

Data Mining

Pendahuluan

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

September 25, 2018

Selayang Pandang

① Administrasi

Tentang Perkuliahan
Referensi

② Konsep Data Mining

Definisi
Etika Data Mining

③ Tugas-tugas dalam Data Mining

Deskriptif
Klasifikasi
Regresi
Supervised Learning
Clustering
Asosiasi dan Sistem Rekomendasi

Administrasi

Mata Kuliah Terkait

Prerequisites

- Statistika & Probabilitas
- Matriks dan Ruang Vektor
- Kalkulus

Paralel/Saran

- Kecerdasan Buatan
- Pemrograman Python

Aturan Perkuliahan

- 1 Materi bisa dilihat di <http://uai.aliakbars.com/data-mining/>
- 2 Kuliah setiap hari Rabu, pukul 07.00-09.30
- 3 Bahasa/teknologi pengantar: Python, pandas, scikit-learn, Jupyter Notebook
- 4 Terdapat 5 tugas
- 5 Kuis yang tidak masuk komponen penilaian
- 6 Ujian Tengah dan Akhir Semester (tidak ada ujian perbaikan)
- 7 Komponen nilai: 50% tugas, 25% UTS, 25% UAS

Referensi

Buku dan materi daring yang bisa dijadikan referensi:

- ① VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media. ([tersedia online](#))
- ② 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](#))
- ③ Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press. ([tersedia online](#))
- ④ Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87. ([tersedia online](#))

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

Konsep Data Mining

Apa itu *Data Mining*?



Gambar: Dari <https://xkcd.com/1838/>

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*

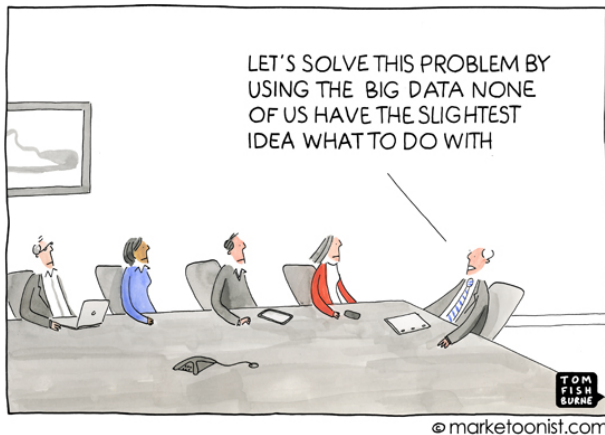
Keterkaitan dengan Machine Learning

- 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

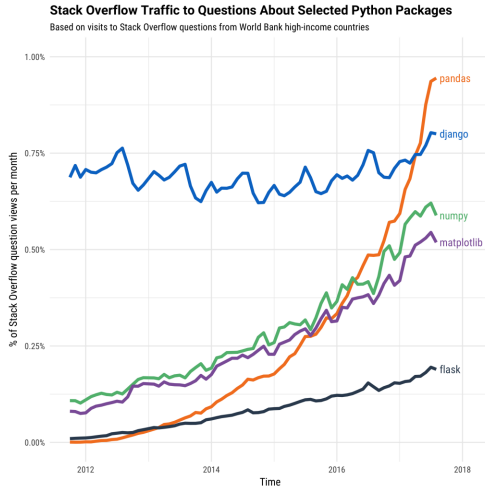
- Dalam prosesnya, algoritma *machine learning* sering digunakan untuk mempermudah proses *data mining*
- *Machine learning* dapat bekerja dengan baik jika pengetahuan yang kita miliki terbatas
- Jika polanya sudah *straightforward*, gunakan saja *if-then-else*!

Data Mining & Big Data



Gambar: Dari <https://marketoonist.com/2014/01/big-data.html>

Tren Data Mining



Gambar: 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

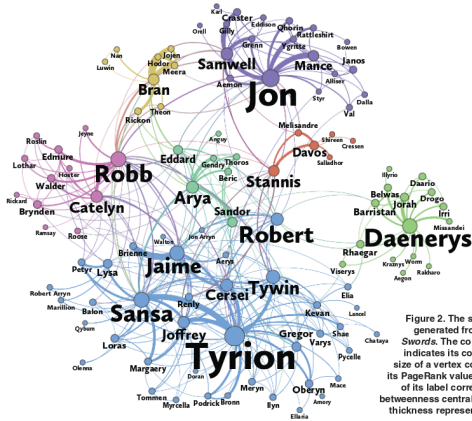
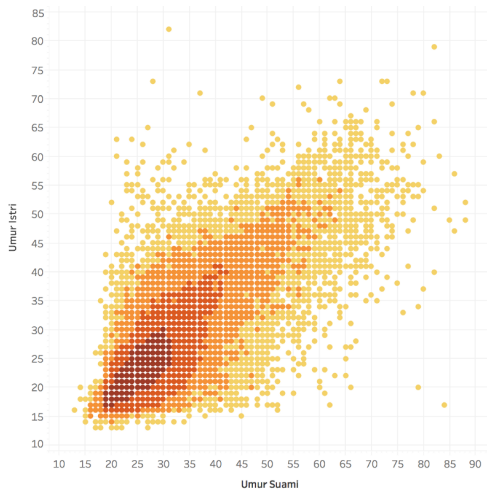


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

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

Heatmap



Gambar: Usia pernikahan [Yanurzha, 2017]

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/>)

Google

About



Google Dataset Search Beta

Search for Datasets



Try [boston education data](#) or [weather site:noaa.gov](#)

Gambar: Pencarian data semakin mudah

Sumber Data

Beberapa situs tidak menyediakan API untuk memberikan data karena:

- ❶ tidak dikembangkan sejak awal;
- ❷ tidak ingin datanya disebar, e.g. Instagram; atau
- ❸ hanya bisa diakses terbatas, e.g. Microdata BPS

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

“visible \neq accessible \neq storable \neq presentable”
[Lavrenko, 2010]

Bahan Bacaan

- 1 `https://blog.gojekengineering.com/tagged/data-science`
- 2 `https://airbnb.io/`
- 3 `https://research.netflix.com/articles`
- 4 `https://ai.googleblog.com/`

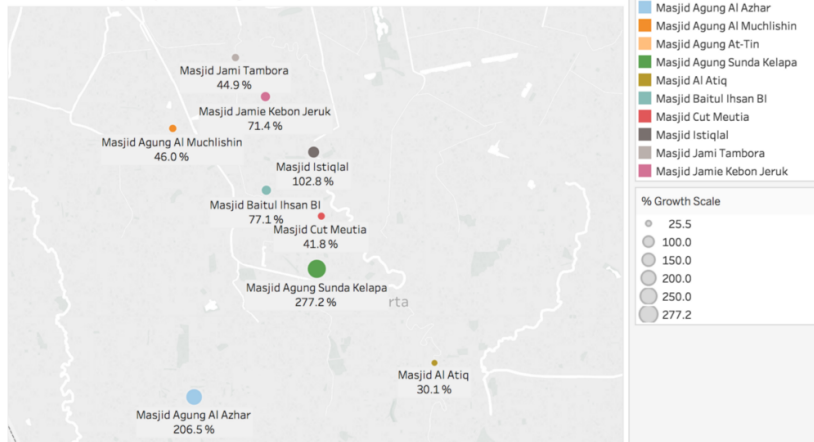
Tugas-tugas dalam Data Mining

Deskriptif

- 1 Melaporkan statistik, e.g. mean, median, modus
- 2 Visualisasi
- 3 Deskripsi perubahan tren

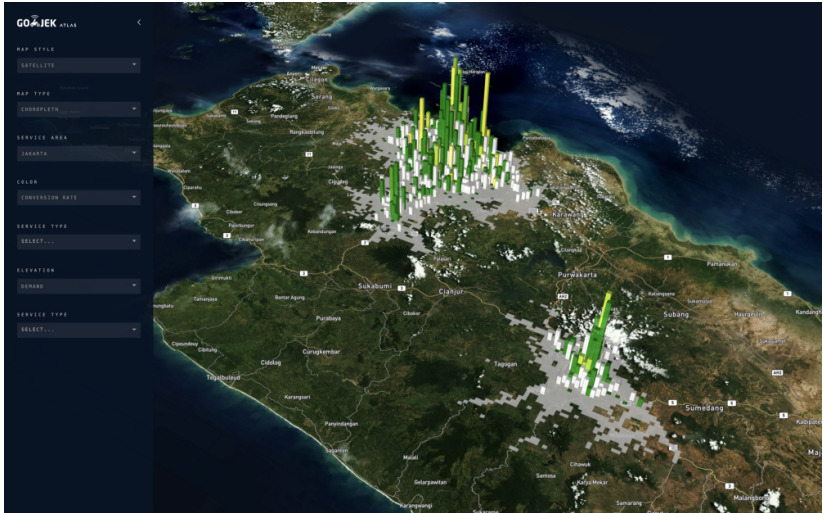
GO-FAST

Growth in Pickup Requests during Ramadan



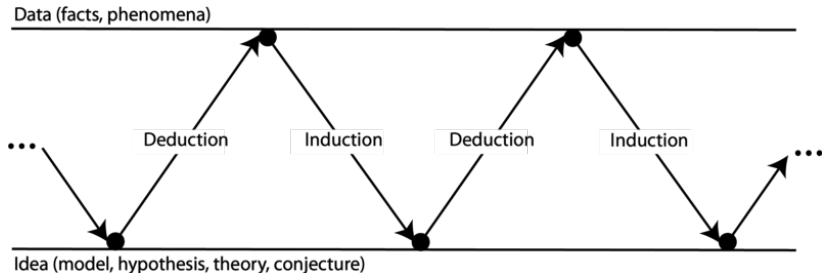
Gambar: Data penjemputan GO-JEK di bulan Ramadan 2017
[Widjaja, 2017]

GO-JEK Atlas



Gambar: Platform visualisasi geospasial *real-time* milik Go-Jek [Suhag, 2018]

Metode Ilmiah Deduksi-Induksi



Gambar: Proses pembelajaran iteratif oleh George Box [Govind, 2017]

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 $\{-1, 1\}$
- 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 $\{-1, 1\}$
- 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

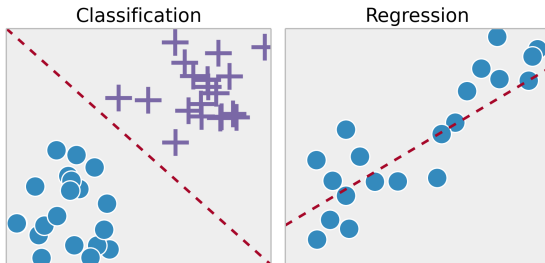
Regresi

- 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: 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

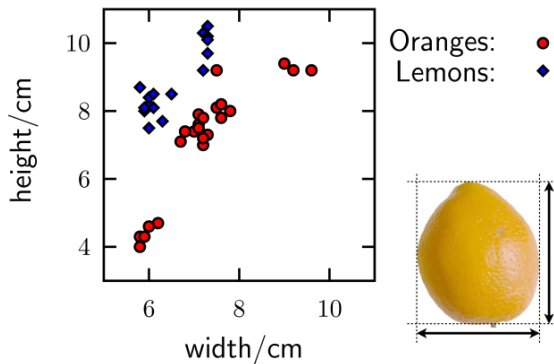
Clustering

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

Clustering

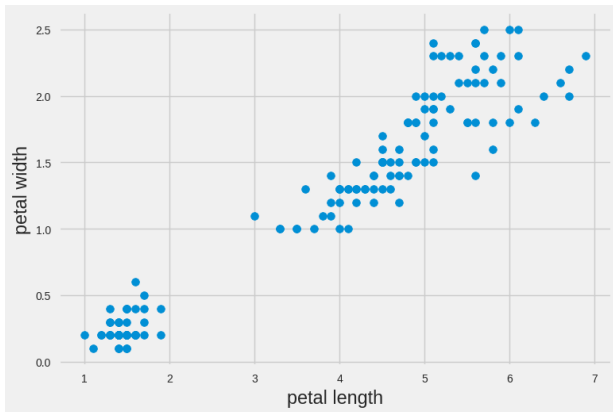
- ① Mencoba memberikan deskripsi terhadap data
- ② Tidak berhubungan dengan label
- ③ Menemukan pola yang “menarik” dalam data
- ④ Tidak mempunyai *evaluation metrics* yang pasti

Contoh Clustering



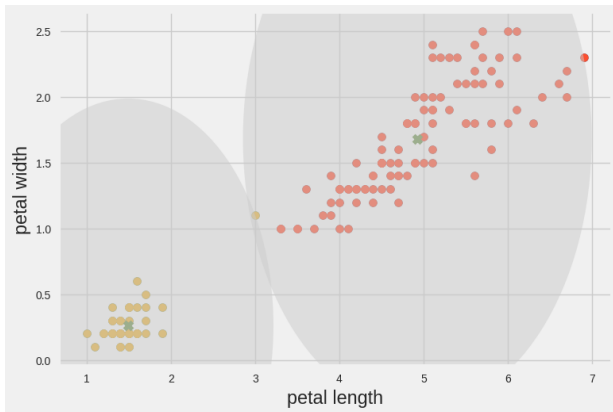
Gambar: *Clustering* buah lemon dan jeruk [Murray, 2011]

Contoh Clustering



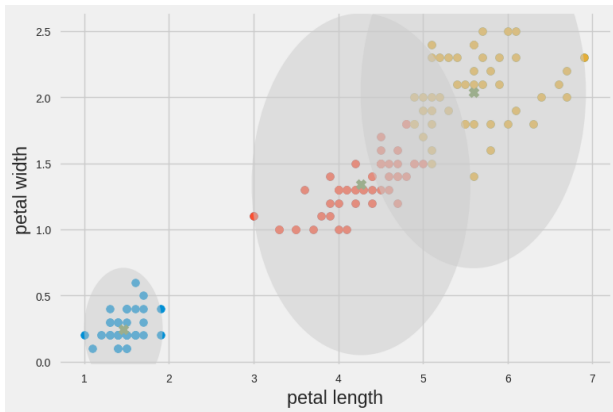
Gambar: Kluster dari dataset Iris

Contoh Clustering



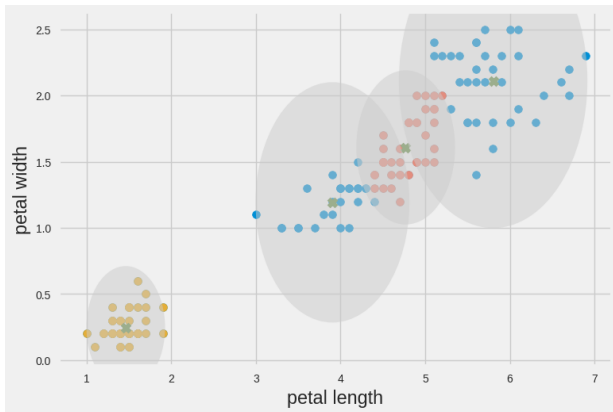
Gambar: Klaster dari dataset Iris

Contoh Clustering



Gambar: Kluster dari dataset Iris

Contoh Clustering



Gambar: Kluster 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

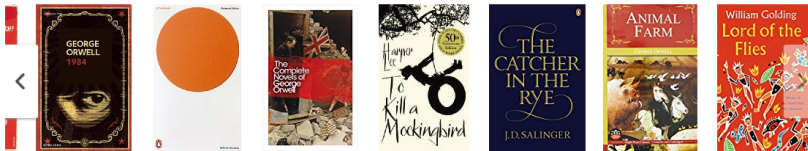
- ① Roti, soda, susu
- ② Bir, roti
- ③ Bir, soda, popok, susu
- ④ Bir, roti, popok, susu
- ⑤ Soda, popok, susu

Aturan yang ditemukan

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

Sistem Rekomendasi

More items to consider [See more](#)



Gambar: 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 ”

Kuis

- 1 Berikan masing-masing dua contoh kasus klasifikasi, regresi, dan *clustering*!
- 2 Apa yang menjadi atribut dan (jika ada) label dari contoh-contoh kasus tersebut?
- 3 Variabel seperti apa yang dapat dipakai oleh sistem rekomendasi berdasarkan konten dari aplikasi seperti Spotify?

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

[//stackoverflow.blog/2017/09/14/python-growing-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>



Victor Lavrenko (2010)

Text Technologies

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

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

Referensi



Crystal Widjaja (22 Juni 2017)

GO-FAST: The Data Behind Ramadan

[https://blog.gojekengineering.com/
go-fast-the-data-behind-ramadan-38037953561b](https://blog.gojekengineering.com/go-fast-the-data-behind-ramadan-38037953561b)



Ravi Suhag (19 Februari 2018)

ATLAS: GO-JEKs real-time geospatial visualization platform

[https://blog.gojekengineering.com/
atlas-go-jeks-real-time-geospatial-visualization-platform-1cf5e168](https://blog.gojekengineering.com/atlas-go-jeks-real-time-geospatial-visualization-platform-1cf5e168)



Nirmal Govind (13 Juni 2017)

A/B Testing and Beyond

[https://medium.com/netflix-techblog/
a-b-testing-and-beyond-improving-the-netflix-streaming-experience-](https://medium.com/netflix-techblog/a-b-testing-and-beyond-improving-the-netflix-streaming-experience-)

Referensi



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/

Terima kasih