Konsep Dasar *Data Science*

Ari Fadli
fadli.te.unsoed@gmail.com

Data science merupakan ilmu pengetahuan multidisiplin yang secara khusus mempelajari data terutama yang sifatnya kuantitatif. Selain itu *data science* dapat pula didefisnisikan sebagai proses penggalian data sehingga dihasikan produk data yang benar atau dengan kata lain. *Data Science* merupakan sebuah proses untuk memproduksi pengetahuan data *(data insight)*. Untuk menghasilkan produk data yang benar *data science* memiliki terdiri dari tiga fase yaitu desain data, pengumpulan data, dan analisis data. *Data Scientist* merupakan seseorang yang melakukan pengolahan data tersebut sehingga menghasilkan pengetahuan.

Pendahuluan

Data is a new currency. Kalimat tersebut akhir-akhir ini santer diperbincangkan dikaitkan dengan pergerakan transformasi digital, mengisyaratkan betapa bernilainya data bagi sebuah langkah strategis bisnis.

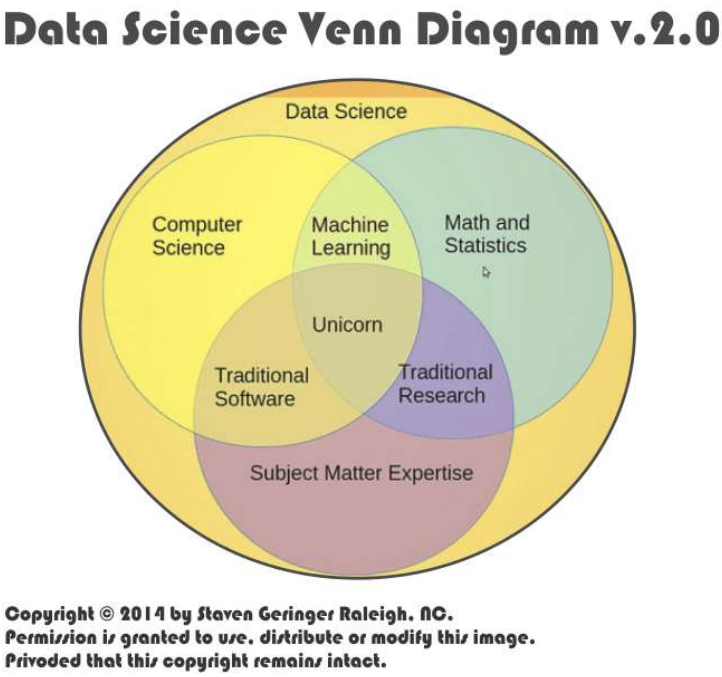
Data science dapat pula didefinisikan sebagai cabang ilmu yang memperlajari teknik ektransi data sehingga bermakna dan logis. Dalam data science ini juga terdiri dari beberapa tahapan kegiatan yaitu penambangan data dan analisis data, dengan menggunakan pengetahuan pada cabang ilmu matematika, statistik, dan teknologi informasi, pemrograman komputer, pengenalan pola, pembelajaran mesin.

Definisi Lain adalah "*Data science starts with data, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables. Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset. The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI). We can further define data science by investigating some of its key features and motivations* [1]".

Sementara itu *Data Scientist* didefinisikan sebagai :

"*A data scientist is simply a person who can write code (in languages like R, Python, Java, SQL, Hadoop (Pig, HQL, MR) etc.) for data (storage, querying, summarization, visualization) efficiently and quickly on hardware (local machines, on databases, on cloud, on servers) and understand enough statistics to derive insights from data so business can make decisions* [2]"

Menurut Staven Geringer Raleigh (2014), pembentuk *data science* dapat diilustrasikan dalam diagram venn berikut :



Gambar-1. Diagram Data Science
Sumber : Data Science Venn diagram. Source: Copyright © 2014 Steven Geringer Raleigh, NC [2]

Berdasarkan Gambar-1 dapat dijelaskan beberapa hal sebagai berikut :

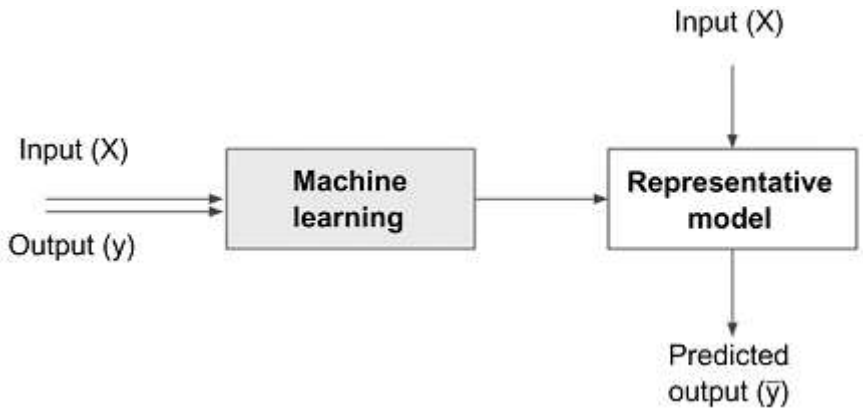
1. *Machine Learning*
   Machine Learning adalah cabang ilmu kecerdasan buatan (*Artificial Intelligence*) yang memepelajari bagaimana dapat memberikan kemampuan belajar pada sebuah mesin (komputer, mini komputer) dengan menggunakan algoritme tertentu.
2. *Traditional Software*
   Merupakan cabang ilmu yang dihasilkan dari irisan cabang ilmu komputer dengan SME (S*ubject Matter Expertise*). SME sendiri merupakan pengetahuan yang digunakan untuk mengembangkan sistem yang dapat membantu proses bisnis pada sebuah instansi. Penerapan *traditional software* ini telah digunakan hampir di seluruh instansi pemerintahan maupun swasta atau pada perusahaan, contohnya *e-learning*, *e-library*, *online banking*, *Point of Sales* (*PoS*)**.**

3. *Traditional Research*
   Traditional research merupakan cabang ilmu yang diperoleh dari irisan pada ilmu matematika dan statistika dengan SME (S*ubject Matter Expertise*). *Traditional research* telah dilakukan diberbagai baik di perusahaan, instansi serta universitas.

Model *Data Science*

Pada Gambar-1 tampak bahwa *data science* akan menemukan pola yang sebelumnya tidak diketahui dalam data dengan menggunakan pembelajaran mesin untuk menghasilkan model representatif. Dalam *Representative Model* (model representatif) akan memberikan gambaran hubungan antar variabel yang ada dalam dalam *dataset* dengan kata lain hal ini menjelaskan bagaimana satu atau lebih variabel dalam data terkait variabel lain.

Dengan kata lain *data science* juga merupakan proses membangun model representatif yang sesuai dengan data pengamatan. Model ini melayani dua tujuan: di satu sisi, ia memprediksi output berdasarkan pada data input baru serta model dapat digunakan untuk memahami hubungan antara variabel output dan semua variabel input.



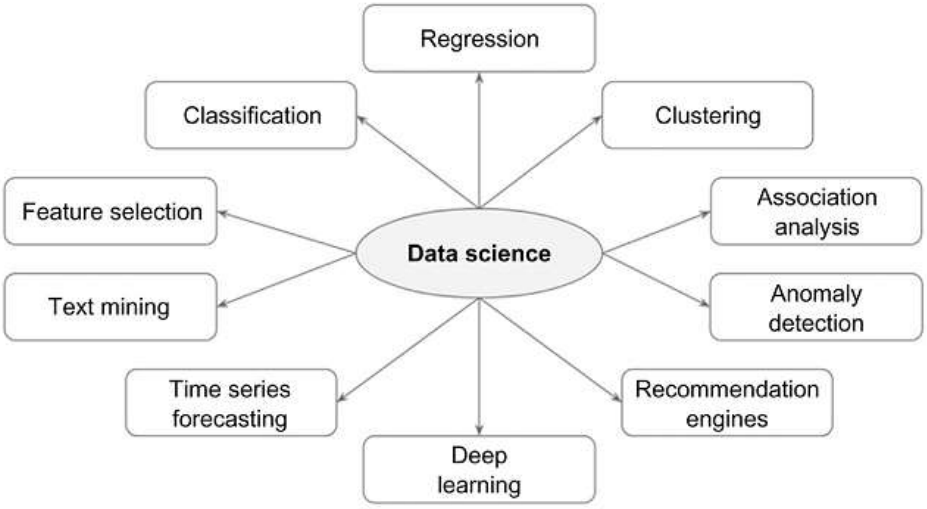
Gambar-2 Model Data Science [1]

Untuk membangun *data science* dapat digunakan beberapa sumber data berikut :

1. *Kaggle* merupakan salah satu situs web untuk Data Science dan Machine Learning yang menyediakan sekitar 6000 *dataset* dalam format CSV.
2. *UCI Machine Learning Repository* merupakan pusat *dataset* yang menyediakan *dataset* yang dapat diunduh secara gratis. Terdapat sekitar 400 *dataset*.
3. *data.gov* merupakan adalah pusat data terbuka milik Pemerintah AS yang terdiri terdiri dari berbagai kategori beberapa diantaranya yaitu Pertanian, Konsumen, Ekosistem, Pendidikan, Energi, Keuangan dan Sains.

Data Science Tasks

Klasifikasi beberapa *task* dalam data science seperti tampak pada Gambar 3



Gambar-3 *Task Data Science* [1]

Berikut adalah deskripsi singkat dari beberapa *task data science* pada Gambar-3

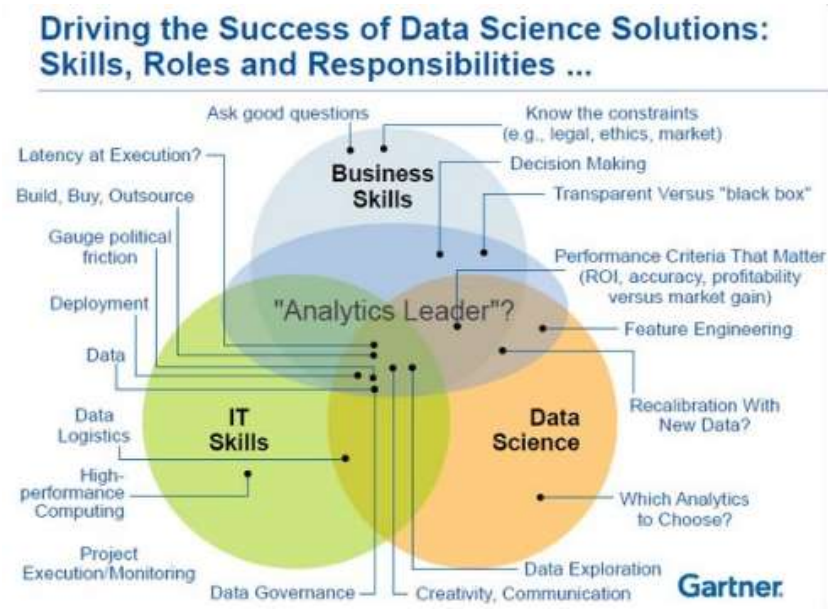
1. Pada *task* klasifikasi dan regresi digunakan untuk memprediksi variabel target berdasarkan pada variabel input. Prediksi yang dibuat ini didasarkan pada model umum yang dibangun dari dataset yang diketahui sebelumnya.
2. *Deep learning* merupakan *artificial neural network* yang bersifat *sophisticated*, penerapan deep learning ini telah banyak diterpkan untuk penyelesaian masalah klasifikasi dan regresi
3. *Clustering* merupakan proses mengindentifikasi pengelompokan data yang dilakukan secara alami berdasarkan pada dataset yang tersedia. Pengelompokan ini didasarkan pada algoritma pembelajaran yang bersifat *unsupervised learning*
4. *Recommendation engines* merupakan mesin yang dibuat agar memiliki kemampuan memberikan rekomendasi kepada pengguna berdasarkan pada preferensi pengguna.
5. *Anomaly or outlier detection* merupakan kemampuan melakukan identifikasi pada titik-titik data diluar dataset yang secara signifikan memiliki sifat yang berbeda dengan dataset.
6. *Time series forecasting* merupakan sebuah proses memprediksi sebuah nilai tertentu berdasarkan pada histori data masa lalu yang kemungkinan akan memberikan sebuah trend / pola tertentu yang sifatnya didasarkan pada waktu (tahunan, bulanan, mingguan atau harian)
7. *Text mining* dikenal juga sebagai dengan nama analisis teks yang merupakan proses mengubah data teks yang tidak terstruktur menjadi informasi yang bermakna dan dapat ditindaklanjuti.

8. *Feature selection* merupakah sebuah proses untuk menyeleksi attribut dalam sebuah dataset, sehingga diperoleh attribut-atribut yang sifatnya penting dan dapat memberikan ciri dari objek tertentu.

Kemampuan Dasar Seorang *Data Scientist*

Berdasarkan definisi *data scientist* yang telah disampaikan sebelumnya, dapat disimpulkan bahwa seoerang *data scientist* dituntut memiliki kreativitas dan kecerdikan dalam menggunakan kemampuan teknisnya untuk membangun dan menemukan solusi yang cerdas untuk setiap permasalahan. Beberapa skil yang wajib dimiliki oleh seorang *data science* tampak pada Gambar 4 dan Gambar 5.

Gambar-4 Skill Dasar *Data Science* Sumber : https://www.simplilearn.com



Gambar 5. Skill, Roles, and Responsibilities *Data Science* [3]

Tools Available to Data Scientists

Beberapa Tools yang tersedia untuk para *Data Scientist* :

1. *Data storage* : MySQL, Oracle, SQL Server, HBase, MongoDB, and Redis
2. *Data querying* : SQL, Python, Java, and R
3. *Data analysis* : SAS, R, and Python
4. *Data visualization* : JavaScript, R, and Python
5. *Data mining* : Clojure, R, and Python
6. *Cloud* : Amazon AWS, Microsoft Azure, and Google Cloud
7. *Hadoop Big Data* : Spark, HDFS MapReduce (Java), Pig, Hive, and Sqoop

Dalam pemrograman *data science* terdapat beberapa paket *software* sebagai berikut :

1. Pandas merupakan sebuah *software library* Phyton yang digunakan untuk melakukan manipualasi dan analisis data.
2. NumPy merupakan *Add-On* dari Python yang mendukung untuk operasi multidimensional arrays and matrices dalam skala besar.
3. SciPy merupakan library dasar untuk *scientific computing*.
4. Matplotlib merupakan tools yang digunakan untuk membuat visualisasi data 2D.
5. Seaborn merupakan varian pemrograman Python untuk melakukan visulisasi data dengan menggunakan library dari matplotlib.

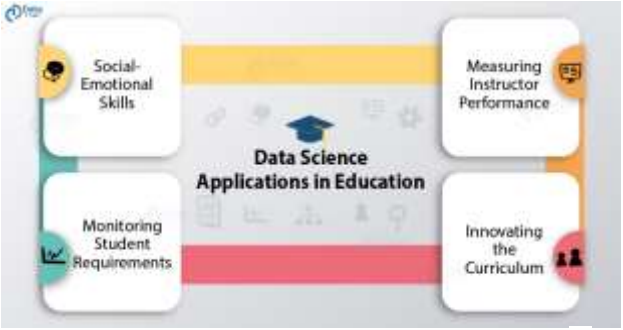
Aplikasi *Data Science* yang Populer

Gambar-6 Penerapan *Data Science* di Keuangan [4]

Gambar-7 Penerapan *Data Science* di Bisnis Ritel [4]

Gambar-8 Penerapan *Data Science* di Bisnis [4]

Gambar-9 Penerapan *Data Science* di Kesehatan [4]



Gambar-10 Penerapan *Data Science* di Pendidikan [4]

Gambar-11 Penerapan *Data Science* di Perbankan [4]

Gambar-12 Penerapan *Data Science* di Indusri Film [4]

Penutup

Data Science merupakan suatu proses yang dilakukan untuk menghasilkan pengetahuan data (*data insight*). Pengetahuan data tersebut merupakan sebuah simpulan yang dapat memberikan rekomendasi atau prediksi untuk kebutuhan tertentu. *Data scientist* adalah seseorang yang harus mampu melakukan *mining* data dengan mengekstraknya hingga menemukan data yang akurat yang dapat digunakan oleh para pemangku kebijakan. Sehingga dengan demikian seorang *data scientist* harus mampu mengidentifikasi permasalahan, mengumpulkan data dari berbagai sumber yang berbeda, mengatur informasi dan menerjemahkan hasil menjadi solusi.

Referensi

[1]. Vijay Kotu Bala Deshpande, Data Science Concepts and Practice Second Edition
[2]. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119126805.ch1[online], diakses pada 25 Mei 2020
[3]. Vincent Granville, https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning [online], diakses pada 25 Mei 2020
[4]. Himani Bansal,https://becominghuman.ai/top-data-science-applications-how-data-science-bought-change-to-the-world-e215c3b25d9d?gi=920fc92643a1 [online], diakses pada 25 Mei 2020