PERTEMUAN 1

Pengenalan Data Science

Why, Where, What, How, Who

Datascience telah ada sejak…

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

W.E. Deming

1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"

1989: "Business Intelligence"

Howard Dresner

1997: "Machine Learning"

2010:"The Data Deluge"

1996: Google

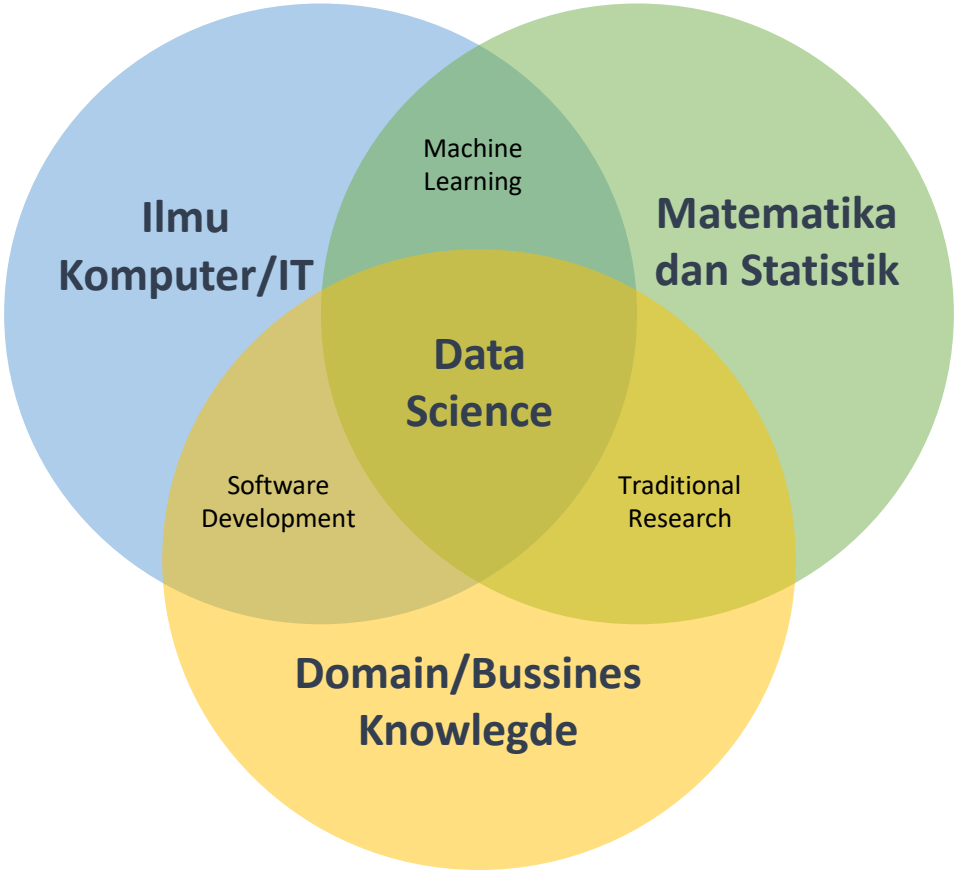
Google

2009:"The Unreasonable Effectiveness of Data"

Data Science

Data science adalah ilmu yang menggabungkan matematika, statisika dengan ilmu komputer dengan tujuan analisa data (data analysis) dari suatu himpunan data baik skala kecil (sampel) maupun besar (populasi) dengan mengaplikasikan algoritma tertentu untuk tujuan menggali data (data mining) dan mendapatkan pola data serta dapat melakukan prediksi data (prediction) dengan cukup akurat yang dapat membantu dalam pengambilan keputusan dan dapat digunakan untuk membuat sistem yang cerdas (AI) yang dapat terus belajar dengan sendirinya (machine learning).

Irisan Bidang Ilmu dalam data Science



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Irisan Bidang Ilmu dalam data Science

Matematika dan Statistika.

Pemrograman (R, Python, dan lainnya)

Database dan Query (SQL dan lainnya) dan pengolahan data.

Analisa data dan visualisasi data.

Pemahaman masalah terkait bisnis atau suatu bidang lainnya

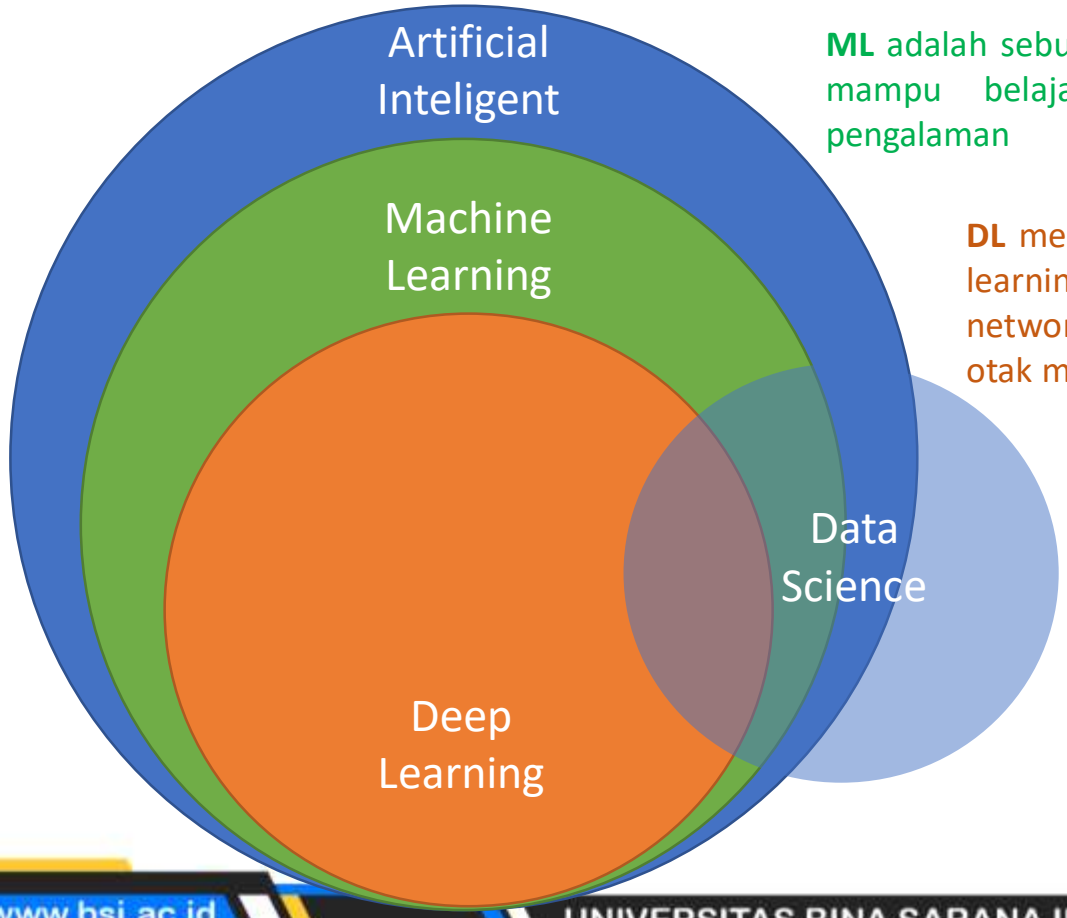
Proses Yang Terlibat dalam Data Science

Data Mining adalah proses pengambilan informasi dari pola data dari himpunan data yang sebelumnya tidak diketahui, kadang disebut juga Data Discovery. Data Mining fokus pada mengekstrak pola menggunakan metode statistik untuk dianalisa dan dapat juga melakukan prediksi.

Machine learning adalah bidang yang merupakan bagian dari Artificial Intelligence (AI) yang digunakan agar sistem komputer secara otomatis dapat belajar dengan sendirinya tanpa diberi instruksi pemrograman dan dapat meningkatkan prediksi yang akurat dan pengunaannya biasanya sifatnya realtime.

Hubungan AI, Machine Learning Deep Learning dan DataScience

AI mensimulasikan Kecerdasan manusia manusia kedalam mesin, khususnya kedalam komputer



ML adalah sebuah algortima dalam computer yang mampu belajar secara otomatis berdasakan pengalaman

DL merupakan salah satu bagian dari dari machine learning, yang berbasis pada artificial neural network, dengan cara kerja mengikuti pola saraf otak manusia

DS adalah sebuah bidang lintas disiplin ilmu yang didalamnya metodologi scientific, process, algoritma dan system yang mengekstraksi pengetahuan untuk mencari pengetahuan baru dari sebuah data baik terstruktur maupun tidak terstruktur. DS berhubungan dengan datamining, deep learning dan big data

Source : @ingliguori

Kenapa datascience begitu menarik

Lebih efektif untuk analisis data

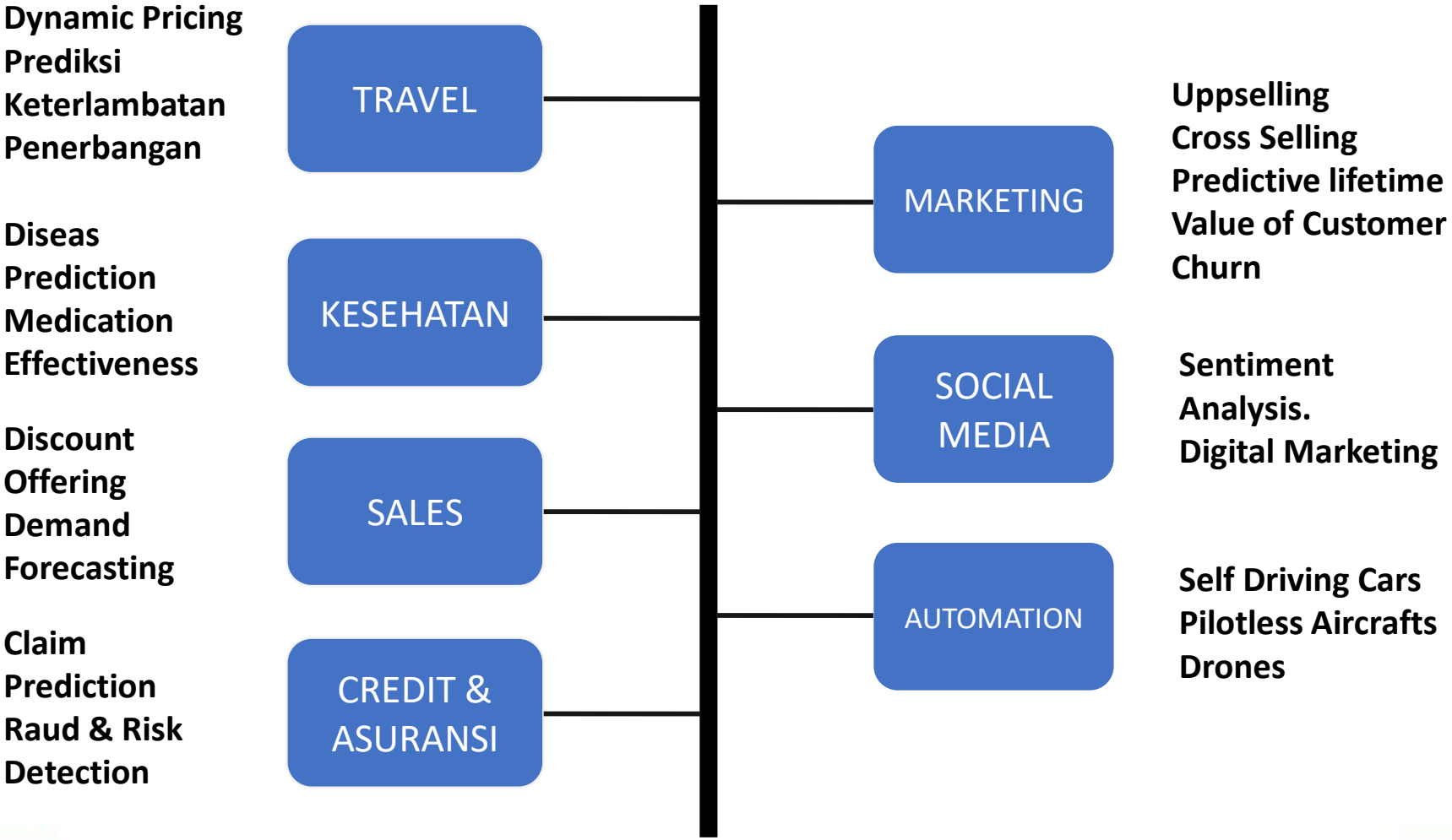
Contoh Implementasi
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

Prediction model is built on
Various data sources,
types and analysis.

Contoh penerapan datascience



Source : https://www.analyticssteps.com/

Darimana data itu berasal

Saat ini data dihasilkan media digital, dalam 1 hari total data yang dihasilkan diperkirakan sebesar **2.500.000.000.000.000.000** byte atu **2,5** *quintillion* byte.

Seditadaknya, dibutuhkan 2.500 buah hardisk berkapasitas 1 Terra Byte untuk menampung data tersebut **DALAM 1 Hari**

Darimana itu berasal

Google memproses 24 Petabyte data setiap harinya

3,5 miliar pencarian setiap harinya

selain search engine goole terdapat 5,5 miliar percarian perhari

Perangkat seluler mengirimkan 1,3 Exabyte setiap harinya (1,3 Exabyte =1.300.000 TB)

Darimana itu berasal Lanjut…

156 Juta email terkirim setiap harinya, 103.447.520 e-mail spam setiap menit

Netflix di tonton sebanyak 64.444 jam per menit

Spotify 18.720 lagu baru teluppload setiap hari

Darimana itu berasal Lanjut…

Twitter 456.000 tweets setiap harinya

Instagram 95 juta post foto dan video 100juta orang menggunakan fitur stories setiap hari

Youtube 4.146.000 tontonan setiap harinya

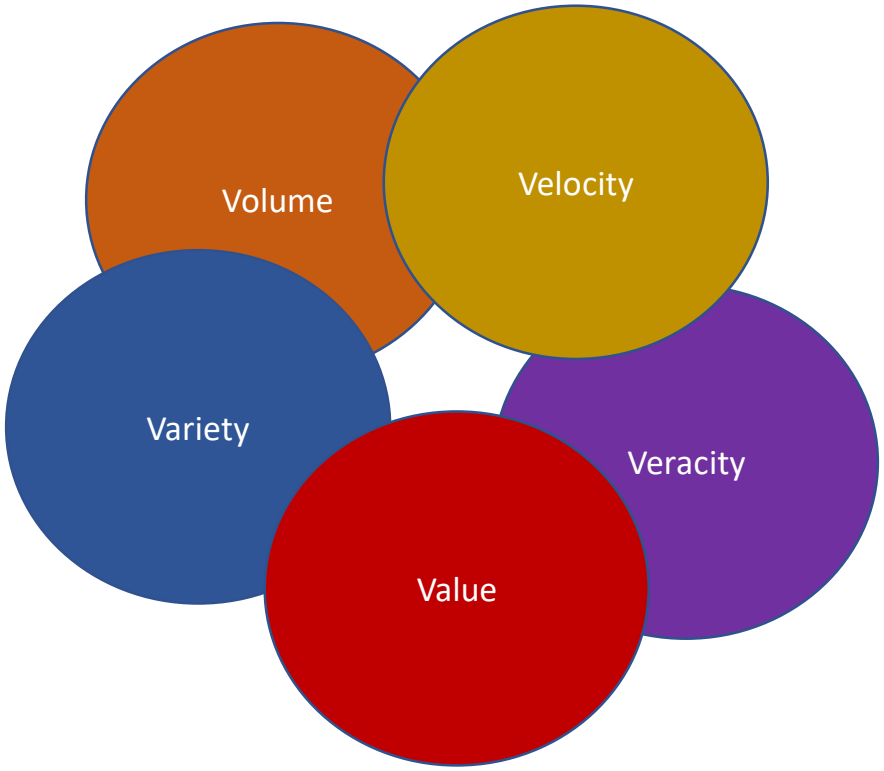
Facebook 510.000 komentar dan 293.000 status per menit

Hubungan Big Data & Data Science

Big Data : lebih basis data/source data

Data science : proses pengolahan data yang terdapat dalam basis data tersebut

Five-V BIG DATA



Sumber : https://www.geeksforgeeks.org/5-vs-of-big-data/

Istilah dalam Big Data

Volume

Mengacu pada sejumlah big data yang dihasilkan setiap detik nya. Artinya sekumpulan data dalam jumlah dan *volume*yang sangat besar dan kadang tidak terstruktur. Contohnya feed Twitter, feed Istagram, data teks chat dan status Whatsapp, alur klik *user* dari halaman *web*. Arus data-data tersebut bisa berukuran hingga ribuan Terrabyte (TB) per detiknya.

Velocity

Data dapat diakses dengan kecepatan yang sangat cepat sehingga dapat langsung digunakan pada detik itu juga (lebih real time). Salah satu buktinya antara lain, adanya sistem operasi *online*berbasis Microsoft Silverlight, aplikasi perkantoran (*office*) berbasis *web* seperti Office365, *cloud storage* seperti Dropbox dan GDrive

Istilah dalam Big Data

Variety

Data bisa disebut sebagai *big data*jika memiliki karakteristik yang bermacam-macam dan tidak homogen, tetapi memiliki banyak sekali variabel dan sangat beragam meliputi berbagai jenis data baik data yang telah terstruktur dalam suatu database maupun data yang tidak terorganisir dalam suatu *database*. Analisis terhadap data yang tidak terstruktur akan memerlukan algoritma yang agak berbeda, seperti data teks, gambar, suara, dan video. Untuk data-data semacam itu akan memerlukan waktu lebih untuk memprosesnya, karena bisa jadi di dalam data yang tidak terstruktur tersebut masih ada data lain atau data baru yang bisa digali.

Istilah dalam Big data

Veracity

Big data memiliki kerentanan dari sisi keakuratan dan kevaliditasan sehingga memerlukan kedalaman untuk menganalisis *big data* agar bisa menghasilkan keputusan yang tepat. Karakter *veracity* mengarah kepada seberapa akurat dan dapat dipercaya suatu data.

Value

Value berarti *big data* memiliki nilai yang sangat tinggi apabila diolah dengan cara yang tepat guna atau dapat juga dikatakan seberapa bernilainya atau bermaknanya suatu data. Contohnya, biodata karyawan suatu perusahaan penjualan bahan baku makanan tidak akan bernilai untuk kepentingan analisis prediksi penjualan bahan baku ke customer. Data tersebut mungkin tidak penting dan tidak bernilai untuk satu hal, namun bisa sangat penting dan sangat bernilai untuk hal lain. Data yang tidak memiliki nilai di bagian mana pun tidak akan terfilter di sistem aplikasi analisis Big data.

Thomas Connolly and Carolyn Begg. 2015.

- Dari data yang ada tersebut apakan kita dapat memanfaatkannya..?

- Maka lahirlah sebuah profesi datascientist

