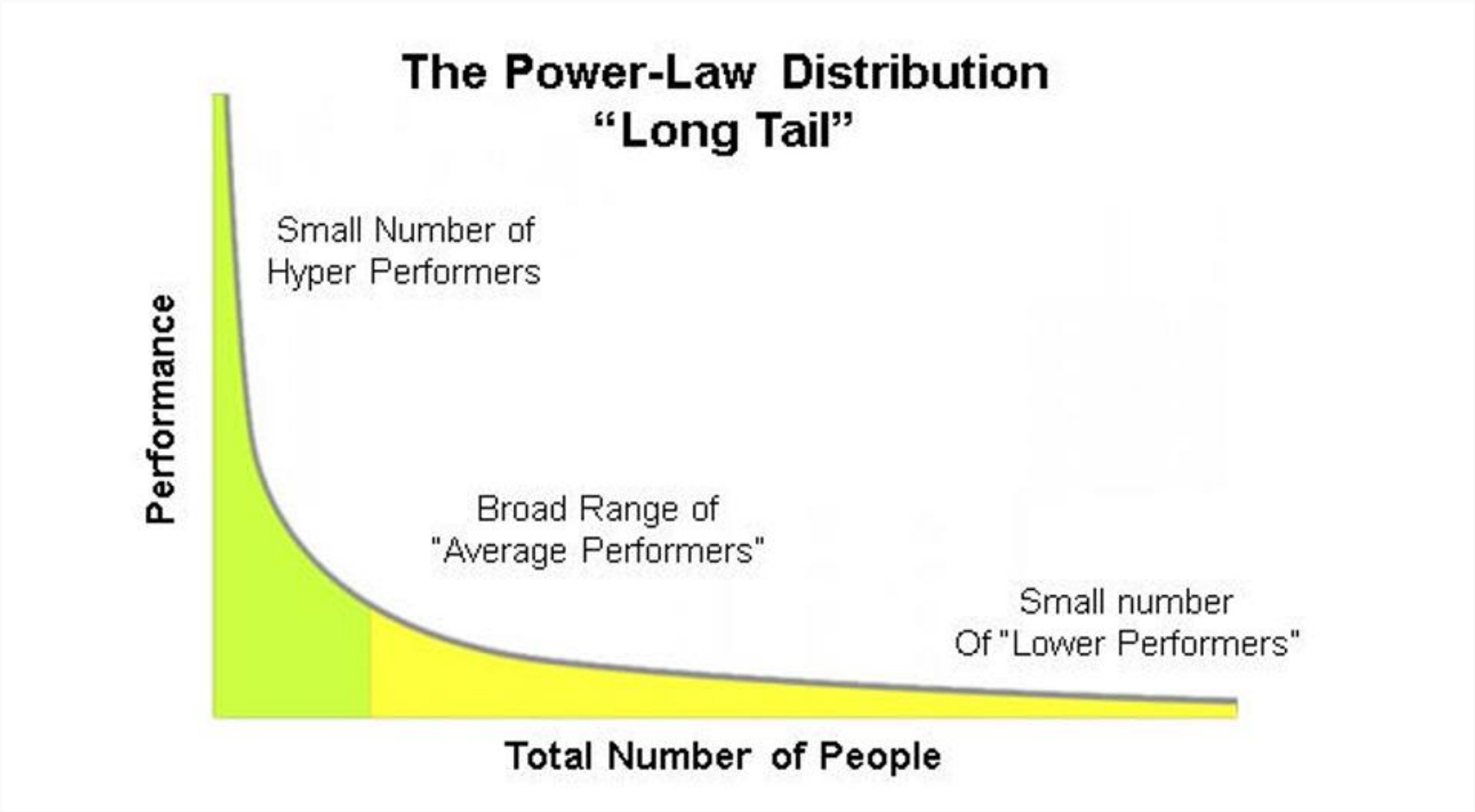
Interval

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])

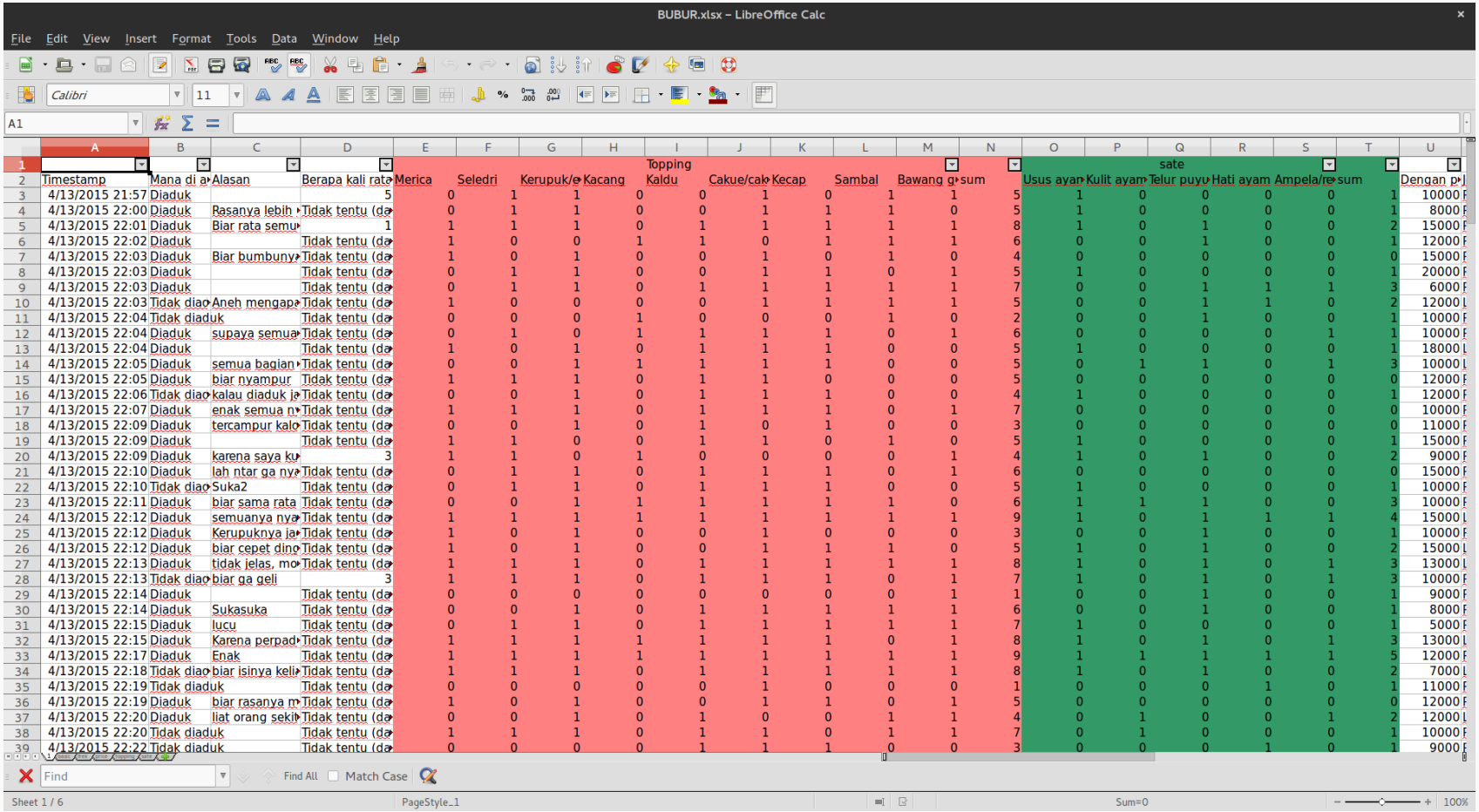
Distribusi yang Condong



Gambar 12: *Power law distribution* [Bersin, 2014]

Tipe Data

Data Matriks



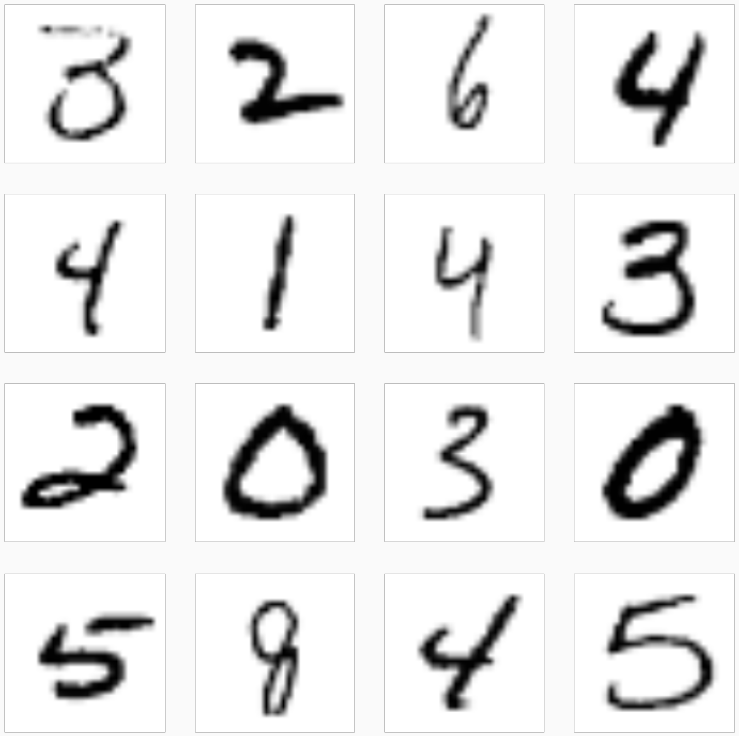
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1					Topping										sate						
2	Timestamp	Mana di a	Alasan	Berapa kali rata	Merica	Seledri	Kerupuk/e	Kacang	Kaldu	Cakue/cak	Kecap	Sambal	Bawang g	sum	Usus ayan	Kulit ayam	Telur puyu	Hati ayam	Ampela/re	sum	Dengan p
3	4/13/2015 21:57	Diaduk		5	0	1	1	0	0	1	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 dino	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 sekit	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

Gambar 13: 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

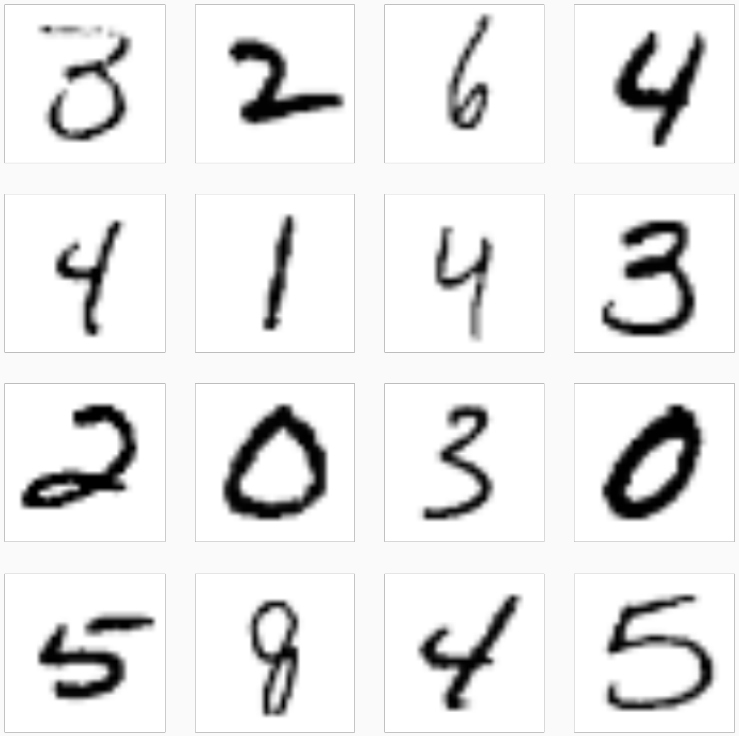
Gambar



Gambar 14: Contoh data MNIST [O'Shea, 2016]

1. Bagaimana cara merepresentasikan gambar?

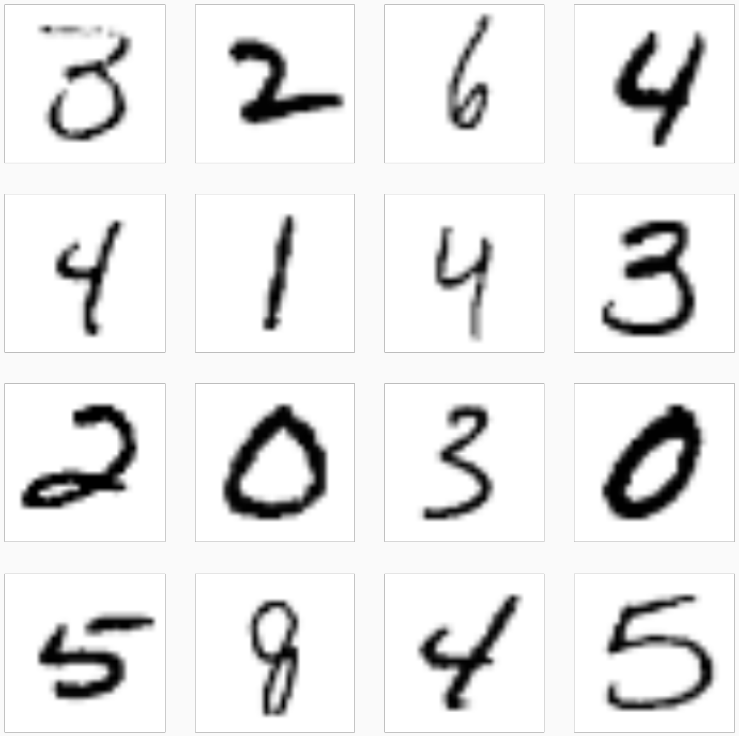
Gambar



Gambar 14: Contoh data MNIST [O'Shea, 2016]

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

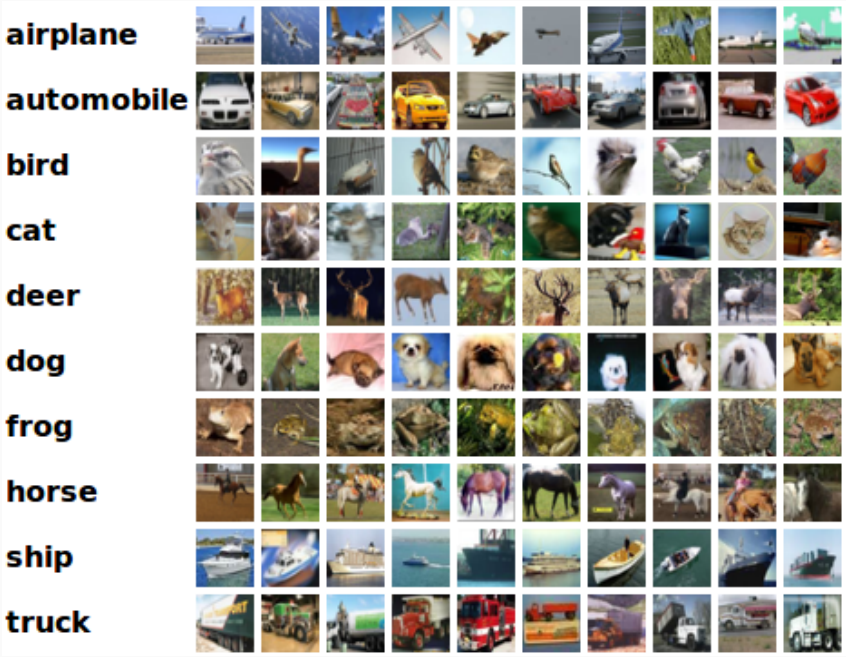
Gambar



Gambar 14: 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?

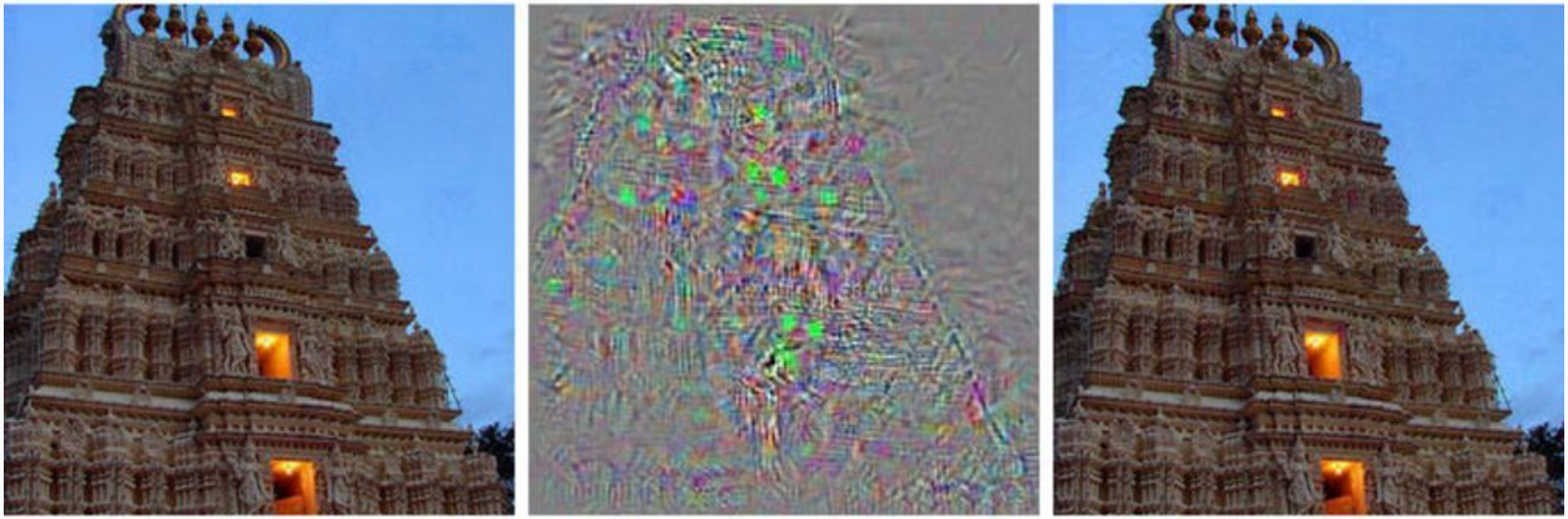
Gambar: Object Recognition



Gambar 15: 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 16: 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

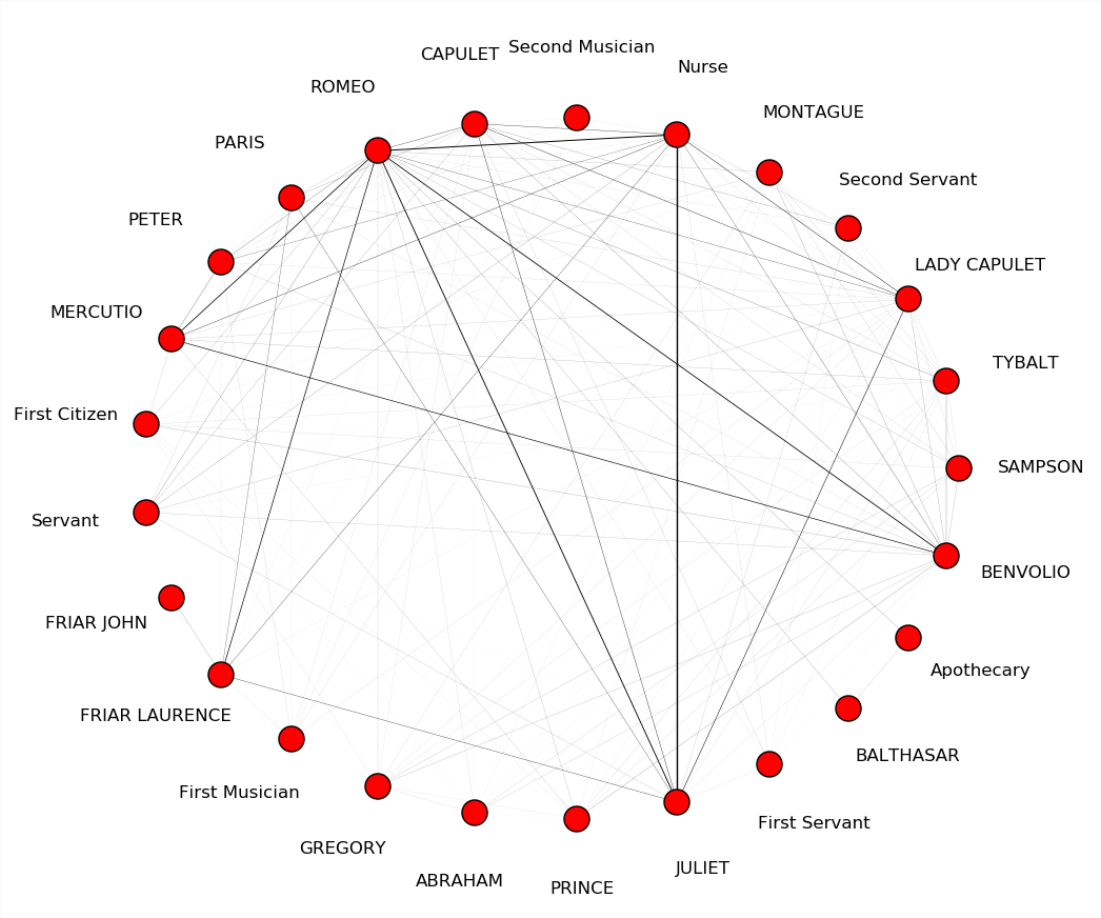
Kata sebagai Atribut Numerik

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

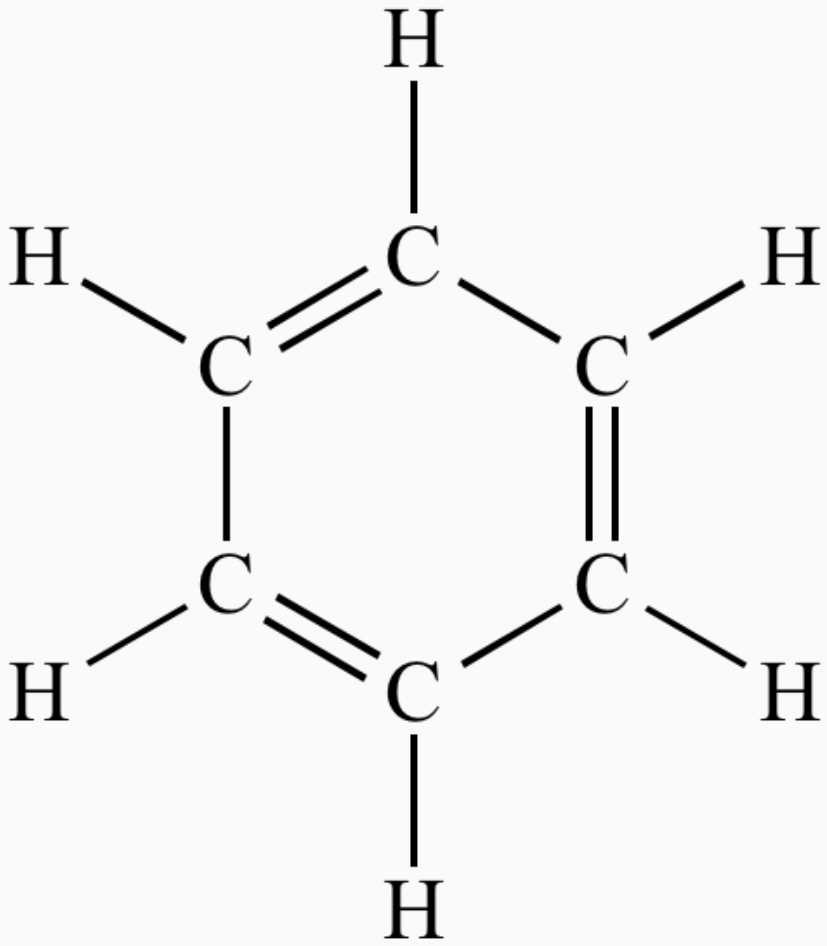
1. Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**
2. Bernilai 1 jika terdapat di contoh teks, 0 jika tidak
3. Dapat diubah menjadi frekuensi atau bobot (TF-IDF)
4. **Catatan:** Dimensinya bisa jadi sangat besar dan matriksnya akan menjadi *sparse*

Jejaring Sosial



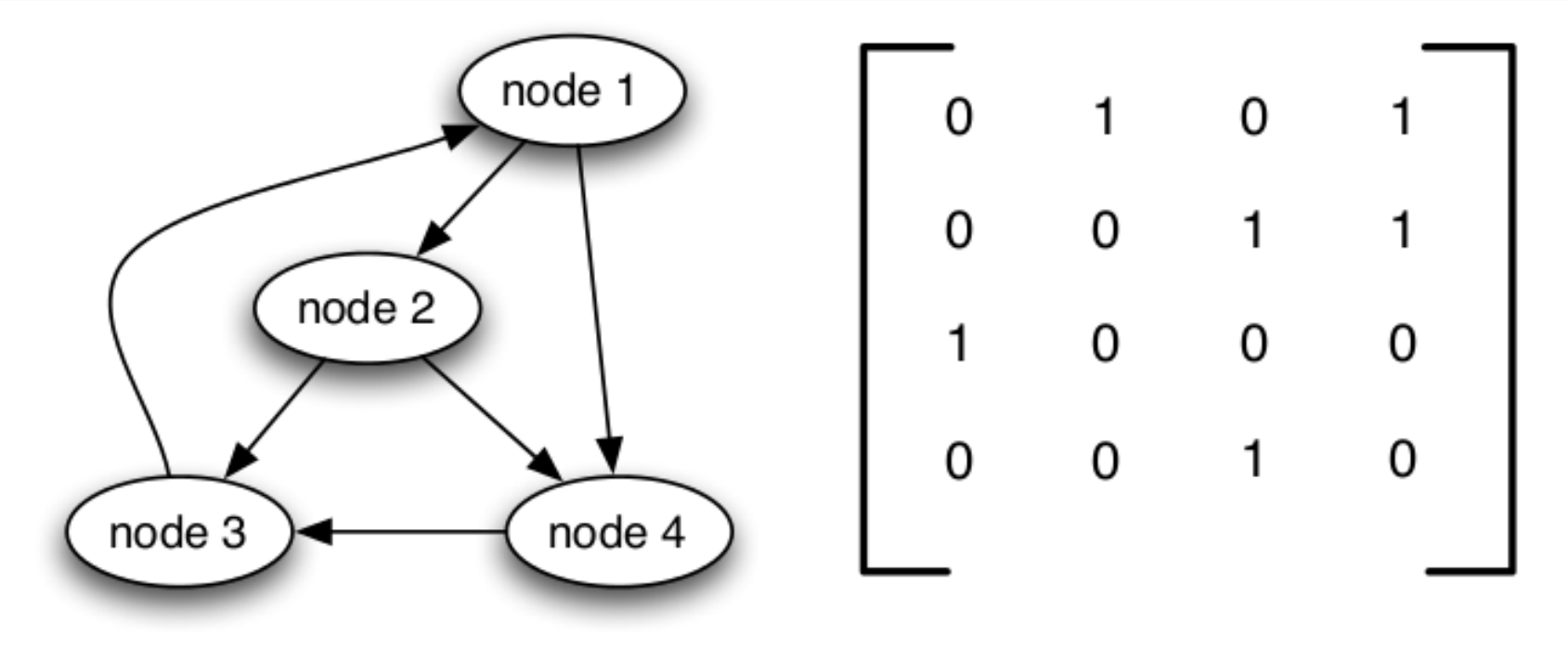
Gambar 17: Graf dari drama Romeo dan Juliet berdasarkan kemunculan karakter di satu babak yang sama

Struktur Kimia



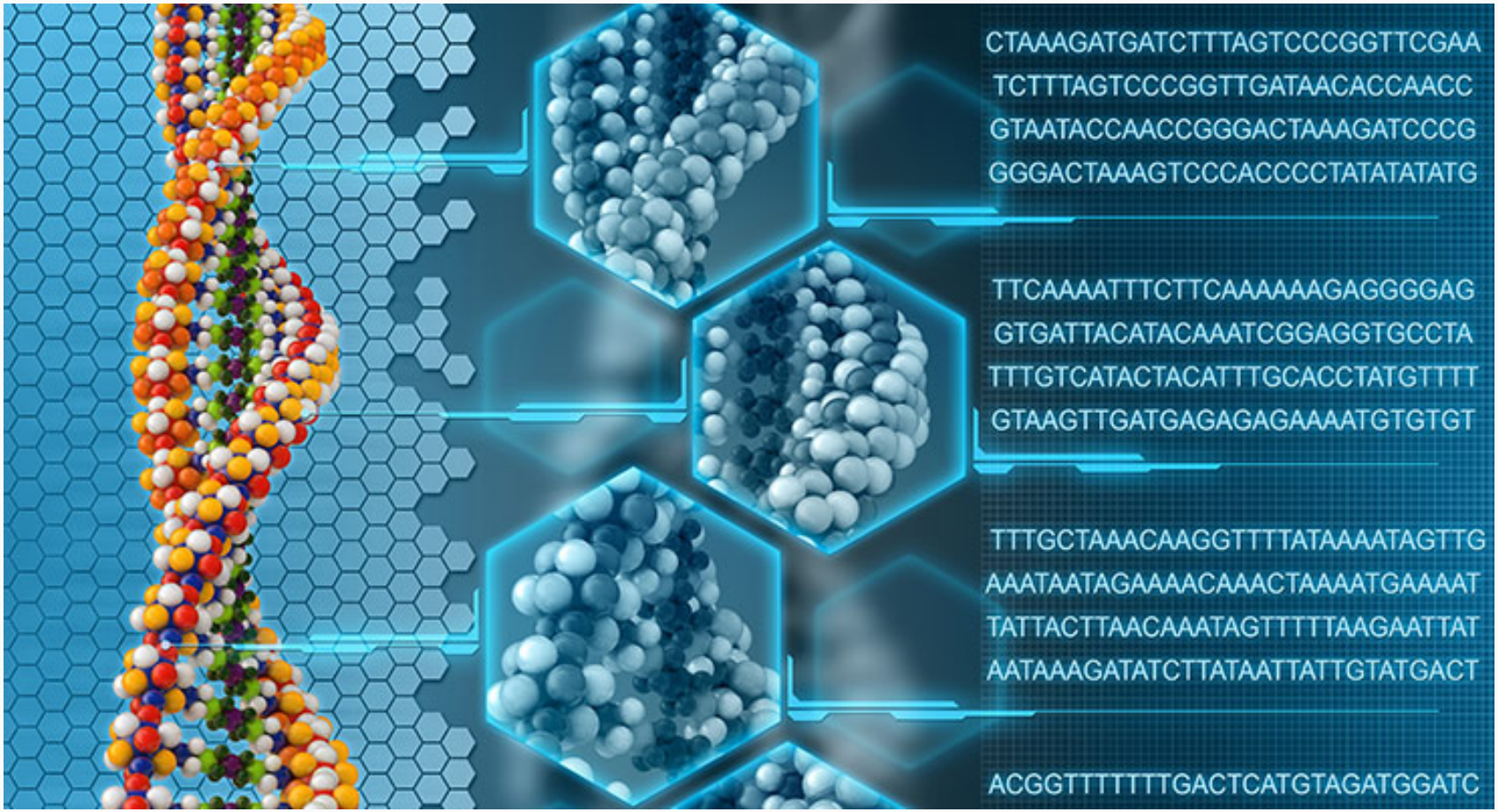
Gambar 18: Struktur kimia benzena [Hardinger, 2017]

Adjacency Matrix



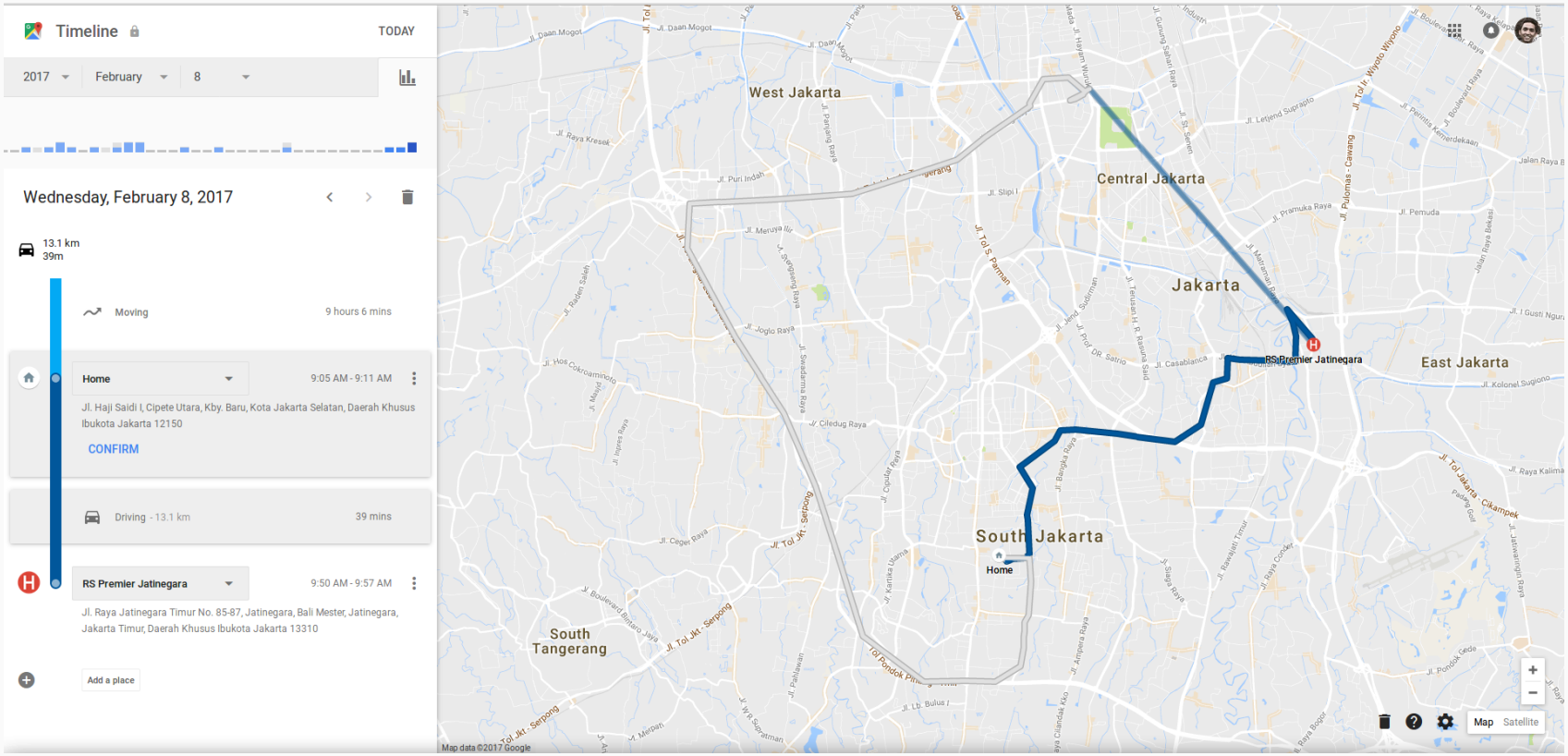
Gambar 19: Adjacency matrix dari graf [Easley dan Kleinberg, 2010]

Genomic Sequence



Gambar 20: Urutan genom [Global Biodefense, 2014]

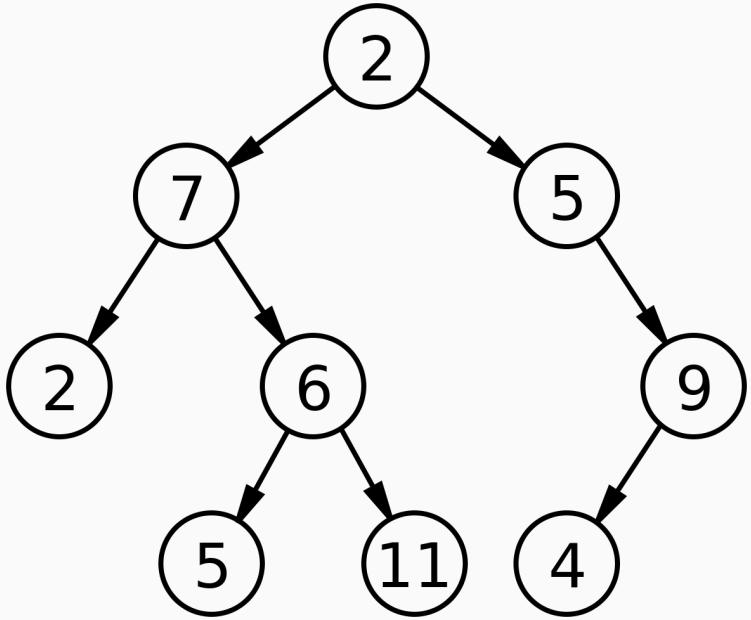
Spatio-Temporal



Gambar 21: Peta perjalanan seseorang yang direkam oleh Google Maps

Masalah pada Data

Dealing with Structures



Gambar 22: 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

Imbalanced 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

Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman (2014)

Mining of Massive Datasets

Cambridge University Press

Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal (2016)

Data Mining: Practical machine learning tools and techniques

Morgan Kaufmann

Andrew Beveridge and Jie Shan (2016)

Network of Thrones

Math Horizons, 23(4): 18-22

Referensi

David Robinson (14 September 2017)

Why is Python Growing So Quickly?

https://stackoverflow.blog/2017/09/14/python-growing-quickly/

Ramda Yanurzha (31 Mei 2017)

Berbagi Nama (Belakang)

https://medium.com/@ramda/berbagi-nama-belakang-f91b75c4aa39

Referensi

Victor Lavrenko (2010)

Text Technologies

http://www.inf.ed.ac.uk/teaching/courses/tts/pdf/crawl-2x2.pdf

Cyrille Rossant (2014)

Introduction to Machine Learning in Python with scikit-learn

http://ipython-books.github.io/featured-04/

Iain Murray (2011)

Oranges, Lemons and Apples dataset

http://homepages.inf.ed.ac.uk/imurray2/teaching/oranges_and_lemons/

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-performe

Tim O'Shea (Juli 2016)

MNIST Generative Adversarial Model in Keras

http://www.kdnuggets.com/2016/07/

Referensi

Alex Krizhevsky (2009)

Learning Multiple Layers of Features from Tiny Images

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

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

Referensi

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