

Big Data Fundamentals: Concept & Ecosystem



Apa itu Big Data

- Big data: banyak data
 - Karena makin banyak 'Jejak digital' □ dianalisis agar tidak hanya menjadi 'sampah'.
- Contoh hasil analisis:
 - Kebiasaan konsumen
 - Minat terhadap jenis buku, musik, film, dll.
 - Pola iklim/cuaca
 - Semuanya disebut 'insight' □ informasi penting berarti.
- Analisis harus selalu memiliki tujuan yang jelas.



Apa itu Big Data

“The basic idea behind the phrase ‘Big Data’ is that everything we do is increasingly leaving a digital trace (or data), which we can use and analyze to become smarter. The driving forces in this brave new world are access to ever-increasing volumes of data and our ever-increasing technological capability to mine that data for commercial insights.”

– Bernard Marr



Apa itu Big Data

“Big Data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation.”

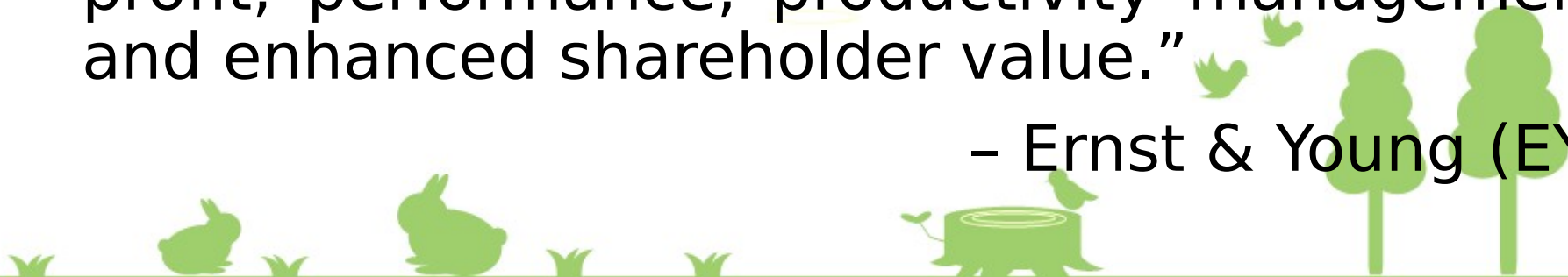
– Gartner



Apa itu Big Data

“Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools and machines. It requires new, innovative, and scalable technology to collect, host and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management and enhanced shareholder value.”

– Ernst & Young (EY)



Apa itu Big Data

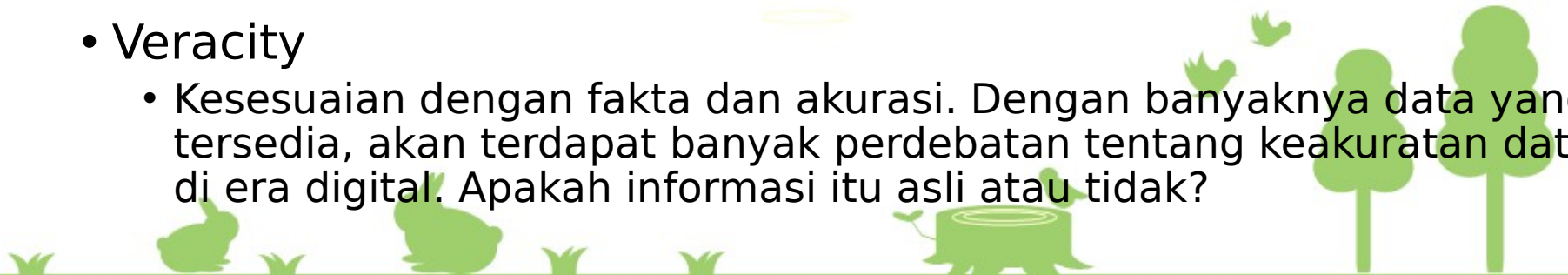
“Big Data is a collection of data from traditional and digital sources inside and outside a company that represent a source of ongoing discovery and analysis.”

– Lisa Arthur

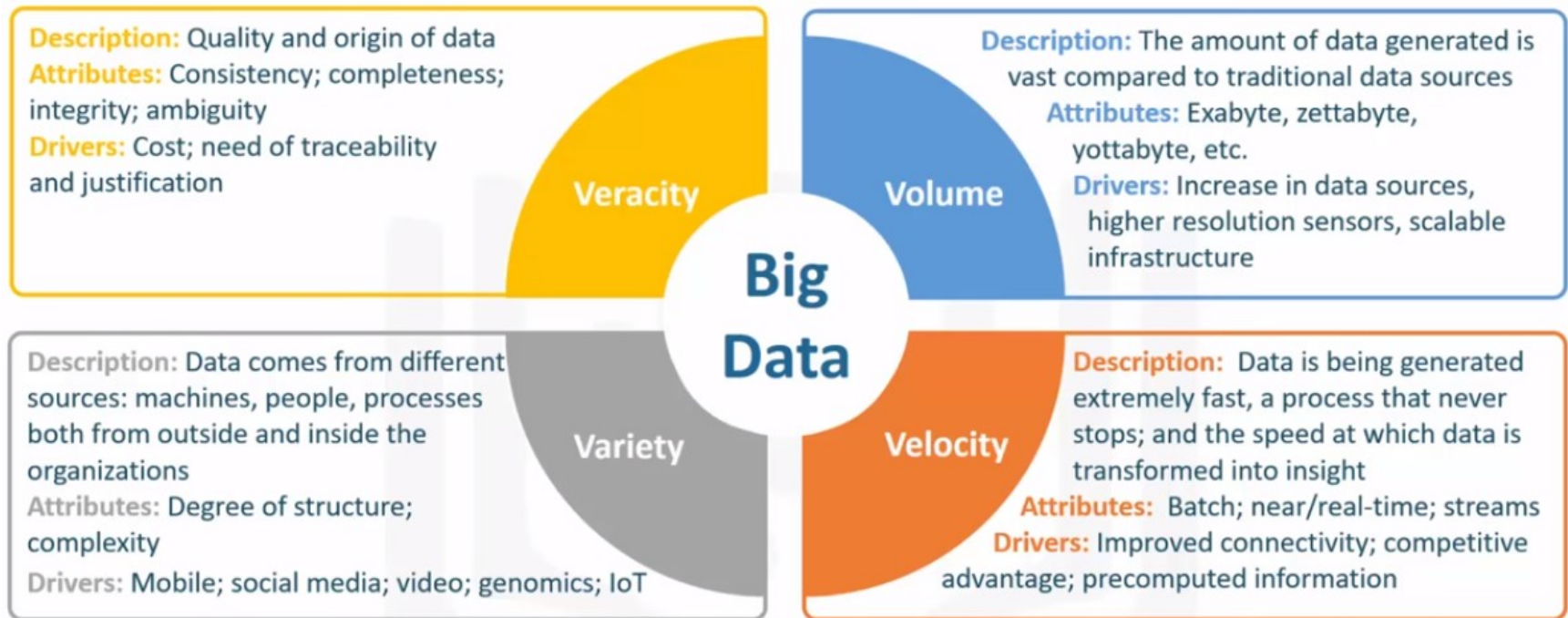


4 V's of Big Data

- Velocity
 - Kecepatan data, atau kecepatan akumulasi data
- Volume
 - Skala data, atau peningkatan jumlah data yang disimpan.
- Variety
 - Keragaman data. Terdapat data terstruktur (structured data) yang dimuat dalam kolom-kolom dan baris-baris (atau database relasional) dan data tidak terstruktur (unstructured data) yang tidak diatur dengan cara yang telah ditentukan sebelumnya, misalnya Tweet, posting blog, gambar, angka, dan bahkan data video.
- Veracity
 - Kesesuaian dengan fakta dan akurasi. Dengan banyaknya data yang tersedia, akan terdapat banyak perdebatan tentang keakuratan data di era digital. Apakah informasi itu asli atau tidak?



5 V's of Big Data



Source: EY, 2014; ITU, 2013

Velocity

- Setiap 60 detik:
 - Video dengan total durasi sekitar 400 jam diupload ke Youtube
 - 2,430,555 likes di Instagram
 - 972,222 swipe di Tinder
- Seluruh data tersebut dihasilkan **setiap menit**.



Volume

- Sebagian besar populasi manusia memiliki perangkat digital yang menghasilkan, menerima, dan menyimpan data.
- Sebagian memiliki lebih dari 1 perangkat (misal: HP, PC, Laptop, Tablet, dll)
- Setiap harinya kita menghasilkan sekitar **2.5 Quintilion Bytes data**.



Variety

- Tipe-tipe data
 - Teks
 - Gambar
 - Video
 - Suara
 - Wearables devices data
 - Data-data lainnya dari IoT device



Veracity

- 80% data yang beredar merupakan unstructured data.
- Data tersebut harus dikategorikan, dianalisis, dan divisualisasikan untuk menentukan apakah data tersebut akurat dan dapat dipercaya.



Value

- Value yang dihasilkan dari penerapan Big Data dapat berupa:
 - Keuntungan materi (profit)
 - Keuntungan non-materi (medical & social benefit)
 - Kepuasan (customer/employee/personal satisfaction)



Pengaruh Big Data

- Pada 2011, McKinsey & Company mengatakan bahwa Big Data akan menjadi kunci utama dari kompetisi yang mendukung pertumbuhan produktivitas dan inovasi.
- Pada 2013, UPS (United Parcel Service) mengumumkan bahwa mereka menggunakan data dari customer, driver, dan juga kendaraan untuk menentukan rute baru yang menghemat waktu, uang, dan bahan bakar.



Pengaruh Big Data

Recommendation Engines

amazon.com

Amazon's recommendations are based on what the user has bought in the past, the items he/she has in the virtual shopping cart, items he/she has rated and liked, and what other customers have viewed and purchased.

NETFLIX



Recommendation engines merupakan salah satu penerapan Big Data yang paling umum, yaitu untuk memberikan rekomendasi berdasarkan preferensi dan kebiasaan pengguna.

Pengaruh Big Data

Virtual personal assistants



Google Now

Siri knows what users mean when they ask her questions, she knows where they are, what time they are talking about and can use this information to look for restaurants of a particular type of food and check whether there reservations are available.

Google Now makes recommendations before users ask for them, especially when it is linked up to the user's calendar and location sensing is enabled on the user's phone. Google Now knows where the user is and where he/she needs to be, it can tell users about things like traffic or the weather before you even ask for it.

Virtual personal assistant juga merupakan penerapan dari Big Data.

Siri menggunakan Big Data untuk menyusun jawaban atas pertanyaan-pertanyaan pengguna.

Google Now memberikan rekomendasi berdasarkan Big Data pada perangkat pengguna.

Pengaruh Big Data

Salah satu perusahaan yang memanfaatkan penggunaan Big Data adalah Netflix. Dengan penerapan Big Data, Netflix dapat melakukan analisis terhadap kebiasaan pengguna yang pada akhirnya dapat dijadikan dasar untuk menentukan film yang disukai pengguna, bahkan memprediksi bagaimana kesuksesan suatu film sebelum dirilis.

Online streaming – Netflix

Data:

- >> Time of day when movies/shows are watched
- >> Record of when users pause, rewind and fast forward
- >> User ratings
- >> Searches
- >> Type of device

HOUSE  of CARDS

Pengaruh Big Data

- E-commerce di memanfaatkan Big Data untuk mendapatkan pangsa pasar, dengan mempertimbangkan kejenuhan pasar dan customer yang selektif.
- Contoh: E-Commerce di China

E-commerce in China

Facts:

- >> E-commerce accounts for more than 13% of China's total retail sales of consumer goods
- >> About 90% of Internet users and 70-80% of consumers as a whole are shopping online in top tier cities
- >> Chinese consumers often visiting 4 to 5 sites before reaching a purchase decision

淘宝网
Taobao.com

1号店



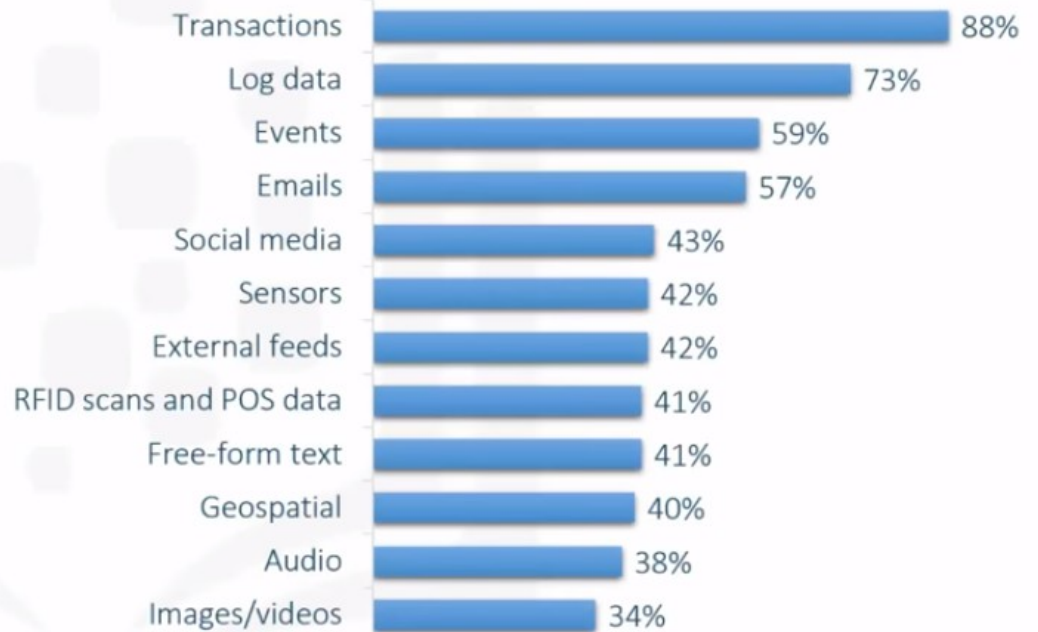
Source of Big Data

There are three major sources of Big Data:

- People-generated data
- Machine-generated data
- Business-generated data

Big data sources

Multiple answers allowed



Big Data Platform

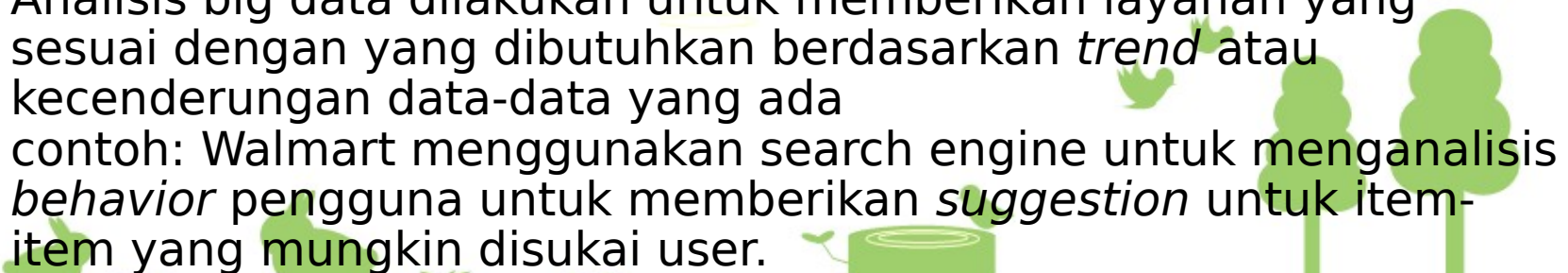
6 aspek utama yang harus ada dalam Big Data Platform:

1. *Integration*

Sebuah platform big data idealnya dapat mengelola seluruh data.
contoh : Hadoop Distributed File System (HDFS) menyimpan data dari banyak lokasi dan membuat sebuah pusat penyimpanan dan pemrosesan seluruh data tersebut.

2. *Analysis*

Analisis big data dilakukan untuk memberikan layanan yang sesuai dengan yang dibutuhkan berdasarkan *trend* atau kecenderungan data-data yang ada
contoh: Walmart menggunakan search engine untuk menganalisis *behavior* pengguna untuk memberikan *suggestion* untuk item-item yang mungkin disukai user.

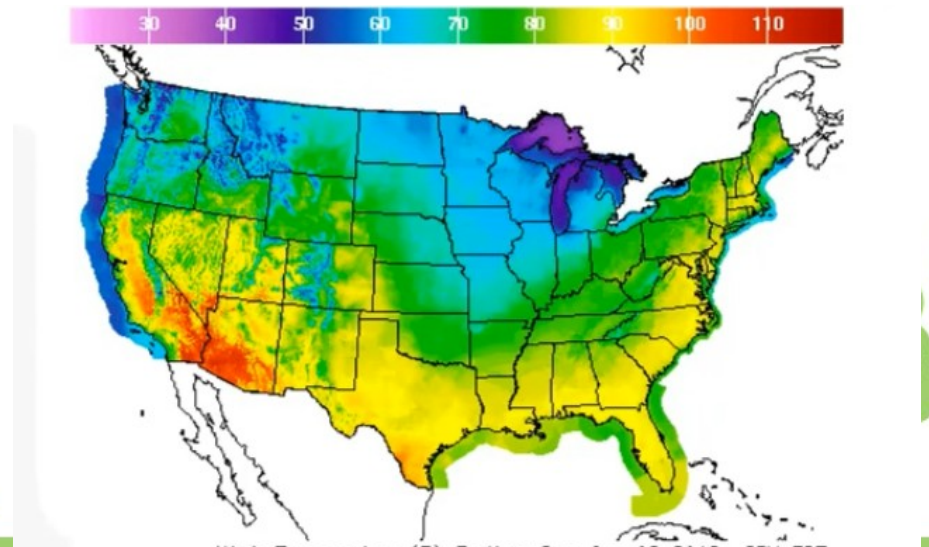


Big Data Platform

3. *Visualization*

Big data harus dapat divisualisasikan dalam bentuk grafis agar dapat mudah dipahami oleh siapapun, terutama untuk para pengguna yang tidak memahami teknis.
Contoh

Visualisasi temperatur suatu daerah pada peta, user akan lebih mudah memahaminya dibandingkan jika hanya menampilkan data dalam tabel



Big Data Platform

4. *Workload Optimization*

Optimalisasi sumber daya *open source* untuk meningkatkan efisiensi penyimpanan dan pemrosesan data

5. *Security*

Privasi data setiap orang harus dijamin dan dilindungi dengan berkembangnya penerapan big data. Dalam penerapan big data, harus selalu dipertimbangkan bagaimana data diperoleh, digunakan, diproses, serta direpresentasikan. Selain itu organisasi/perusahaan juga harus menerapkan aturan control dan *privacy policies* untuk menjamin keamanan data dan user.



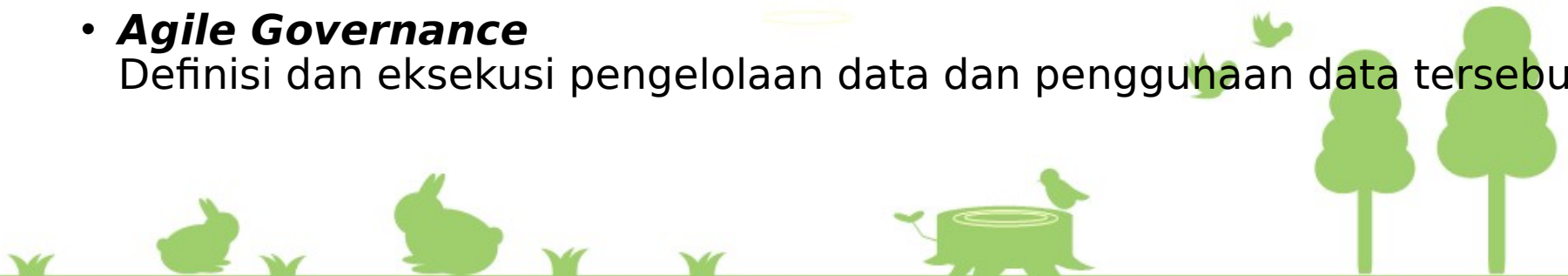
Big Data Platform

6 aspek utama dalam Big Data Platform (cont'd):

6. Governance

Pengelolaan dan pengorganisasian dalam big data memerlukan 3 hal:

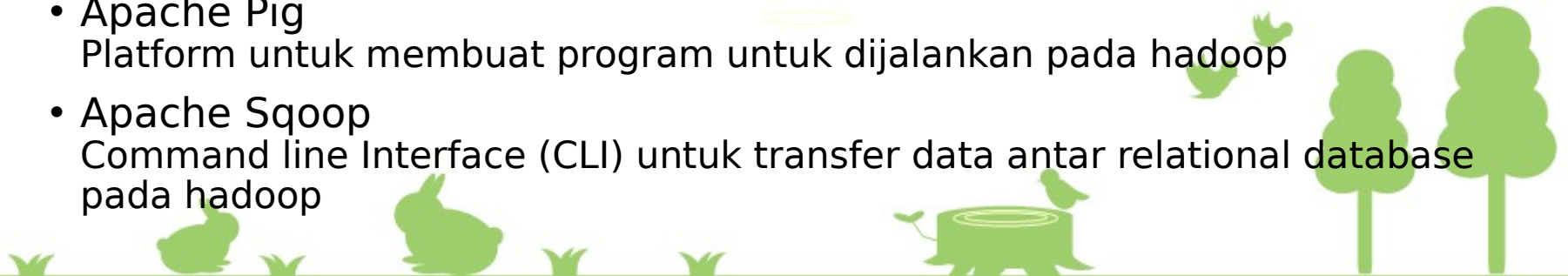
- ***Automated integration***
memberikan akses data yang mudah dimanapun lokasi data tersebut
- ***Visual context***
memberikan kemudahan pengkategorian, *indexing*, dan *discovery* pada big data untuk mengoptimalkan penggunaannya
- ***Agile Governance***
Definisi dan eksekusi pengelolaan data dan penggunaan data tersebut



Big Data and Data Science

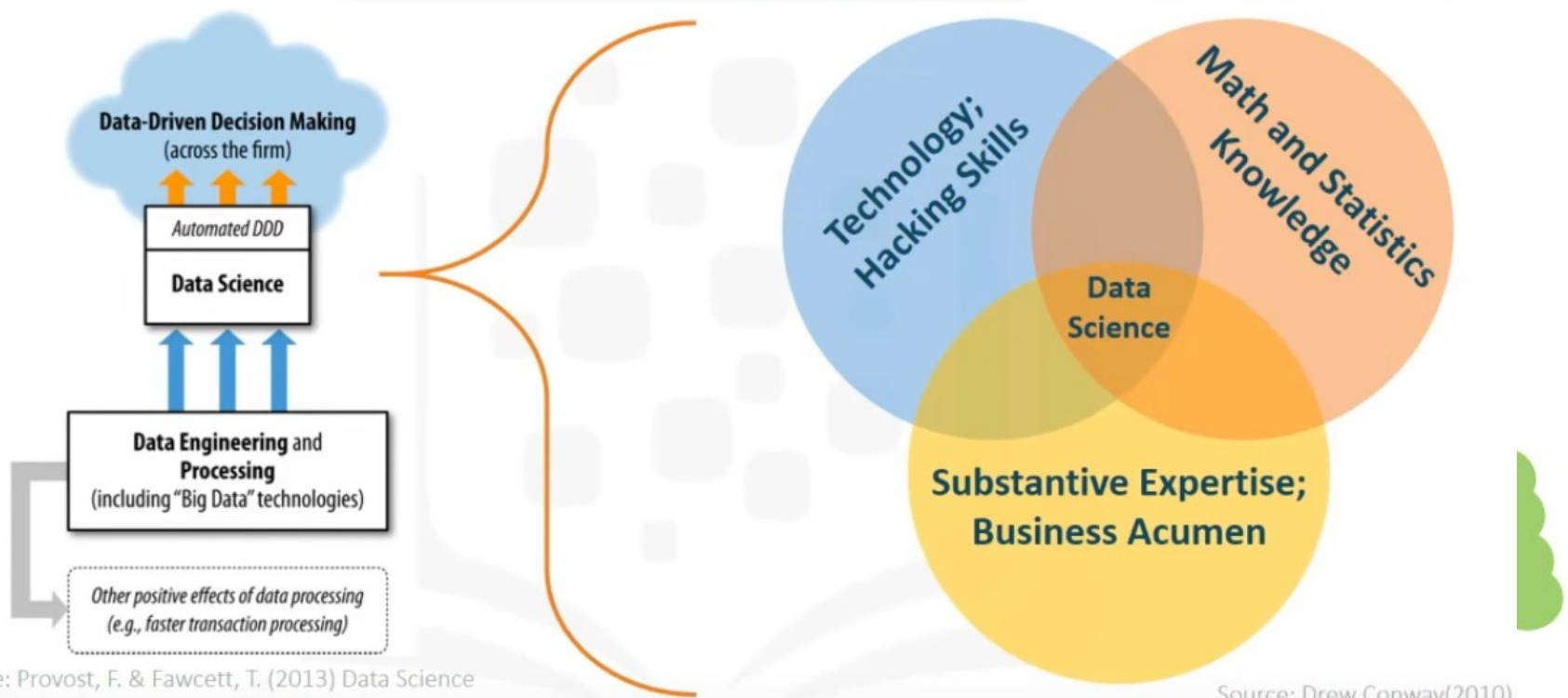
Beberapa Technical Skills untuk Data Scientists:

- Hadoop
open source platform untuk distributed massive data computing
- Apache Oozie
Scheduling system berbasis server untuk mengelola jobs pada hadoop
- Apache Hive
data warehouse software project untuk data query dan analisis
- Apache Flume
software untuk log data collection, aggregation, migration
- Apache Spark
open-source distributed general-purpose cluster-computing framework (programming interface)
- Apache Pig
Platform untuk membuat program untuk dijalankan pada hadoop
- Apache Sqoop
Command line Interface (CLI) untuk transfer data antar relational database pada hadoop



Data Science Process

“Data science is the process of cleaning, mining, and analyzing data to derive insights of value from it”

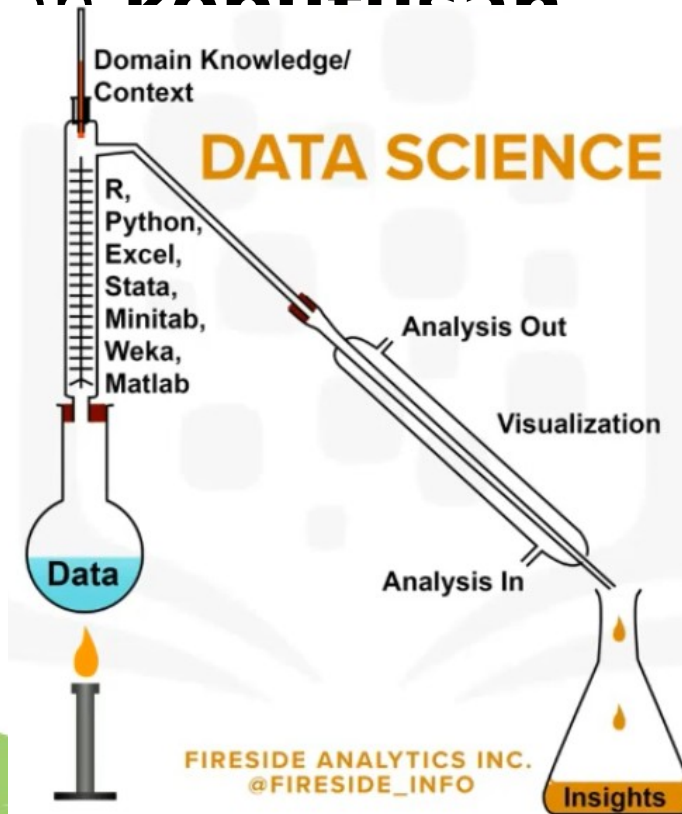


Source: Provost, F. & Fawcett, T. (2013) Data Science for Business. Sebastopol, CA: O'Reilly Media

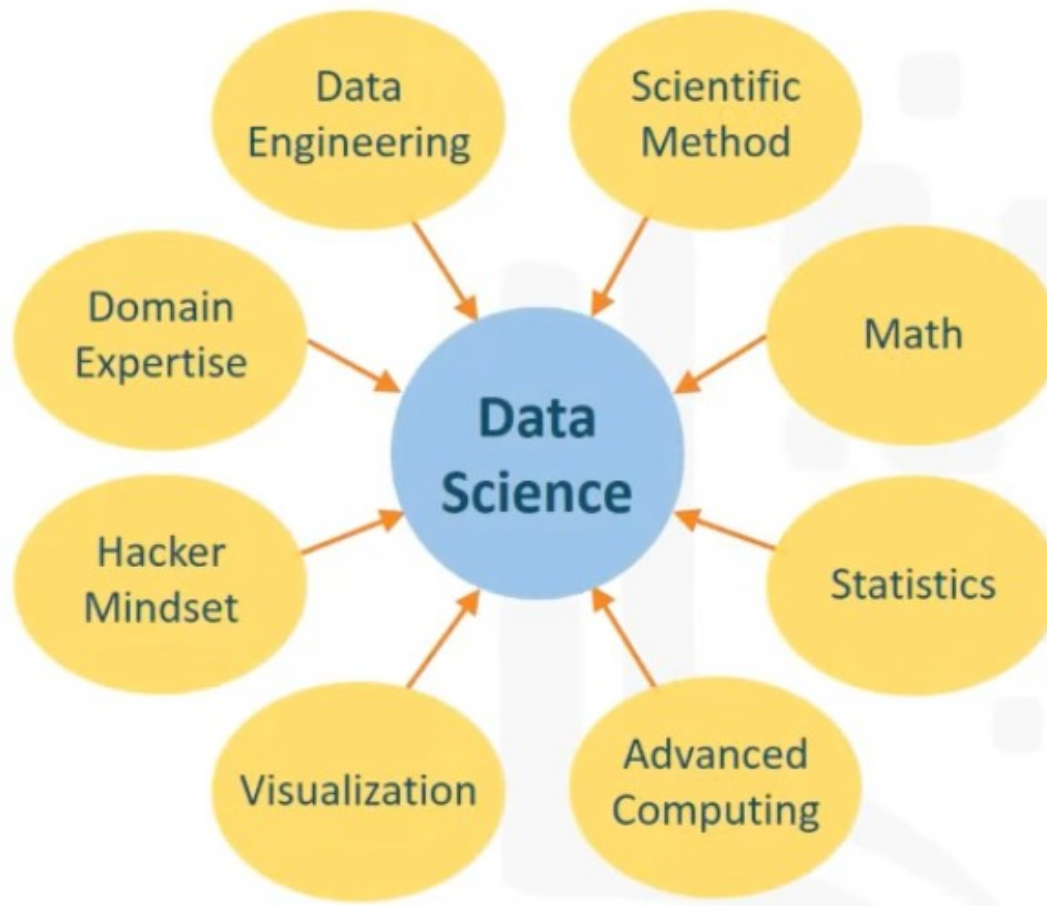
Source: Drew Conway(2010)

Data Science Process

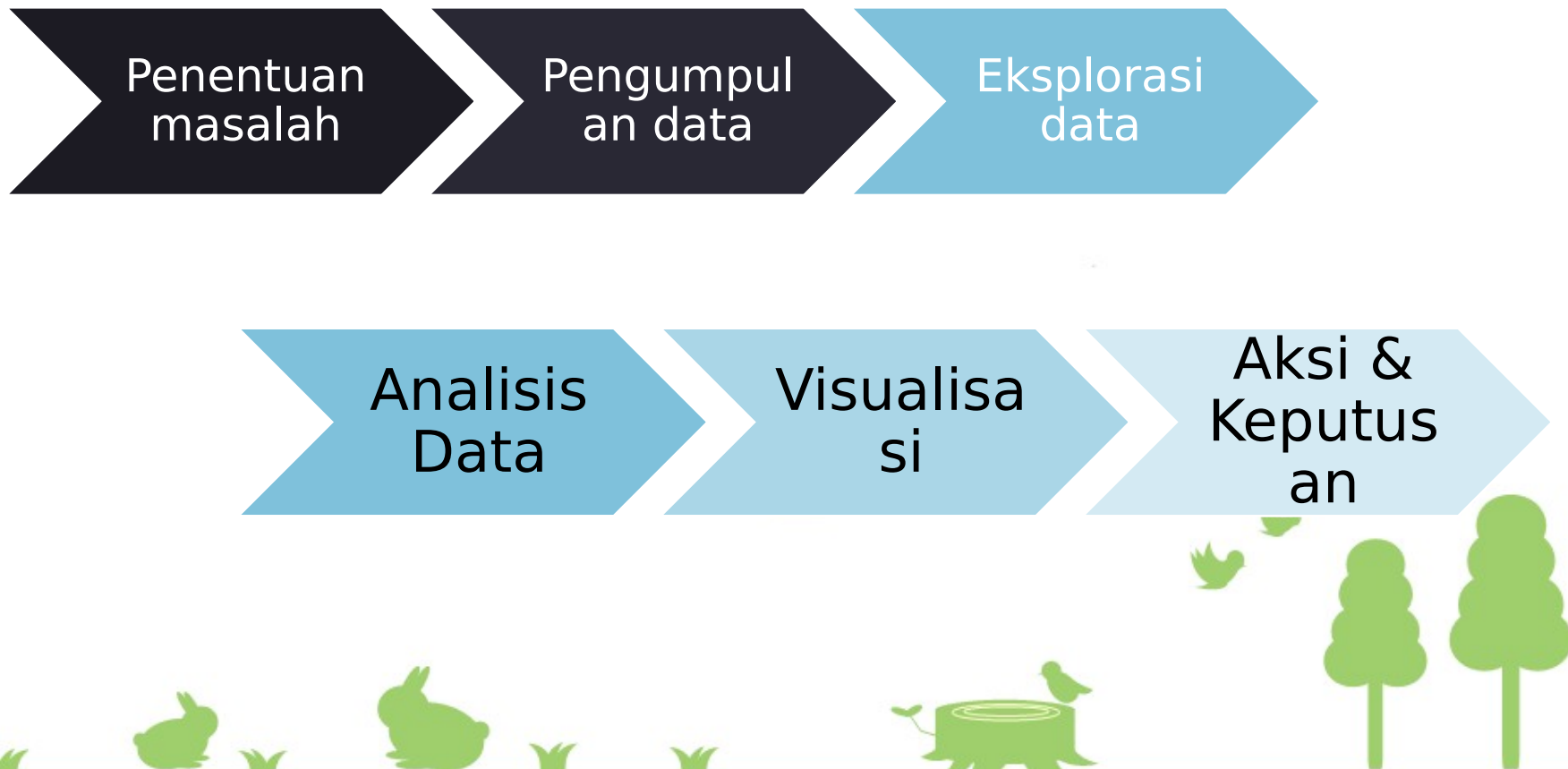
- Menyarikan **insight** dari **data**, untuk pengambilan keputusan



Data Science Skills



Data Science Process



Data Science Process (cont'd)

1. Penentuan Masalah

- Apa masalah bisnis yang ada?
- Apa tujuan dari proyeknya?
- Apa yang akan dilakukan jika semua data sudah didapatkan?

2. Pengumpulan Data

- Data mana yang relevan?
- Apakah ada masalah privasi?

3. Eksplorasi Data

- Plot data
- Apakah ada pola tertentu dari data tersebut?



Data Science Process (cont'd)

4. Analisis Data

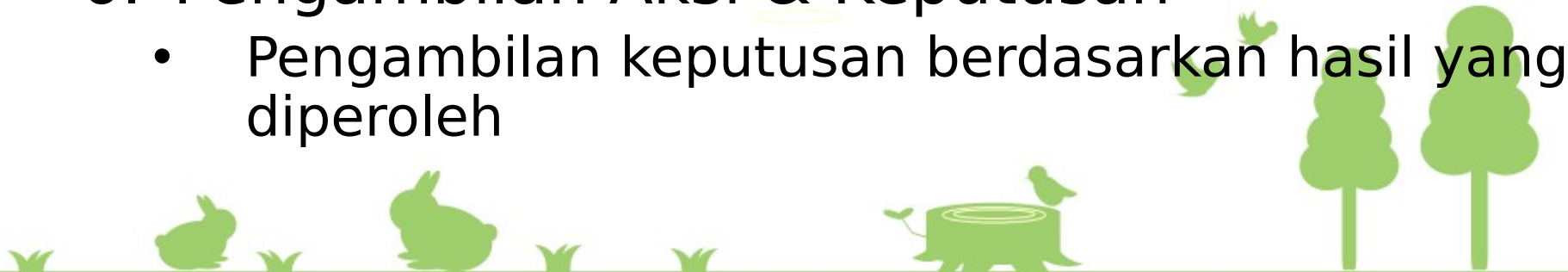
- Membuat model
- Mencocokkan dan memvalidasi model

5. Visualisasi

- Apakah hasilnya masuk akal?
- Ceritakan hasil yang diperoleh

6. Pengambilan Aksi & Keputusan

- Pengambilan keputusan berdasarkan hasil yang diperoleh



Use cases of big data

Nilai yang dapat ditingkatkan dari penerapan big data:

1. Big Data Exploration

Pencarian, visualisasi, dan pemahaman big data untuk meningkatkan pengetahuan tentang bisnis

2. Enhanced 360° View of the Customer

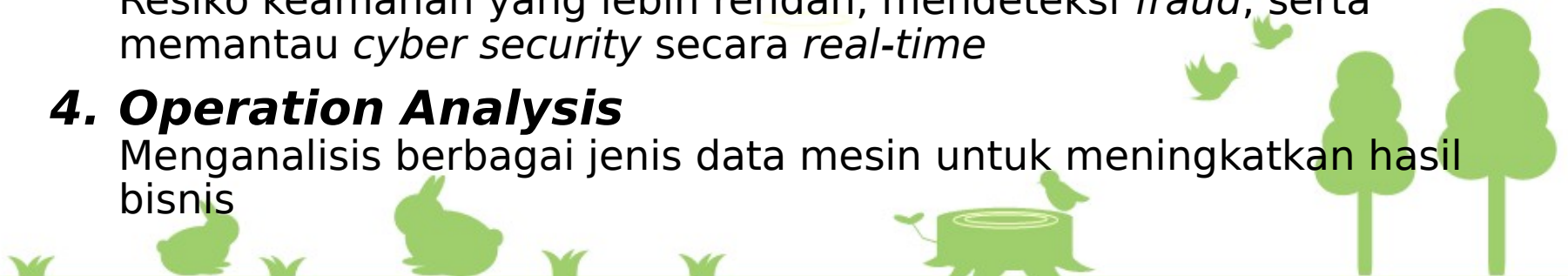
Mendapatkan suatu pandangan yang seragam tentang pelanggan, dan menggabungkan sumber daya internal dan eksternal

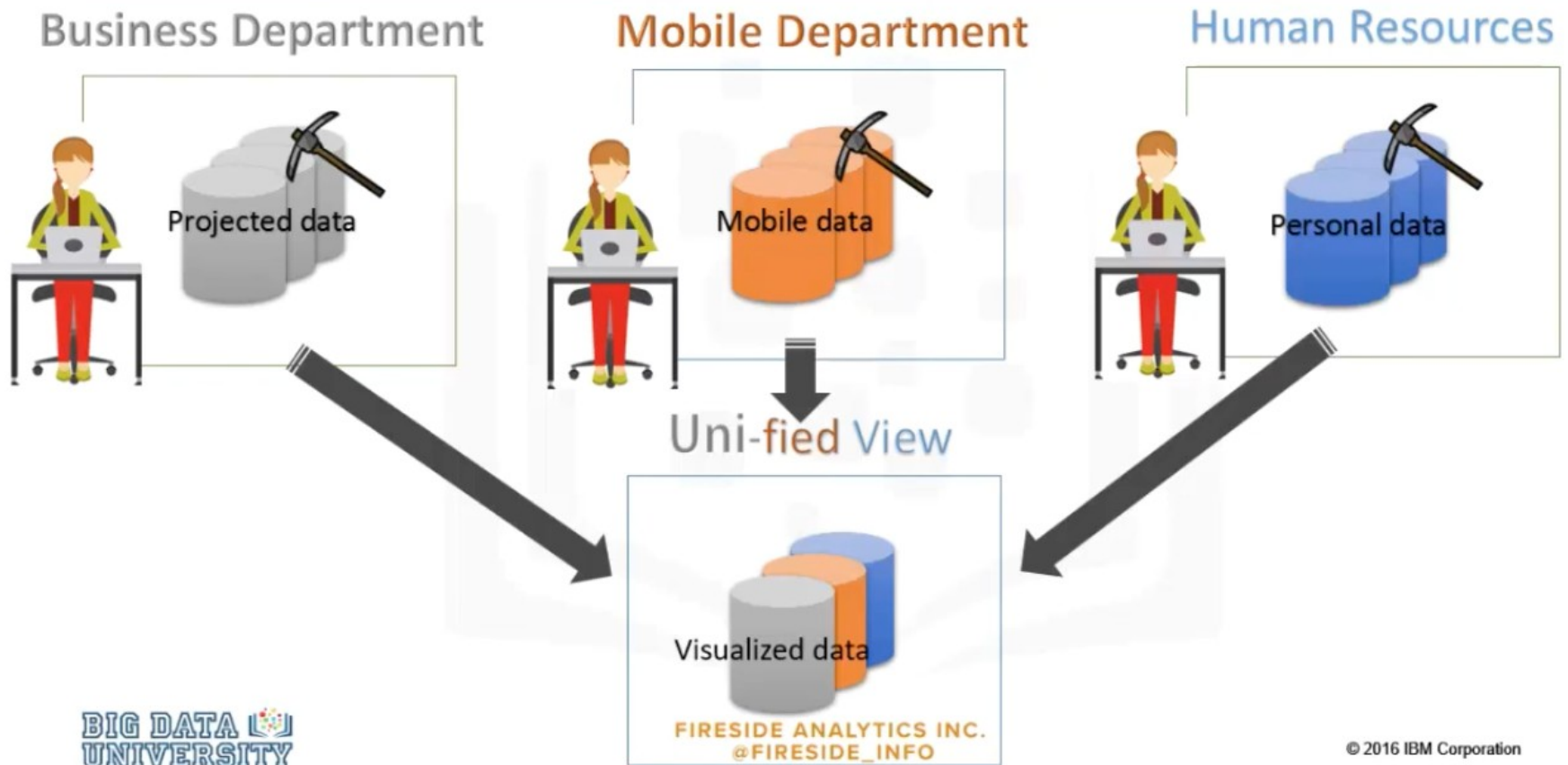
3. Security/Intelligence Extension

Resiko keamanan yang lebih rendah, mendeteksi *fraud*, serta memantau *cyber security* secara *real-time*

4. Operation Analysis

Menganalisis berbagai jenis data mesin untuk meningkatkan hasil bisnis





Dengan **big data exploration**, kita dapat mengeksplorasi big data untuk mencari, memvisualisasi, dan memahami semua data untuk membantu proses pengambilan keputusan.

Dengan membuat sebuah *unified view* dari informasi pada seluruh sumber data, baik di dalam, ataupun di luar organisasi, kita dapat mendapatkan *value* yang lebih dan juga pandangan baru.

Mengurangi Kemacetan Lalu Lintas

- *Real-time smarter traffic system* dapat memprediksa dan memperbaiki aliran lalu lintas
- Menganalisis data secara *stream real-time* yang dikumpulkan dari kamera-kamera pada titik masuk dan keluar kota, data GPS dari taksi dan truk, serta informasi cuaca



Mengurangi Kemacetan Lalu Lintas (cont'd)

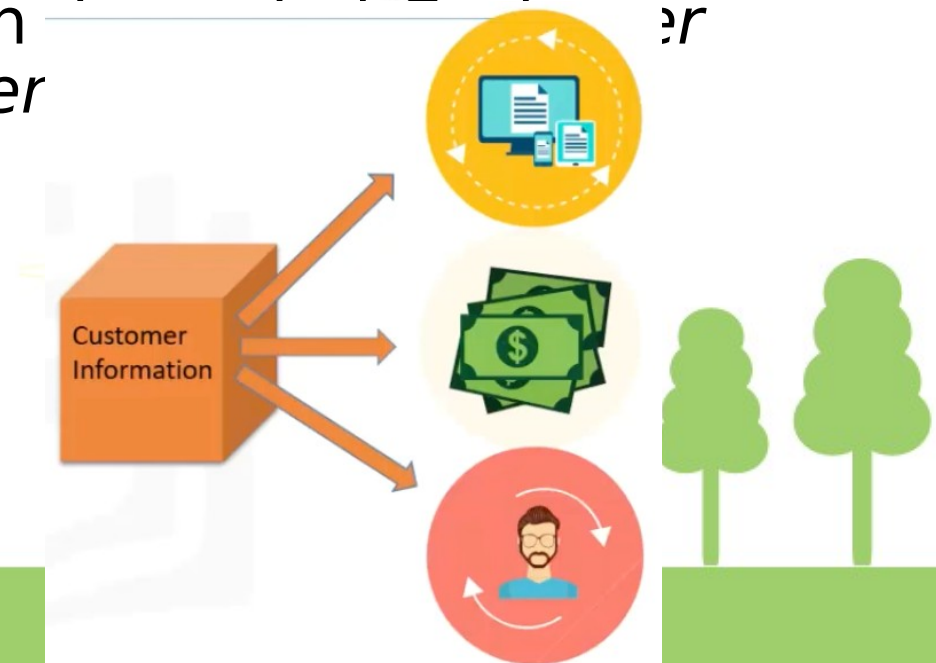
Keuntungannya:

- Dapat menganalisis dan memprediksi lalu lintas secara lebih cepat dan lebih akurat
- Memberikan pandangan baru pada mekanisme yang mempengaruhi sebuah sistem lalu lintas yang kompleks
- Menghasilkan lalu lintas yang lebih cerdas, efisien, serta ramah lingkungan
- Memberikan data informasi untuk *autonomous car* dapat mempelajari *behavior* pengemudi mobil biasa pada kondisi lalu lintas yang berbeda-beda



Enhanced 360° View of the Customer

- Pada era digital seperti ini, organisasi/perusahaan harus mempertimbangkan semua informasi penting yang ada tentang konsumen/pelanggan untuk meningkatkan *revenue* serta hubungan *loyalty* jangka panjang dengan konsumen
- Hal tersebut merupakan *Relationship Managemer*



Enhanced 360° View of the Customer (cont'd)

Retail : Meningkatkan Hubungan Konsumen

- Mempersiapkan untuk peningkatan jumlah konsumen pada *rush hour/rush time*
- Mempelajari dan menganalisis preferensi konsumen
- Merencanakan kebiasaan konsumtif dari konsumen (diskon, kupon, dll)

Hasilnya:

- Peningkatan *revenue*
- Peningkatan efisiensi
- Membangun loyalitas konsumen




Security/Intelligence Extension

Meningkatnya jumlah kejahatan siber (*cyber crime*), terorisme siber (*cyber terrorism*), spionase, dll menjadi tantangan bagi organisasi/perusahaan yang harus diatasi dan juga dicegah.

Hal tersebut dapat dilakukan dengan memproses dan menganalisis tipe data baru seperti media social, email, dan menganalisis rekaman video, menganalisis *data in motion/at rest*.

Dengan demikian organisasi dapat mendapatkan asosiasi baru, atau menemukan pola dan fakta baru untuk meningkatkan *intelligence*, *security* dan *law enforcement*

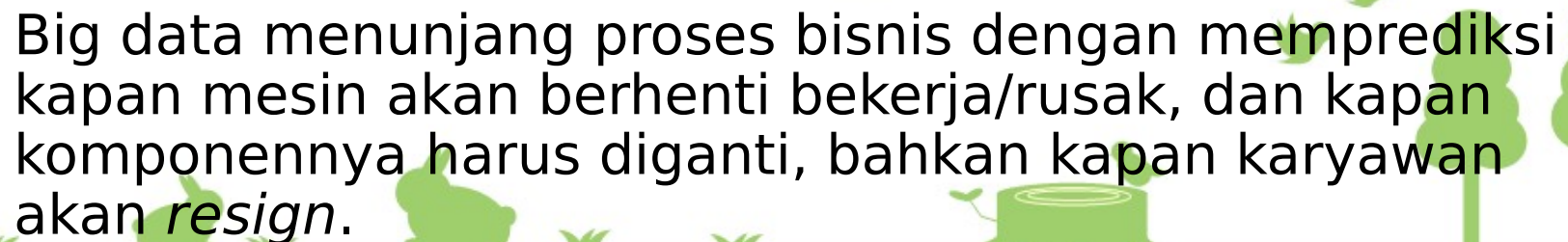


Operation Analysis

Berfokus pada analisis data mesin, yang dapat mencakup sinyal, sensor, dan log data dari perangkat GPS, dimana data-data tersebut meningkat secara eksponensial, dalam jumlah yang besar, dan format yang berbeda-beda

Dengan menggunakan big data untuk menganalisis operasi, organisasi dapat melihat secara *real-time* : operasi, *customer experience*, transaksi, dan *customer behavior*

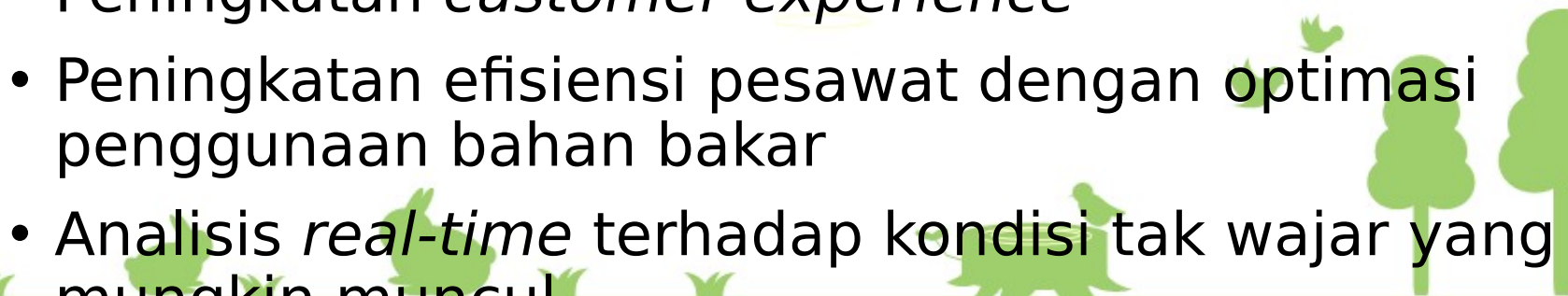
Big data menunjang proses bisnis dengan memprediksi kapan mesin akan berhenti bekerja/rusak, dan kapan komponennya harus diganti, bahkan kapan karyawan akan *resign*.

A decorative green landscape at the bottom of the slide. It features several stylized green trees of different shapes and sizes, a small green bird in flight, and a green stump with a small plant growing from it. The entire scene is set against a solid green background.

Meningkatkan Keamanan Penerbangan (*Aviation*)

- Mengumpulkan dan menganalisis data dalam jumlah besar dari turbin pesawat, sensor, dan GPS
- Memvisualisasikan data dengan analisis big data kompleks

Hasilnya:

- Pantauan *real-time* terhadap operasi pesawat
 - Peningkatan *customer experience*
 - Peningkatan efisiensi pesawat dengan optimasi penggunaan bahan bakar
 - Analisis *real-time* terhadap kondisi tak wajar yang mungkin muncul
- 
- A decorative green landscape at the bottom of the slide, featuring stylized green trees of varying heights and a small green bird flying in the sky.

Use cases of big data

Pengaruh Big Data Terhadap Proses Bisnis

Objective

“We want to know what every product in the world is. We want to know who every person in the world is. And we want to have the ability to connect them together in a transaction.”

– Neil Ashe, CEO of Global E-commerce at
Walmart



Use cases of big data (cont'd)

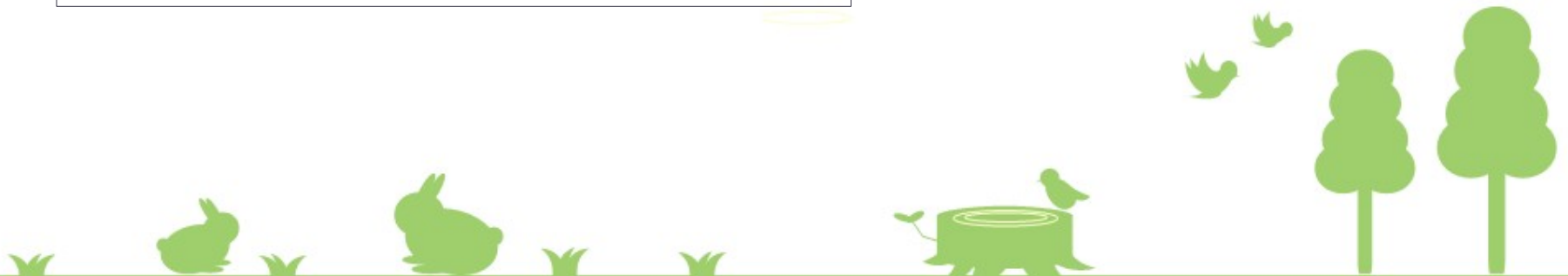
Pengaruh Big Data Terhadap Proses Bisnis



Dalam 1 jam, Walmart mengumpulkan data sebesar 2.5 petabytes dari 1 juta konsumennya



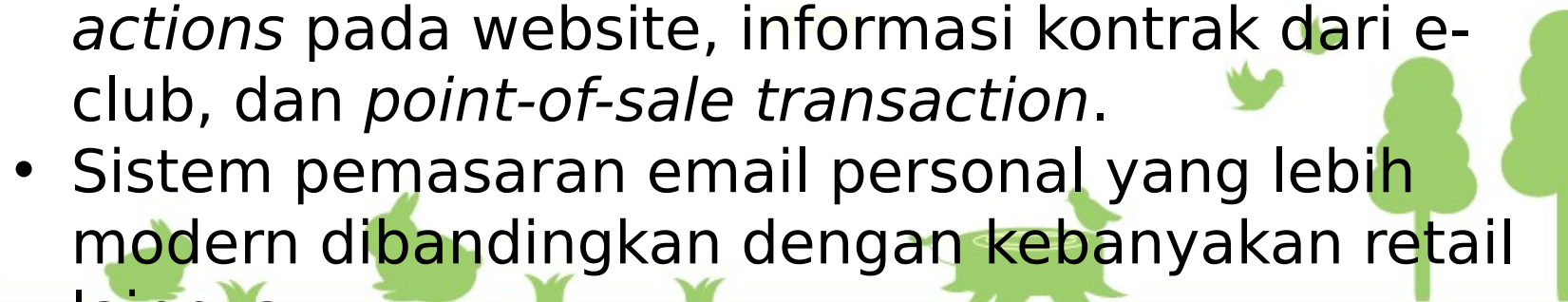
BIG DATA



Use cases of big data (cont'd)

Pengaruh Big Data Terhadap Proses Bisnis

Aplikasi Big Data:

- “menambang” (*mining*) data penjualan membantu Walmart untuk menemukan pola yang dapat digunakan untuk memberikan rekomendasi produk dan promo ke konsumen
 - Menghubungkan *in-store* dan *online customer behavior* dengan menggunakan data dari *click actions* pada website, informasi kontrak dari e-club, dan *point-of-sale transaction*.
 - Sistem pemasaran email personal yang lebih modern dibandingkan dengan kebanyakan retail lainnya
- 

Diskusi (1 meja 1 topik)

- Tentukan topik berkaitan dengan big data.
- Tentukan dan jelaskan volume, velocity, variety dan veracity dari topik tersebut.
- Apa keputusan/kesimpulan penting yang bisa diambil dari big data tersebut?
- Bagaimana cara memudahkan perolelahan keputusan/kesimpulan tersebut? (visualisasi, clustering, klasifikasi, regresi dll)

