Data Mining for Unstructured Data

Pertemuan 2
MK Data Mining II

M. N. Fakhruzzaman, S.Kom., M.Sc.
Ratih Ardiati Ningrung, S.Si., M.S., M.Stat.
Malikhah, S.Kom., M.Kom.

Program Studi S1 Teknologi Sains Data
Fakultas Teknologi Maju dan Multidisiplin
Universitas Airlangga Indonesia

Type of Data Structure

Data Terstruktur

outlook	temperature	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Data Tidak Terstruktur

Pada data terstruktur sebuah record memiliki fitur atau kolom yang jumlahnya sama dengan record lainnya. Setiap fitur pada record memiliki nilai yang bisa berupa angka atau kategori. Angka bisa berupa bilangan bulan atau desimal. Sedangkan contoh nilai kategori adalah pria dan wanita, benar dan salah, dan lain-lain. Setiap fitur atau kolom akan bernilai dengan tipe data yang sama, sebagai contoh untuk kolom fitur1 berisi nilai angka saja. Sehingga tidak mungkin dalam sebuah fitur berisi dengan tipe data yang berbeda, sebagai contoh angka dan kategori.

Data Semi Terstruktur

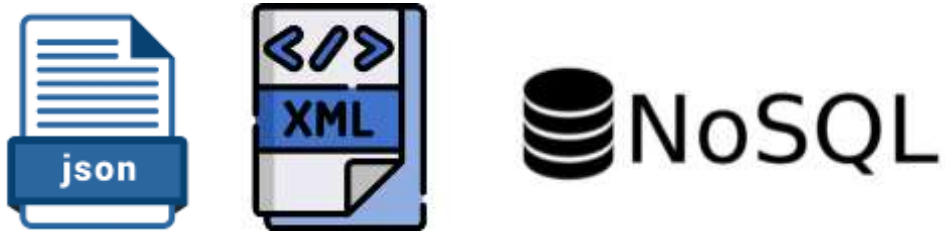
```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

Type of Data Structure

Data Terstruktur

Data Tidak Terstruktur

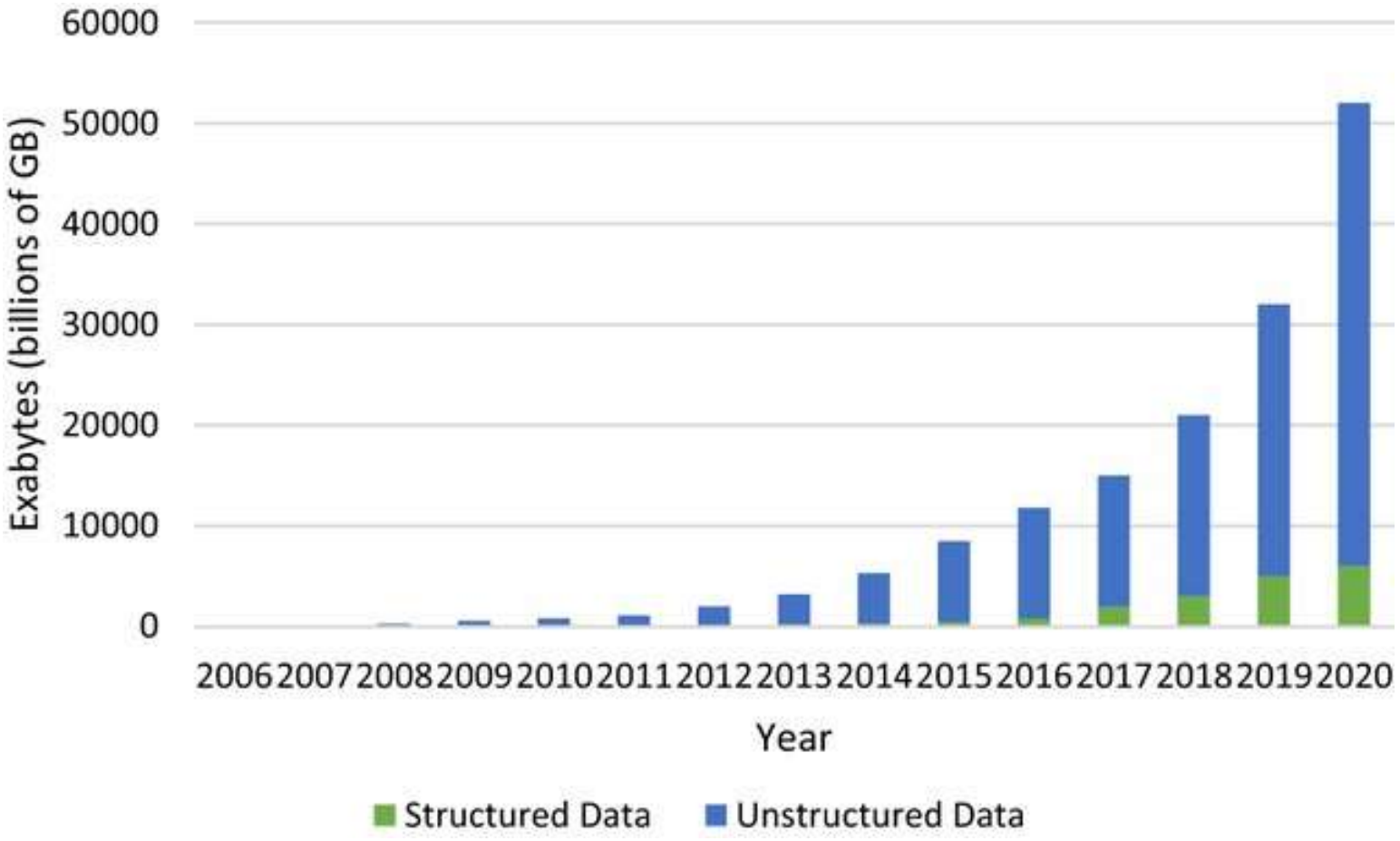
Data Semi Terstruktur



Key Differences

Data Terstruktur	Data Tidak Terstruktur
Defined Data	Undefined Data
Qualitative Data	Quantitative Data
Stored in Data Warehouses	Stored in Data Lakes
Easy to analyze	Hard to analyze
Predefined format	Variety format

Problem with Unstructured Data



Sumber : The role of structured and unstructured data managing mechanisms in the Internet of things by Poupak Azad et al.

Which One is Data Mining?

a. Mencari nomor kontak di HP
b. Mengelompokkan kontak di HP berdasarkan Namanya
c. Mengunduh dokumen di Google
d. Mengelompokkan dokumen berdasarkan hasil pencarian di Google berdasarkan konteksnya
e. Mengunduh gambar buah-buah dari Internet
f. Mengelompokkan buah berdasarkan gambar buah yang sudah diunduh
g. Mengunduh lagu di internet
h. Mengelompokkan lagu berdasarkan genre lagu

Challenges in Data Mining

- Scalability
- High Dimensionality
- Data yang kompleks dan Heterogen
- Kepemilikan dan Distribusi Data
- Analisis data tidak terstruktur

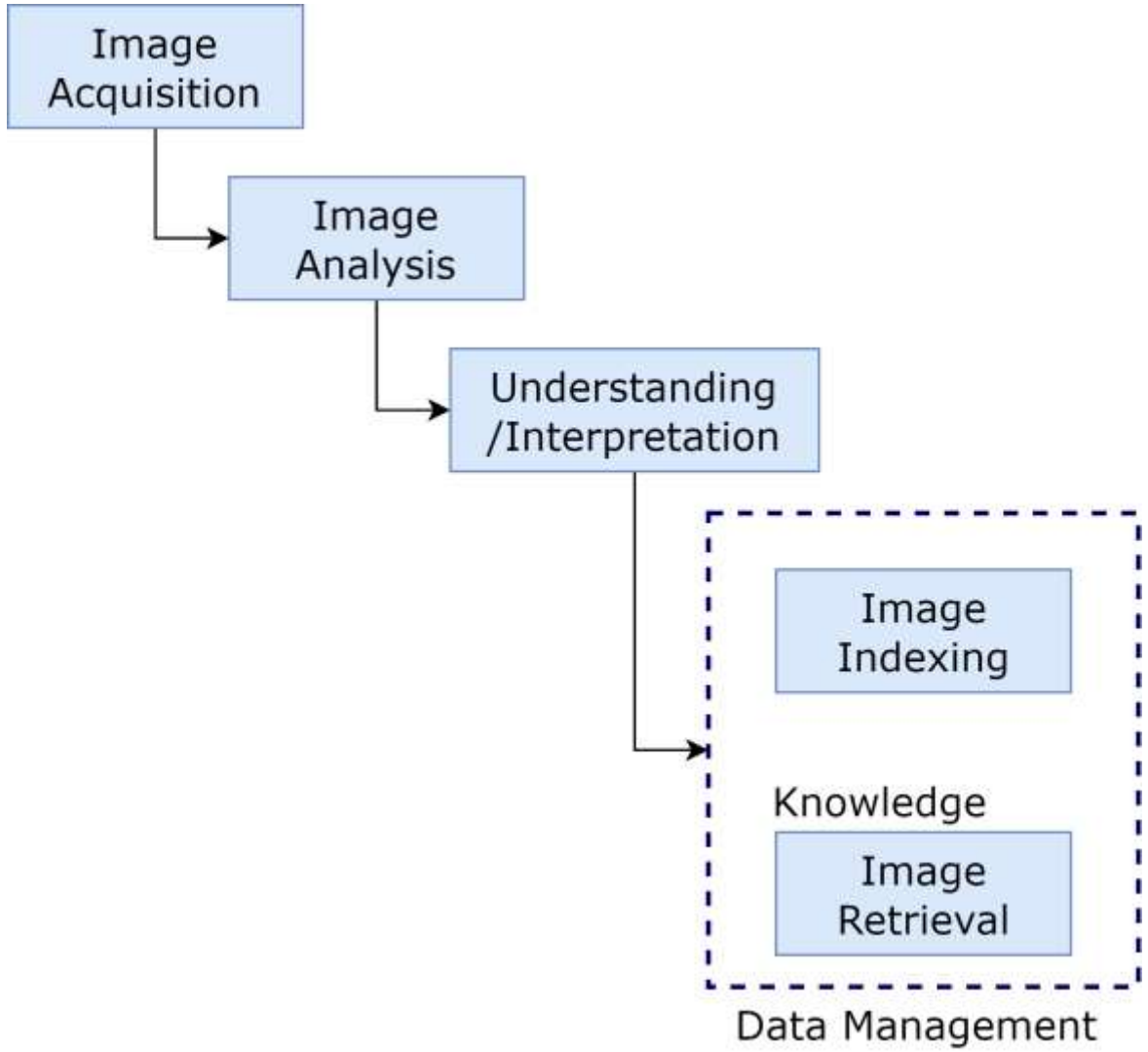
Data Mining for Unstructured Data

- Text Mining
- Image Processing and Data Mining (Image Mining)
- Audio Data Mining
- Social Media Data Processing and Analytics
- ...

Image Mining

Image Mining

- Citra digital adalah citra yang terdiri dari picture element (pixel), dimana pada tiap pixel terdapat nilai diskrit dari intensitas citra
- Image Mining adalah proses ekstraksi citra, mendapatkan hubungan antar citra, dan mendapatkan pola dari suatu citra untuk mendapatkan informasi yang berguna



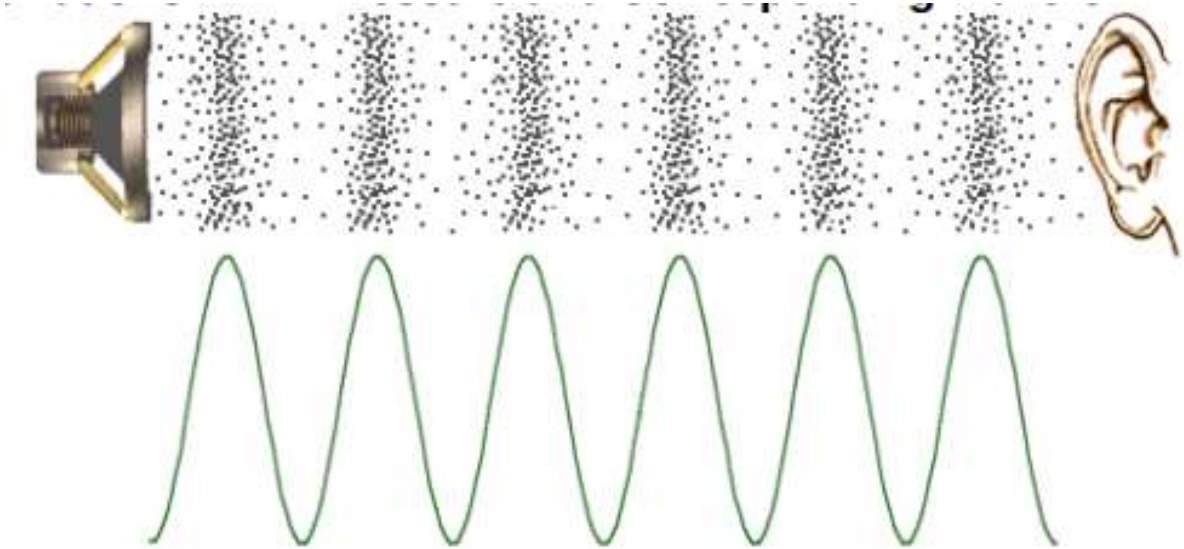
Application of Image Processing

- Fingerprint recognition
- Satellite Images
- Meteorology
- Radiology
- Ultrasonic Imaging
- Surveillance
- Face Recognition
- ...

Audio Data Mining

Audio Data Mining

- Audio adalah suara atau hasil reproduksi suara.
- Menggunakan data sinyal audio untuk mendapatkan pola atau fitur agar mendapatkan informasi yang berguna



Audio Data Mining

- Image Processing atau visual data mining menggunakan data citra yang membutuhkan konsentrasi yang lebih untuk mengamati pola dan fitur yang ada didalamnya, hal tersebut tentu saja sangat melelahkan.
- Jika pola dapat diubah menjadi suara dan music, kita bisa mendengar nada, ritme, dan melodi tanpa harus mengamati gambar untuk mengidentifikasi pola. Hal ini dapat meringankan beban konsentrasi visual dan lebih santai daripada visual mining. Oleh karena itu audio mining merupakan pelengkap untuk visual mining.
- Tantangan: banyaknya background noise dan cross talk.

Audio Data Mining Process

- 4 komponen:
  1. Audio Indexing: pencarian data audio yang efisien
  2. Speech Processing and Recognition: mengidentifikasi unit kata atau fonem yang mungkin mucul.
  3. Feature extraction: ekstraksi fitur audio (zero crossing, spectral centroid, spectral roll-off, spectral flux, dll)
  4. Mining: Classification, Clustering, dll

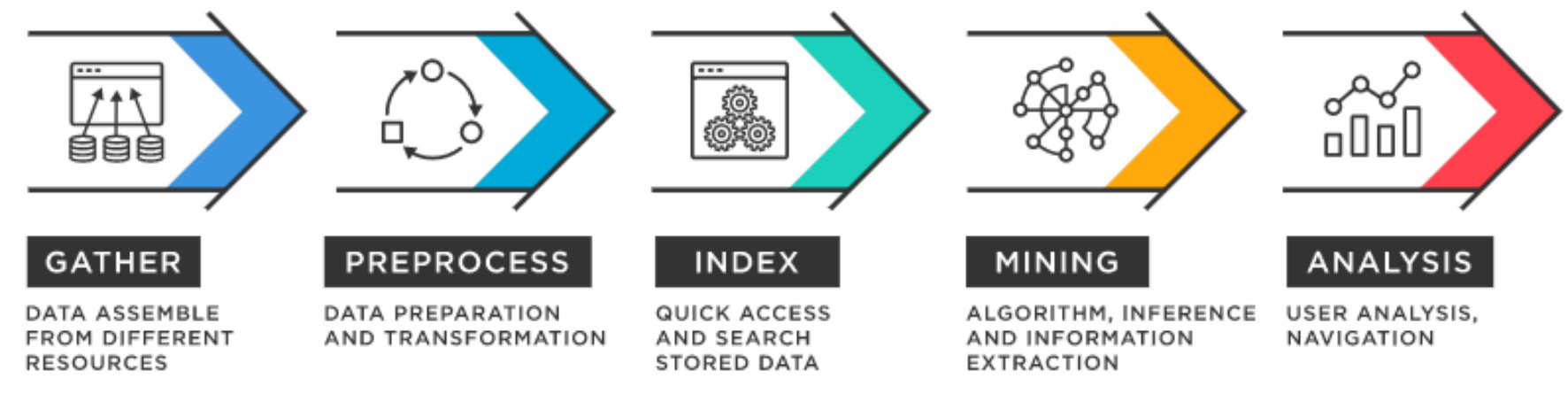
Application of Audio Data Mining

- Automated Transcription of Audio
- Understanding Customer opinions accurately
- Captioning (subtitling) of Video (youtube, dll)
- Speech recognition system
- Voice dialer on car phones
- Voice personal assistant (Amazon Alexa, Google Assistant, Android Cortana, etc)

Text Mining

Text Mining

- Jumlah data teks berkembang dengan pesat.
- Proses mendapatkan informasi berkualitas tinggi dari teks.



Natural Language Processing

- Text Mining menggunakan Natural Language Processing (NLP) untuk mengubah data tidak terstruktur menjadi data terstruktur untuk dianalisis lebih lanjut.
- NLP juga membantu text mining agar mesin bisa ‘membaca/mengerti’ teks.

The Importance of Text Mining

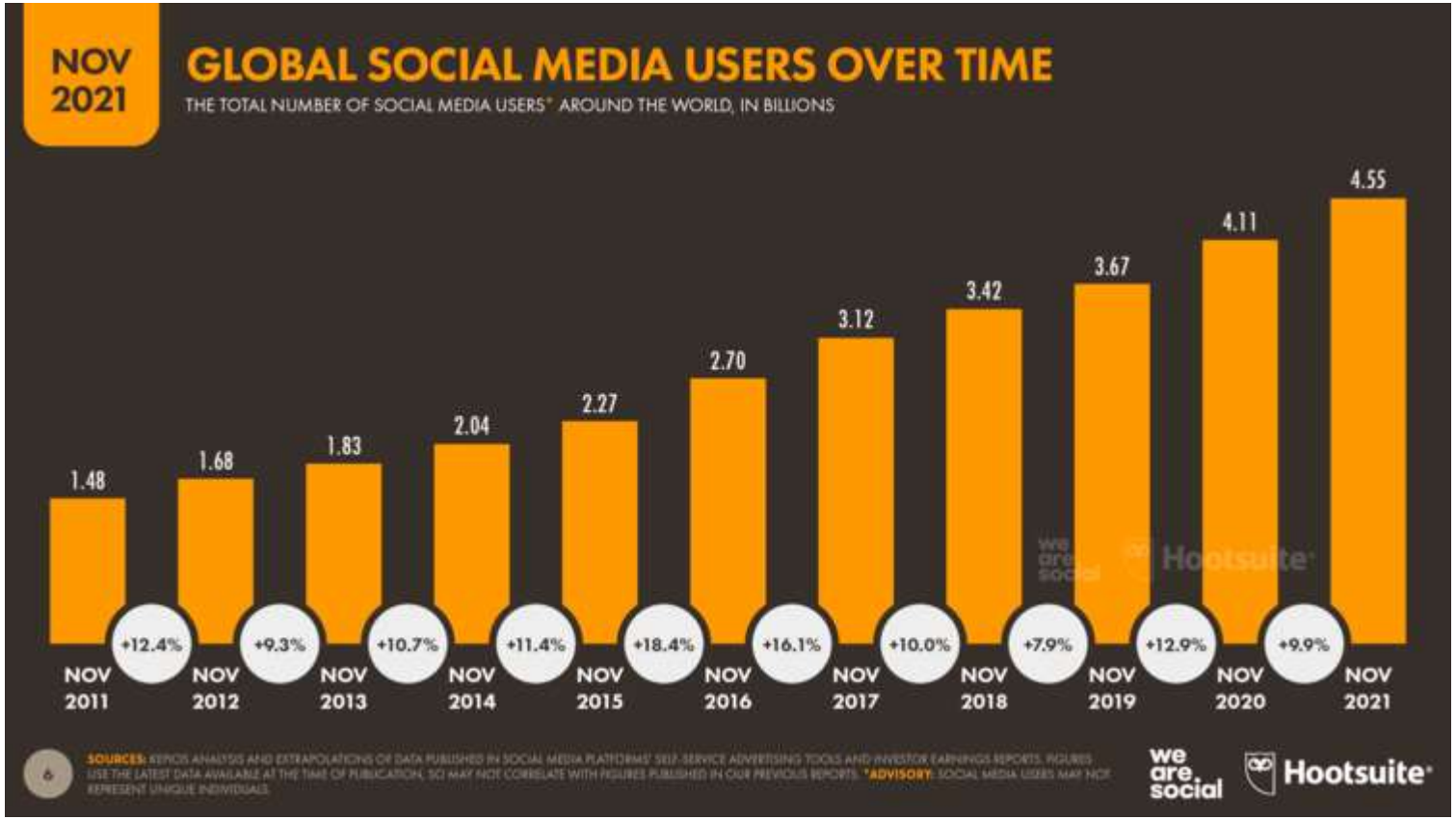
- Persaingan yang semakin ketat di bidang bisnis, yang membuat organisasi untuk mencari lebih banyak solusi untuk menjadi terdepan dalam persaingan
  - Amazon menggunakan text mining untuk mengidentifikasi fitur dari suatu produk
  - Google dan Bing menggunakan Text Mining untuk mengidentifikasi spam dan melakukan filter konten

Text Mining Use Case

Social Media Data Mining and analytics

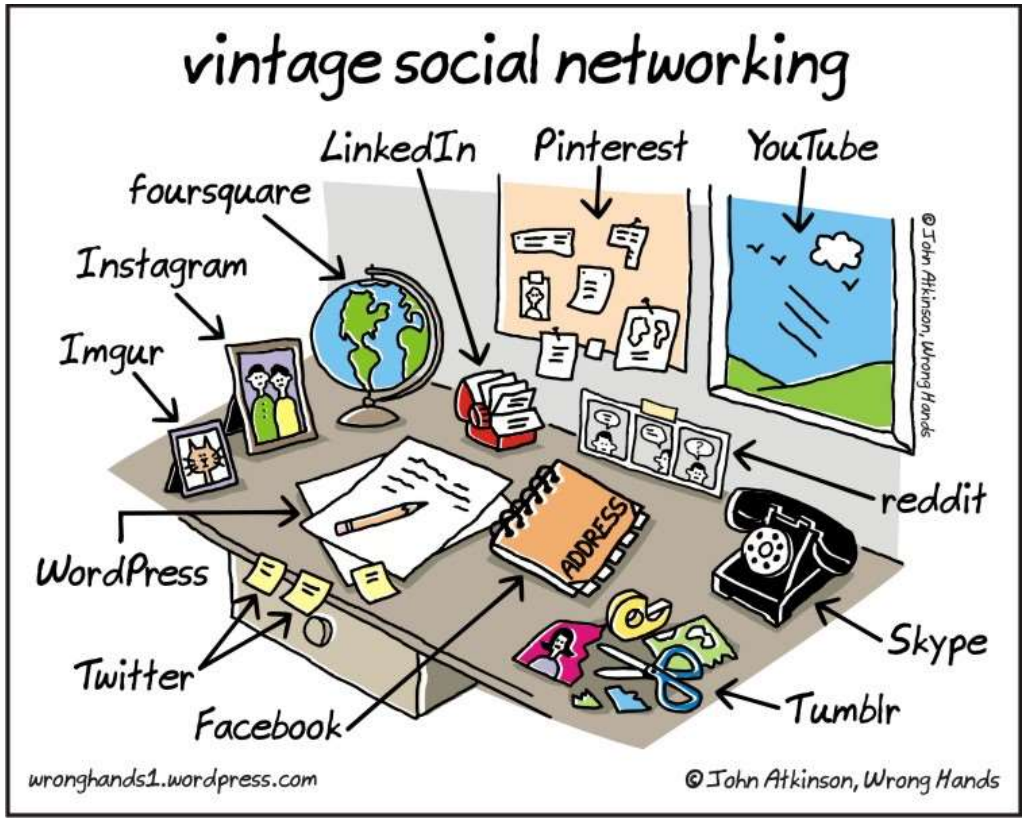
Social Media Growth

Sumber: https://www.javatpoint.com/social-media-data-mining



Sumber: https://datareportal.com/reports/a-decade-in-digital

Social Media for Business



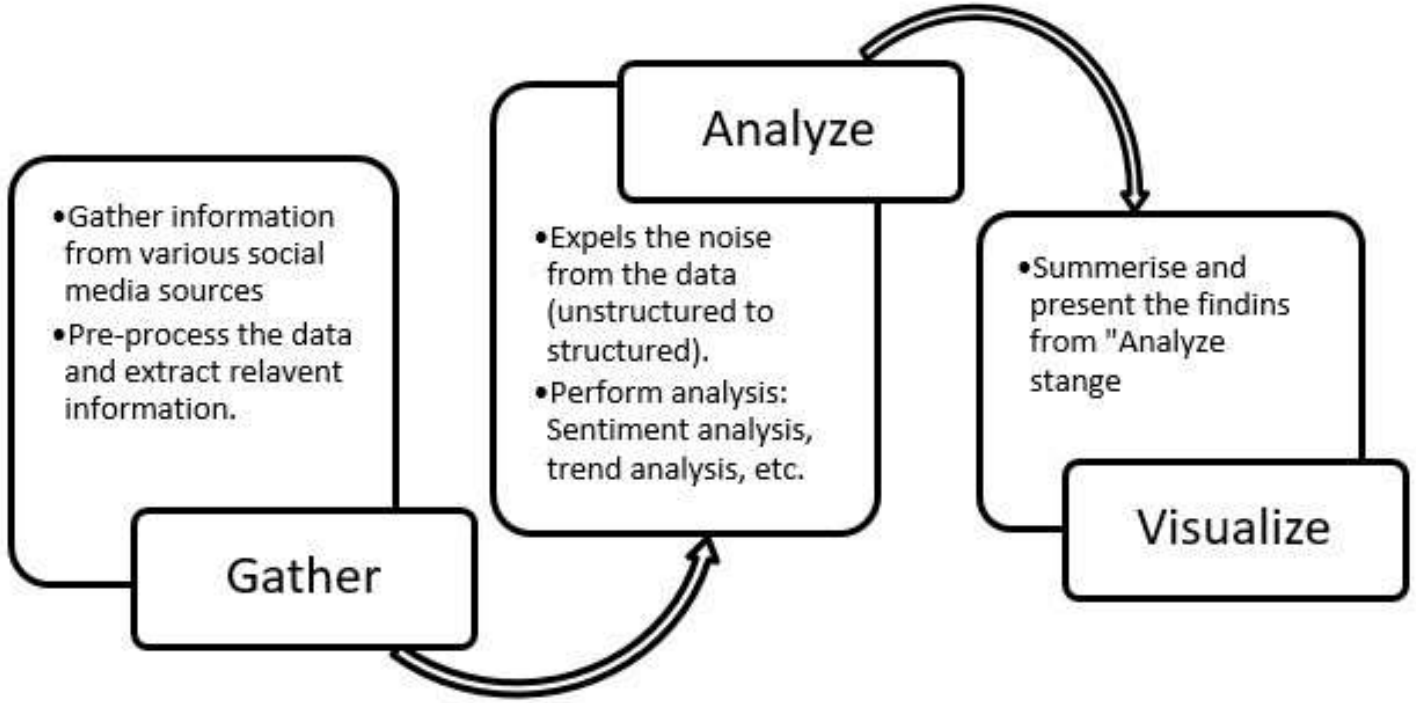
Sumber: https://www.nationalbusinesscapital.com/blog/leverage-social-media-promote-brand/

Motivations for Data Mining in Social Media

- Bagian dari Text Mining
- Data yang dapat diakses melalui platform social media dapat memberi insight tentang jejaring sosial dan masyarakat yang sebelumnya tidak diketahui.
- Tantangan utama pada social media data:
  - Social media data berukuran besar, contoh: user aktif dari facebook berjumlah 2.41 milyar.
  - Kumpulan data pada social media banyak mengandung noise, contoh: blog spam, tweet tidak penting, dll
  - Social media data sangat dinamis. Modifikasi dan pembaruan terjadi dalam waktu singkat.
- Tanpa menerapkan teknik data mining akan sulit melakukan pengumpulan social media data dan menganalisanya.

Social Media Data Mining and Analytics

- Proses mengumpulkan data dari berbagai sumber social media dan bagaimana memprosesnya agar mendapatkan informasi yang berguna



Sumber: The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R by Ahmed Imran Kabir et al

Social Media Data Mining and Analytics Use Case

- Targeted marketing campaign
- Market research
- Predictive analytics
- Influencer Marketing
- Monitoring of brand reputation
- Trend Analysis
- ...

Terima Kasih

Referensi:

Camastra, F. dan Vinciarelli, A., 2015, Machine Learning for Audio, Image, and Video Analysis Theory and Applications 2nd edition, Springer, London.
Marinai, S. dan Fujisawa, H., 2008, Machine Learning in Document Analysis and Recognition, Springer, Berlin Heidelberg.
Bird, S., Klein, E., dan Loper, E., 2009, Natural Language Processing with Python, O'Reilly Media Inc., USA.
Soumya Sen, Anjan Dutta, dan Nilanjan Dey, 2019, Audio Processing and Speech Recognition Concepts, Techniques and Research Overviews, Springer.
Dengsheng Zhang, 2019, Fundamentals of Image Data Mining, Springer.