Bab 1
Data Science dan Data Scientist

Oleh:

Veronica S. Moertini

1.1. Data Abad 21

Bagi mayoritas orang, terlebih lagi yang belum berkecimpung di dunia kerja, barangkali data dianggap tidak penting. Data bisa jadi dianggap berkonotasi dengan "tumpukan" angka-angka yang membosankan dan "meaningless". Data dianggap menjadi urusan perusahaan atau pemerintah, sehingga merupakan hal yang "jauh" dari kehidupan sehari-hari. Maka, meskipun "data science" atau ilmu data dan profesi data scientist sudah "terlahir" sejak beberapa tahun yang lalu, dapatlah dipahami bahwa masih banyak orang yang bertanya-tanya tentang apa itu data science, juga apa yang dikerjakan data scientist.

Sejatinya dalam kehidupan sehari-hari kita sudah memanfaatkan atau bahkan "menikmati" hasil data science atau buah karya dari para data scientist. Misalnya:

- Saat kita browsing di toko online, lalu kita klik salah satu item produk, di bawah browser akan diberikan produk-produk lain yang dibeli bersamaan atau yang mungkin kita sukai juga. Sama halnya ketika kita browsing di penyedia streaming lagu dan video. Kita juga akan disuguhi dengan rekomendasi item-item lain untuk didengar atau dilihat. Tidak jarang setelah melihat item-item tersebut, kita jadi "tergoda" untuk melihat satu atau lebih item yang direkomendasikan. Bahkan, bisa berujung pada transaksi pembelian, jika item tersebut dijual.
- Buat kita yang tinggal di kota besar dengan trafik padat, adakah yang belum pernah "ngecek" kemacetan di jalan-jalan kota kita? Kita mungkin jadi batal pergi ke tempat tujuan jika jalan di situ dan sekitarnya berwarna "merah". Ketika kita memilih jalur tercepat dari satu tempat ke tempat lainnya, mesin Google akan memanfaatkan informasi kepadatan lalu-lintas di tiap alternatif jalur untuk memilih yang tercepat. Warna hijau, kuning, oranye dan merah di peta Google telah menjadi informasi penting buat kita!

- Apa saja yang sedang "hot" dibicarakan di dunia maya? Berbagai trending di Twitter menjadi salah satu jawabannya. Di situ juga bisa kita dapatkan informasi sentimen atau persepsi, apakah positif atau negatif, terhadap pesan tertentu.
- Saat kita bepergian, terlebih lagi ke negara 4 musim dimana di suatu wilayah cuacanya dapat berubah dengan cepat (dalam hitungan jam), ponsel kita menjadi sumber informasi yang penting. Kita bisa cek di sekitaran objek wisata yang akan kita kunjungi, pada hari tanggal dan jam kita berada di sana, cuacanya bagaimana. Apakah akan turun hujan/salju? Angin kencang? Suhu *super* dingin atau sangat panas? Dari situ, kita bisa menentukan fashion bagaimana yang cocok untuk kita kenakan. Bisa juga kita batal pergi ke objek itu.
- Pernah membandingkan hasil search di Google dengan keyword tertentu dari satu orang ke orang lain? Bisa beda. Hasil yang diberikan oleh mesin pencari Google akan dibuat sedemikian rupa, dibuat relevan dengan "kebiasaan" pencarian dan browsing kita di Internet.

Contoh-contoh di atas, baru segelintir dari yang sudah dihasilkan para data scientist. Sampai di sini, mungkin para pembaca sudah dapat merasakan atau menduga bahwa untuk menghasilkan tiap layanan di atas, data scientist bekerja dengan data tertentu. Misalnya, untuk menghasilkan rekomendasi item produk, dia menganalisis data transaksi di toko online (e-commerce). Untuk memberikan trending pembicaraan di dunia maya, data yang diproses adalah pesan-pesan Twitter, sedangkan untuk prediksi cuaca, yang diproses adalah data cuaca yang direkam oleh sensor-sensor di berbagai stasiun cuaca di bumi.

Pada abad ke-21 ini data sudah terbuat dan/atau terkumpul dari berbagai sumber (lihat Gambar 1.1). Pembuat data bisa jadi kita sendiri, yang lalu direkam di berbagai sistem, seperti media sosial, penyedia layanan email, chat, blog, review, foto dan video. Dapat juga berupa data bisnis atau data di organisasi (misalnya transaksi pembelian online, supermarket, perbankan, rumah sakit, instansi pemerintah, sekolah, pabrik, dan masih banyak lagi lainnya). Berbagai sensor (misalnya sensor cuaca dan video perekam di jalan, rumah dan perkantoran) dan satelit di angkasa juga berkontribusi banyak menghasilkan rekaman data. Berbagai alat IoT (Internet of Things), misalnya jam yang kita pakai, alat-alat rumah tangga dan mesin industri, juga senantiasa merekam data. Dari banyak jenis sumber tersebut, dapat dikatakan "data tidak pernah tidur". Data terbuat terus dari detik ke detik dalam 24 jam dalam sehari. Pada tahun 2020 ini, diprediksi dihasilkan sekitar 35 zettabytes (10^{21} bytes atau 1.000.000.000.000.000.000.000 bytes) dari seluruh dunia (IBM Cognitive Class-2, 2020).

Dengan sumber-sumber data yang beragam di atas, sudah dapat kita duga bahwa bentuk atau format data yang direkam juga bermacam-macam. Untuk data bisnis atau di organisasi-organisasi, umumnya data terekam dalam format "tabular", seperti data yang kita buat di sheet-sheet Excel. Data berbentuk teks dapat berasal dari email, chat, blog, review maupun medsos.

Data suara dan video, dapat berasal dari medsos maupun sensor. Aliran data "numerik" (berupa angka-angka) dengan format/susunan tertentu diproduksi oleh sensor-sensor. Setiap satelit yang berada di ruang angkasa memiliki tujuan dan kegunaan tertentu, sehingga data yang direkam pun sesuai dengan kegunaannya. Secara umum, data yang direkam adalah "sinyal digital" berupa angka-angka yang contohnya dapat merepresentasikan lokasi, suara dan citra hasil penginderaan satelit itu.

Gambar 1.1. Contoh sumber data.

Barangkali pembaca sudah mendengar atau membaca istilah "big data". Apa itu big data? Apakah data yang berukuran sangat besar? Dengan banyaknya sumber data, apakah jaman sekarang semua data menjadi big data? Belum tentu. Ulasan mengenai big data dengan lebih jelas dapat dibaca di Bab 10.

Sebagian data yang dibahas di atas tersedia di cloud dan dapat diunduh dengan gratis (misalnya data dari media sosial, cuaca dan sebagian data satelit). Ada juga yang dapat dibeli dengan harga terjangkau.

1.2. Apa itu Data Science?

Setelah mengenal contoh pemanfaatan hasil data science, berbagai sumber dan keragaman data, dapatlah diduga bahwa orang-orang yang "ngoprek" data yaitu data scientist, dibutuhkan di berbagai bidang. Bahkan, pada abad ke-21 ini, dimana semua sistem teknologi informasi telah menghasilkan data, data scientist telah dan akan dibutuhkan di semua bidang (industri,

perdagangan, transportasi, layanan, kesehatan, pariwisata, pendidikan, dll). Tapi, apa itu data science?

Sesuai dengan namanya, data science melibatkan data dan sains atau ilmu (yang dibutuhkan untuk memproses data). Data science mulai didengungkan pada tahun 80-an dan 90-an, namun baru benar-benar dipublikasikan pada tahun 2009 atau 2011. Para ahli perintisnya antara lain adalah Andrew Gelman[1] dan DJ Patil[2].

Ada berbagai pendapat tentang definisi data science tapi Profesor Murtaza Haider dari Ryerson University di Kanada memiliki definisi yang cukup mudah dimengerti:
Secara sederhana dapatlah dikatakan bahwa data science "terjadi" ketika kita bekerja dengan data untuk menemukan jawaban atas pertanyaan-pertanyaan (tentunya yang relevan dengan data tersebut). Penekanannya lebih ke data itu sendiri dan bukan tentang sains atau ilmunya (yang dibutuhkan untuk menganalisisnya). Jika kita memiliki data, lalu kita memiliki *curiousity* (rasa ingin tahu) tentang "kandungan" atau "isi" data (yang bermanfaat), lalu untuk menjawab rasa ingin tahu tersebut kita mempelajari data, melakukan eksplorasi terhadap data itu, "memanipulasi"-nya, melakukan berbagai hal untuk menganalisis data tersebut dengan memanfaatkan ilmu dan teknologi tertentu untuk mendapatkan jawaban, itulah data science!

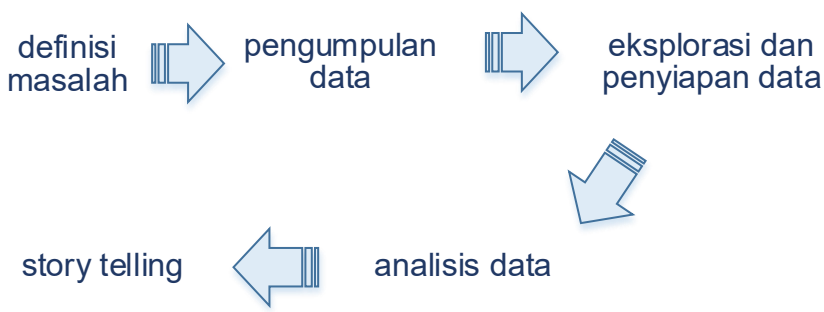
Tujuan akhir dari data science adalah untuk menemukan **insights** dari data. Data science dapat dipandang sebagai proses untuk mendestilasi atau mengekstraksi atau menggali insights dari data. Data yang diolah dapat berukuran sedang hingga sangat besar. Insights tersebut dapat diibaratkan sebagai emas atau berlian, yang meskipun hanya sedikit atau berukuran kecil, namun tetap berharga. Insights dapat berupa informasi penting maupun model-model yang dibuat dari data yang akan bermanfaat dalam mengambil keputusan. Insights yang ingin digali dari data perlu dimulai dengan rasa keingin-tahuan yang kuat dari diri sendiri atau dari organisasi tempat dia bekerja (berupa kebutuhan karena ada masalah yang ingin diselesaikan dengan memanfaatkan data). Berbekal ini, seorang data scientist lalu melakukan berbagai aktivitas dengan memanfaatkan ilmu dan teknologi yang sesuai untuk mendapatkan insights yang disasar.

[1] Profesor di bidang statistik dan ilmu politik dari AS yang telah menulis beberapa buku di bidang data science.
[2] Ilmuwan di bidang matematika dan ilmu komputer dari AS yang telah menulis beberapa buku di bidang data science.

1.3. Apa Saja yang Dikerjakan Data Scientist?

Ibaratnya menambang emas dari gunungan tanah yang melalui proses-proses yang berbelit dan membutuhkan berbagai mesin dan peralatan, untuk menemukan insights dari data (yang dapat berukuran sangat besar juga) pun demikian. Seorang data scientist mengerjakan berbagai pekerjaan dengan alat-alat (tools) pada beberapa tahap untuk untuk mendapatkan insights.

Umumnya data scientist dibutuhkan oleh organisasi-organisasi yang telah memiliki sistem-sistem teknologi informasi operasional sebagai sumber data (lihat Gambar 1.1). Karena "data telah menumpuk" lalu ada kesadaran untuk mendapatkan insights yang bermanfaat. Untuk organisasi bisnis (misalnya perusahaan e-commerce, bank, transportasi dan pariwisata), insights bisa ditujukan untuk memperbaiki organisasi. Perbaikan itu misalnya karyawan menjadi lebih produktif, proses bisnis menjadi lebih efisien sehingga menurunkan biaya operasional, penjualan produk/jasa meningkat sehingga menaikkan keuntungan, layanan ke pelanggan menjadi lebih memuaskan sehingga pelanggan lebih loyal. Untuk organisasi pemerintah yang memberikan layanan kepada masyarakat, misalnya untuk meningkatkan produktivitas pegawai dan memperbaiki layanan. Untuk organisasi riset di bidang sains, kebutuhan akan berbeda, misalnya untuk menemukan model dari data yang bermanfaat untuk melakukan prediksi di masa depan. Model itu misalnya model prediksi panen tanaman, bencana, kebutuhan energi, kebutuhan transportasi penduduk, kerusakan lingkungan, dsb.



Gambar 1.2. Tahapan data science.

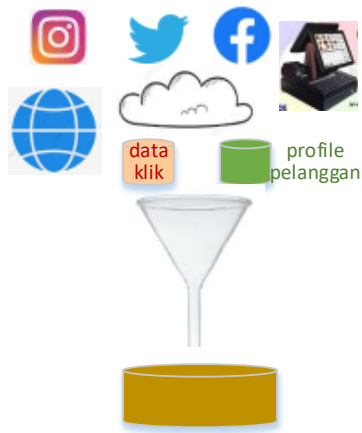
Disarikan dari (EMC, 2015), ketika seorang data scientist bekerja di organisasi-organisasi di atas, secara umum yang dilakukan adalah (lihat Gambar 1.2):

Pertama, tahap pendefinisian masalah. Data scientist mendapatkan kebutuhan organisasi yang harus dicarikan jawaban atau solusi dari data, misalnya menurunkan biaya produksi dan

membuat pelanggan belanja lebih sering (Gambar 1.3). Dapat juga dia menerima insights spesifik yang akan digali dari data. Jika kebutuhan organisasi bersifat umum (misalnya menurunkan biaya produksi), maka data scientist harus mampu untuk merumuskan insights spesifik yang akan digali. Mulai tahap ini, curiosity menjadi bekal yang penting. Adanya curiousity akan memberikan motivasi diri yang kuat yang dibutuhkan untuk menghadapi berbagai tantangan dan kesulitan dalam menggali insights.

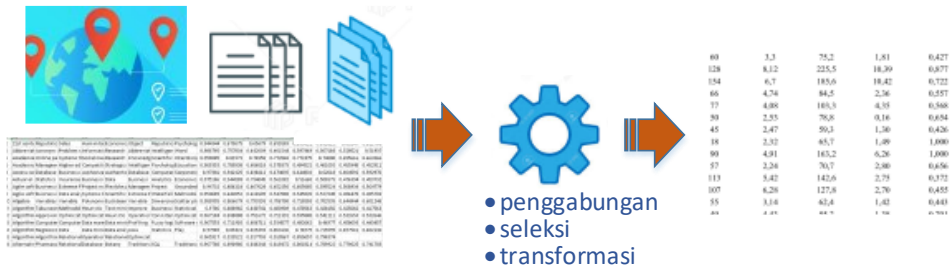
Gambar 1.3. Hal-hal berharga (insights) apa yang dapat digali dari data?

Kedua, tahap pengumpulan data. Berdasar insights yang akan digali, data scientist perlu merumuskan data apa saja yang dibutuhkan. Data itu dapat saja sudah tersedia semua atau baru sebagian. Jika baru sebagian, misalnya baru tersedia data transaksi sedangkan untuk menggali insights dibutuhkan data profile pelanggan dan Twitter, maka data scientist perlu mencari dan mengumpulkan data, yang dapat berasal dari satu atau lebih sumber (Gambar 1.4). Dalam hal tugas pengumpulan data ini kompleks atau berat karena harus dilakukan dengan mengakses berbagai sumber data pada sistem yang besar (dan kompleks pula), data scientist akan membutuhkan bantuan praktisi lain, khususnya data engineer yang tugasnya lebih berfokus dalam infrastruktur dan sistem pengelolalaan data untuk organisasi. Jika sebagian data belum terekam di sistem organisasi namun tersedia di luar organisasi (misalnya data harga saham, kependudukan, cuaca, satelit, yang tersedia di cloud), data scientist (bisa dengan bantuan data engineer) perlu "mengambil" data tersebut. Jika data belum tersedia di sistem organisasi maupun di luar, kemungkinan data scientist perlu untuk "mengadakan" data tersebut, misalnya melalui survei. Semua hal yang dilakukan tersebut harus disertai dengan pertimbangan terhadap isu privasi. Tahap ini dapat dikerjakan dengan cepat atau lama, bergantung kepada ketersediaan data.



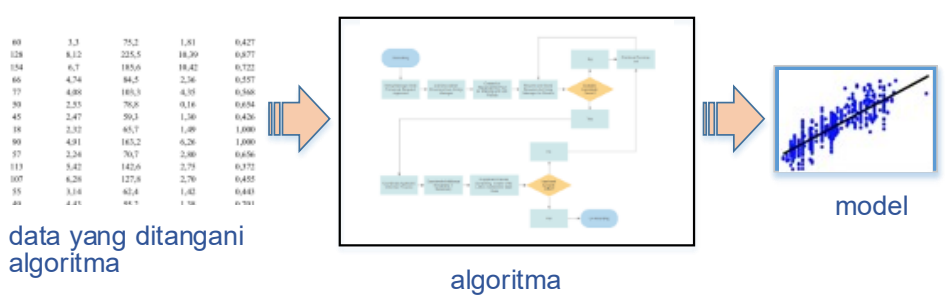
Gambar 1.4. Ilustrasi pengumpulan data kompleks dari berbagai sumber.

Ketiga, tahap eksplorasi dan penyiapan data. Setelah data terkumpul, seluruh komponen data perlu dipelajari dengan seksama. Misalnya, jika data berbentuk sebuah tabel, maka makna dan nilai tiap kolom harus dipahami. Untuk memahami data yang cukup kompleks dan berukuran besar, seringkali perlu dibuat visualisasi, kadang juga perlu komputasi statistik untuk mendapatkan ringkasan data (mencari rata-rata, median, minimum, maksimum juga distribusi data). Data juga harus diperiksa, karena seringkali data hasil pengumpulan tersebut masih "kotor", berisi nilai yang salah atau ada yang hilang. Maka data perlu dicek, apakah semua nilai konsisten, benar atau tidak menyimpang. Jika data perlu diperbaiki, dalam kasus-kasus tertentu perbaikan data dapat dilakukan dengan memanfaatkan konsep statistika. Untuk data tertentu, mungkin juga perlu dilakukan "transformasi", yaitu mengubah nilai data ke bentuk yang dibutuhkan dengan tidak menghilangkan maknanya. Untuk menyiapkan data final (berupa fitur-fitur yang siap untuk diumpankan ke teknik atau algoritma analisis data yang akan digunakan), seringkali dia juga perlu memilah-milah, memilih data (detil ulasan dapat ditemukan di (Han & Kamberlin, 2012)). Ilustrasi pembuatan fitur diberikan pada Gambar 1.5. Jika data kompleks, pekerjaan di tahap ini bisa makan waktu lama dan sumberdaya yang banyak.



Gambar 1.5. Ilustrasi penyiapan data: Berbagai data diintegrasikan, dipilih yang relevan, dan/atau diubah menjadi fitur data yang siap diumpankan ke sebuah algoritma analisis data.

Keempat, tahap analisis data. Jika data yang disiapkan sudah bagus, tahap ini dapat dilakukan dengan relatif lebih mudah, asalkan data scientist sudah menguasai teknik/algoritma, teknologi atau tools yang akan digunakan. Berdasarkan insights yang akan digali, di sini dipilih teknik atau algoritma yang sesuai (dapat berasal dari algoritma Machine Learning yang merupakan subset dari Artifical Intelligent atau Kecerdasan Buatan). Data scientist perlu memahami data yang ditangani, "behavior", prinsip kerja, kelebihan dan kekurangan berbagai algoritma agar dapat memilih algoritma yang tepat. Jika tujuannya untuk membuat model, algoritma lalu dijalankan untuk mengolah data yang telah disiapkan agar dihasilkan model, misalnya model klasifikasi atau prediksi (Gambar 1.6). Model lalu diuji apakah sudah memenuhi standar tertentu. Dalam menguji model, misalnya menguji keakuratan dari model prediksi, data scientist perlu menguasai teknik-teknik pengukuran model (yang biasanya berbasis konsep statistika) dan memilih teknik yang tepat. Hasil uji lalu dievaluasi. Jika kualitas belum memenuhi syarat, model berpotensi tidak dapat dimanfaatkan, karena itu pembuatan model perlu diulangi lagi. Salah satu kemungkinan adalah dengan menyiapkan data masukan yang berbeda. Jadi, tahap pertama perlu diulangi lagi dan dilanjutkan ke tahap berikutnya, sampai didapatkan hasil analisis data yang memuaskan.



Gambar 1.6. Ilustrasi analisis data untuk mendapatkan model.

Kelima, storytelling. Seorang data scientist harus mampu untuk mengkomunikasikan proses dan hasil temuan analisis data dengan sistematis, menarik, tidak ambigu dan mudah dipahami bagi orang-orang (yang berkepentingan dengan proses maupun hasil itu). Bergantung kebutuhan di organisasi tempat data scientist bekerja, komunikasi dapat dilakukan secara tertulis (dalam bentuk laporan) maupun tatap-muka pada rapat atau seminar (Gambar 1.7). Ibaratnya "mendongeng" (telling a story), pembaca atau audiens harus dibuat "terpesona" (impressed) dan percaya dengan hasil-hasil temuannya. Agar menarik dan mudah dipahami, paparan perlu dituangkan dalam bentuk-bentuk visual (yang merepresentasikan data, metoda, model, hasil uji model, dll) yang tepat. Karena itu, data scientist harus mampu menyusun laporan yang sistematis, jelas, berkualitas bagus dan menguasai teknik presentasi yang efektif. Insights yang ditemukan akan menjadi dasar pengambilan keputusan yang bisa jadi berdampak luas, karena itu pihak-pihak yang berkepentingan harus dapat diyakinkan tentang kebenaran temuan itu.

Gambar 1.7. Storytelling dengan berbagai visualisasi.

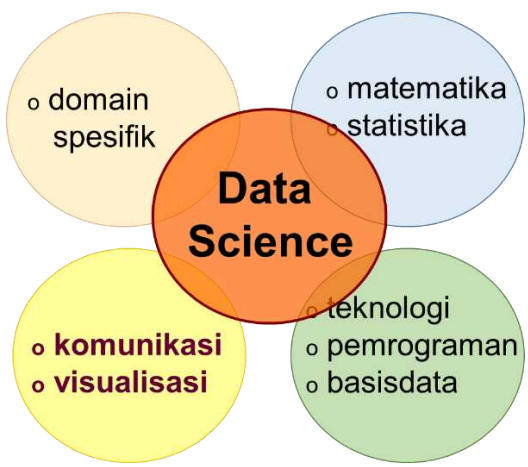
Setelah melakukan storytelling, harapannya tentu saja temuan insights-nya akan dimanfaatkan, menjadi kebijakan, program kerja ataupun *actions* yang tepat terap bagi organisasi. Untuk itu, data scientist perlu memberikan berbagai dukungan yang dibutuhkan. Sesudah hasil temuannya dimanfaatkan, kemungkinan akan muncul masalah-masalah baru yang perlu dicari penyelesaiannya melalui analisis data lagi. Dengan demikian, peran data scientist akan dibutuhkan lagi dan pekerjaan data scientist merupakan pekerjaan yang berkelanjutan.

Jika temuan data scientist berupa model, misalnya yang bermanfaat untuk memprediksi atau memberikan rekomendasi, lalu model tersebut akan "diluncurkan" di aplikasi atau website atau sistem informasi di organisasi, data scientist seringkali perlu bekerja-sama dengan tim pengembang aplikasi/sistem tersebut (karena umumnya pengembangkan aplikasi/sistem informasi tidak menjadi ranah kerja para data scientist). Model yang dihasilkan tersebut kemungkinan juga perlu penyesuaian atau pengembangan dari waktu ke waktu seiring dengan perubahan ataupun bertambahnya data yang dianalisis. Jadi, di sini peran data scientist juga berkelanjutan.

1.4. Keahlian dan Skill Data Scientist

Agar dapat melaksanakan kelima tahap data science itu dengan sukses, bekal ilmu, keahlian dan ketrampilan apa saja yang dibutuhkan untuk menjadi seorang data scientist? Untuk menjadi seorang data scientist, orang harus belajar apa saja?

Secara ringkas, data scientist perlu menguasai beberapa ilmu, keahlian dan ketrampilan yang dapat dikelompokkan menjadi empat (IBM Cognitive Class-2, 2020), yaitu (lihat Gambar 1.8): keahlian substansi di bidang khusus tertentu; matematika dan statistik; teknologi, pemrograman dan basisdata; serta komunikasi dan visualisasi. Keterangan dari setiap kelompok tersebut diberikan di bawah ini.



Gambar 1.8. Keahlian dan skill multi-disiplin data scientist.

Keahlian pada Domain Spesifik

Pada abad 21 ini nyaris tidak ada bidang yang tidak membutuhkan data scientist (lihat Subbab 1.2). Masing-masing organisasi yang bergerak di bidang tertentu (misalnya manufaktur, ritel, transportasi, pariwisata, kesehatan dan pendidikan) memiliki data yang spesifik dan kebutuhan unik yang terkait dengan organisasi mereka. Data scientist harus mampu memahami data dan kebutuhan organisasi tempat dia bekerja agar dapat menggali insights yang tepat dari data yang nantinya bermanfaat bagi organisasi tersebut. Itu sebabnya seorang data scientist perlu memiliki keahlian pada bidang atau domain yang spesifik.

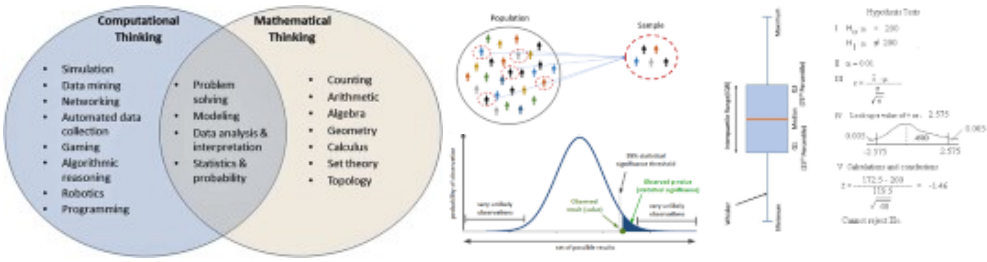
Sebagai contoh, jika seseorang ingin menjadi data scientist bagi perusahaan e-commerce, maka dia membutuhkan ilmu dan skill yang relevan dengan manajemen pelanggan, pemasaran digital, budaya netizen, media sosial dan web analytics. Jika untuk pabrik, misalnya, dia membutuhkan pemahaman terhadap produk yang dibuat, proses produksi, manajemen rantai pasokan, logistik dan pemasaran. Jika untuk pemasaran bidang pendidikan (di universitas), dia harus paham tentang bidang-bidang pendidikan di universitas, pemasaran digital, hubungan manajemen pelanggan untuk dunia pendidikan dan perilaku siswa sekolah menengah. Keahlian khusus yang dibutuhkan data science di bidang kedokteran, lingkungan (yang terkait dengan bumi dan permasalahannya), lembaga antariksa yang mengelola satelit dan perusahaan penyedia transportasi udara, dapat dibayangkan, akan sangat berbeda dengan masing-masing contoh tersebut.

Walaupun keahlian dan ketrampilan data scientist dapat digolongkan ke dalam 4 kelompok, namun dengan menentukan bidang khusus yang tertentu, nantinya seorang data scientist akan membutuhkan bagian ilmu matematika yang tertentu, juga menguasai teknologi, tools, algoritma dan pemrograman yang tertentu pula. Sebagai contoh, teknologi, teknik-teknik atau algoritma-algoritma yang digunakan untuk menganalisis data satelit, secara umum akan berbeda dengan yang digunakan untuk mengolah data transaksi perusahaan e-commerce dan data klik pengunjung website.

Matematika dan Statistik

Sebelum data science ditemukan, orang sudah memanfaatkan statistik untuk menganalisis data. Misalnya, statistik dimanfaatkan untuk mendapatkan distribusi atau sebaran data, “ringkasan” data (seperti frekuensi kemunculan, rata-rata, median, minimum, maksimun, percentile 25-75%, dsb), pengujian hipotesis, juga membuat sampel data dan melakukan analisis multivariat. Pada saat mempelajari dan mengeksplorasi data, data scientist seringkali menggunakan statistika untuk memahami data. Jika kemudian dia mendapati ada data yang salah atau tidak

konsisten, data scientist juga perlu menangani hal ini (istilahnya, "membersihkan data") antara lain dengan memanfaatkan statistika. Statistika juga dibutuhkan ketika data scientist perlu mengubah satu nilai ke nilai lain (istilahnya, "mentransformasi data"). Bergantung kepada insights yang akan digali dari data, kadang analisis data juga dapat dilakukan dengan statistika (beserta visualisasi hasilnya). Penguasaan statistika juga dibutuhkan ketika data scientist menguji insights yang berupa model, untuk mengukur tingkat kebenaran model atau membandingkan berbagai model yang didapatkan untuk dipilih yang terbaik.



Gambar 1.9. Ilustrasi matematika dan statistik untuk data scientist.

Matematika di sini konteksnya luas, termasuk kemampuan berpikir secara logis, sistematis, dan matematika diskret. Jadi, tidak hanya ilmu matematika seperti aritmatika, aljabar, kalkulus, himpunan, geometri, dsb. Jika insights yang akan digali dari data berupa model, misalnya model yang dapat digunakan untuk melakukan prediksi di masa depan, maka Machine Learning perlu digunakan. Setiap algoritma Machine Learning (seperti pengelompokan, klasifikasi data, regresi, analisis aturan asosiasi, outlier, dll.) dirancang berbasiskan matematika dan statistik. Karena itu, penguasaan matematika menjadi dasar bagi data scientist dalam memahami berbagai algoritma Machine Learning. Berbekal pemahaman yang memadai terhadap algoritma-algoritma itu, data scientist lalu dapat memilih algoritma-algoritma yang cocok untuk digunakan dalam menganalisis data yang sudah disiapkan. Ilustrasi untuk kelompok bidang ini diberikan pada Gambar 1.9.

Teknologi, Pemrograman dan Basisdata

Data yang akan dianalisis pastilah tersimpan di suatu (atau beberapa) tempat penyimpanan data. Sistem yang menyimpan dan mengelola data dinamakan sistem basisdata. Sistem ini dapat mengelola data berformat terstruktur (bertipe tabular), semi terstruktur (misalnya data dengan format HTML, CSV, JSON dan XML, juga data spasial atau data geografis), maupun tidak terstruktur (misalnya dokumen, email, foto dan video). Berdasarkan format yang macam-macam tersebut, sudah dapat dibayangkan bahwa sistem basisdata yang mengelola tiap tipe

data juga berbeda. Misalnya, sistem basisdata relasional, menangani data terstruktur, sedangkan basisdata NoSQL utamanya menangani data semi-trestruktur dan tidak terstruktur. Sistem basisdata juga ada yang berjalan di atas sistem big data (misalnya, Hadoop dan Spark) maupun di cloud. Seorang data scientist harus mampu untuk "mengambil" dan memanipulasi data yang tersimpan di basisdata. Maka, dia harus menguasai konsep basisdata dan teknologi basisdata yang menyimpan data yang akan dianalisisnya. Selain itu, dalam mengambil, memilih, memeriksa data dan menyimpan hasil data yang disiapkan ke sistem basisdata, dia juga harus mampu memprogram dengan bahasa pemrograman yang digunakan oleh sistem basisdata itu, misalnya SQL pada basisdata relasional (MySQL, Oracle, SQL Server, dll), HQL pada basisdata berbasis objek, PostgreSQL pada Postgres, HiveQL pada Hive (yang berjalan di atas Hadoop), SparkSQL untuk Spark dan BigQuery untuk datawarehouse Google Cloud (lihat Gambar 1.10).

Gambar 1.10. Berbagai teknologi dan tools analisis data.

Dalam melakukan eksplorasi, menyiapkan maupun menganalisis data yang telah disiapkan, data scientist dapat menggunakan software atau tools yang sesuai dengan data yang diprosesnya. Tools untuk data bisnis yang berformat tabular, akan berbeda dengan tools untuk data teks, citra maupun spasial. Untuk data berukuran kecil sampai sedang, misalnya, Excel dapat digunakan untuk visualisasi, penyiapan data sampai analisis. Namun, jika data scientist perlu menganalisis data teks (misalnya pesan Twitter), dia membutuhkan tools lain. Kemudian, walaupun berformat tabular, tapi jika ukuran data sangat besar (bergiga-giga) dan sudah tersimpan di sistem big data, maka analisis perlu dilakukan dengan software untuk big data (misalnya Hive dan SarkSQL). Berbagai software analisis data maupun layanan cloud sudah menyediakan fitur-fitur Machine Learning. Untuk menganalisis data dengan algoritma tertentu, data science dapat memanfaatkan fitur yang sudah disediakannya. Sekarang sudah tersedia berbagai tools baik untuk data kecil maupun sangat besar, baik yang berjalan di komputer desktop, jaringan maupun cloud, juga untuk berbagai jenis data. Data scientist harus mampu memilih satu atau lebih tools yang tepat dan menggunakannya dengan baik untuk melaksanakan tugasnya.

Tools untuk menganalisis data saat ini cukup banyak yang dapat diperoleh dengan gratis.

Gambar 1.11. Contoh bahasa pemrograman bagi data scientist.

Dalam mengumpulkan, mempelajari, menyiapkan data, seorang data scientist seringkali harus memprogram ("ngoding"). Bahkan, di tahap analisis data, jika tidak ada tools yang memiliki fitur yang tepat untuk digunakan, dia juga perlu memprogram (untuk mengimplementasikan algoritma analisis data yang khusus). Untuk dapat memprogram, dia harus mampu berfikir secara sistematis dan terstruktur dan memahami cara bekerja sistem komputer. Dia harus mampu berpikir analitis agar dapat merancang langkah-langkah pada program atau algoritma program. Dia juga harus memiliki pehamanan terhadap matematika dan statistika yang kuat agar dapat menerjemahkan rumus-rumus menjadi program dengan tepat dan benar. Terdapat berbagai pilihan bahasa pemrograman, masing-masing memiliki kegunaan, kelebihan dan kekurangannya sendiri (Gambar 1.11), misalnya Python, R, Java, dan yang digunakan pada sistem basisdata yang sudah dibahas di atas (SQL, HQL, PostgreSQL, dll). Jika dia bekerja menganalisis big data yang juga tersimpan pada sistem big data, dia perlu memprogram dengan salah satu atau lebih dari pilihan ini: MapReduce pada Hadoop, Scala (untuk memanfaatkan library Machine Learning pada Spark) dan SparkSQL (untuk mengakses data terstruktur pada Spark), HiveQL (untuk mengakses data terstruktur pada Hive). Jika data tersimpan di cloud, dia perlu memprogram dengan bahasa yang digunakan di layanan cloud itu, misalnya BigQuery.

Komunikasi, Visualisasi dan Softskill Lainnya

Sebagaimana dipaparkan pada tahap-tahap data science, setelah menemukan insights dari data, data science harus mampu untuk mengkomunikasinnya (baik secara tertulis maupun tatap-muka) dengan efektif, menggunakan berbagai bentuk visual yang menarik, bergaya story-telling. Maka, keahlian story-telling dan visualisasi harus dikembangkan terus-menerus oleh data scientist. Karena dia harus mampu merancang bentuk-bentuk visual dengan menerapkan seni, maka dia harus menguasai berbagai tools untuk visualisasi data (misalnya Excel, Tableau atau

lainnya seperti ditunjukkan pada Gambar 1.12) atau mampu memprogram untuk menghasilkan bentuk visual khusus yang menarik (misalnya distribusi data pada peta).
Dalam menjalankan tahap-tahap analisis data yang seringkali penuh tantangan dan harus berkoordinasi dengan berbagai pihak, seorang data scientist perlu memiliki passion (kecintaan) terhadap yang dikerjakan, curious terhadap data, hacker-mindset, problem-solver, berpikir strategis, bersikap proaktif, kreatif, inovatif dan kolaboratif.

Gambar 1.12. Contoh tools untuk membuat visualisasi.

1.5. Era Industri 4.0 dan Data Science

Pada awal abad ke 21 ini dunia memasuki era revolusi Industri 4.0. Data Science seringkali dikaitkan dengan era ini. Lalu, apa itu sebenarnya Industri 4.0? Dilansir dari sebuah artikel pada majalah bisnis Forbes, berikut ini ulasan ringkasnya (Marr, 2009):

Sampai saat ini, revolusi industri sudah terjadi 4 kali, yaitu:

- Pertama, terjadinya mekanisasi peralatan industri dengan memanfaatkan tenaga air dan uap.
- Kedua, pabrik-pabrik mampu melakukan perakitan atau produksi barang secara masal dengan menggunakan tenaga listrik.
- Ketiga, industri mulai mengadopsi komputer-komputer dan proses otomatisasi dengan memanfaatkan sistem cerdas, menggunakan data dan algoritma-algoritma Machine Learning.
- Selanjutnya, Industri 4.0 terjadi seiring dengan ketersediaan berbagai sistem teknologi informasi, peralatan Internet of Things (IoT) dan Internet yang makin mudah diakses dan digunakan. Pada era 4.0, berbagai sistem teknologi informasi tersambung secara digital sehingga mampu berkomunikasi untuk saling berbagi data dan informasi. Peralatan dan mesin-mesin makin pintar karena dilengkapi dengan kemampun untuk menangkap dan memproses atau menganalisis data. Dengan memanfaatkan jaringan dan peralatan yang serba pintar tersebut, sebuah sistem juga dimungkinkan membuat keputusan tanpa campur tangan manusia. Kegiatan industri jadi makin efisien, para produsen makin produktif.

Tiga contoh penerapan Industri 4.0 diberikan di bawah ini:

Identifikasi peluang: Idustri 4.0 menawarkan peluang bagi pabrik-pabrik untuk meningkatkan efisiensi a.l. dengan mempercepat proses produksi. Ini dimungkinkan karena masalah (penting) yang terjadi dapat diidentifikasi dengan cepat dan segera dicari solusinya. Pada era ini, mesin-mesin saling terhubung sehingga dapat mengumpulkan berbagai data dalam jumlah yang besar, dimana setelah diproses dapat memberikan informasi yang terkait tentang pemeliharaan, kinerja dan masalah lain yang dapat segera ditindak-lanjuti. Selain itu, data yang terkumpul juga dapat dianalisis untuk mencari pola-pola dan insights (informasi berharga) yang tidak mungkin "digali" secara manual oleh manusia.

Optimasi logistik dan rantai-pasokan (supply chain): Sistem rantai-pasokan yang terintegrasi dapat dibuat lebih responsif atau adaptif. Sistem dapat segera melakukan penyesuaian atau perubahan ketika menerima sebuah informasi baru. Misalnya, ketika sistem menerima informasi tentang terjadinya cuaca buruk yang menghambat pengiriman barang pada bagian delivery (sehingga stok barag menumpuk), sistem "bersikap" proaktif, segera mengubah prioritas produksi di bagian prabrik untuk mengatasinya.

Internet of Things (IoT): Komponen kunci pada Industri 4.0 adalah IoT yang dicirikan dengan saling tersambungnya peralatan-peralatan IoT dan pemanfaatan cloud untuk menyimpan data yang dikirim dari peralatan IoT (secara instant atau real time). Pelaku industri lalu dapat memanfaatkan layanan hasil analisis data di cloud tersebut untuk membuat operasi peralatan-peralatan mereka menjadi lebih efisien (tanpa harus melakukan analisis data sendiri yang dapat membutuhkan sumber daya yang tidak terjangkau).

Dari ulasan ringkas di atas, dipaparkan bahwa di era Industri 4.0, mesin-mesin atau berbagai alat atau sistem dibuat menjadi pintar (seolah-olah mampu berpikir dan memutuskan sendiri) karena dilengkapi dengan kemampuan untuk mengambil data, menganalisis data, atau mengambil informasi penting hasil analisis data dari mesin lain atau dari cloud. Informasi ini lalu digunakan sebagai dasar untuk bertindak (melakukan aksi). Jadi, Industri 4.0 tidak terlepas dari analisis berbagai data yang hasilnya dimanfaatkan berbagai mesin dan alat.

1.6. Kebutuhan Data Science

Beberapa laporan hasil survei dan analisis dari berbagai lembaga menyampaikan bahwa data scientist telah menjadi kebutuhan global maupun di Indonesia.

Untuk lingkup global, berikut ini informasi dari beberapa sumber:

- McKinsey & Company, penyedia layanan konsultasi manajemen dan strategi bisnis, melaporkan bahwa teknologi Artifical Intelligent (AI) yang termasuk Machine Learning makin banyak dibutuhkan karena memberikan keuntungan-keuntungan di bidang bisnis (McKinsey, 2018).
- World Economic Forum (WEC) melaporkan kebutuhan data scientist yang meningkat pada berbagai bidang, misalnya pada industri yang berbasis teknologi informasi, media dan hiburan, layanan finansial dan investasi, layanan profesional, pemerintah, dll. (WEC, 2019).
- Asia-Pacific Economic Cooperation (APEC) pada laporan tahun 2017 menuliskan: *Data Science and Analytics (DSA) skills are in high demand, but supply is critically low with employers facing severe shortages* (APEC, 2017).
- LinkedIn, organisasi yang mengelola jaringan para profesional di Internet yang terbesar, pada laporan tahun 2020 menempatkan data scientist di *top 10 emerging jobs* di berbagai negara, seperti Amerika Serikat (Linkedin-US, 2020), Kanada (Linkedin-CA, 2020), Australia (Linkedin-Aus, 2020). Demikian juga di kawasan ASEAN, seperti Singapore (Linkedin-Sing, 2020) dan Malaysia (Linkedin-Malay, 2020).

Bagaimana dengan di Indonesia?

Banyak perusahaan mencari data scientist atau pakar artificial intelligence, maupun data engineer. Semua itu terkait dengan pengolahan data. Hal tersebut dapat kita temui di lowongan-lowongan pekerjaan di banyak perusahaan Indonesia. Pada laporan yang dirilis LinkedIn tahun 2020, kebutuhan data scientist di Indonesia menempati urutan ke empat (Linkedin-Indon, 2020). Mengapa kebutuhannya begitu besar? Seperti disampaikan oleh Taufik Susanto[3], doktor pada bidang data science lulusan Queensland University of Technology Australia dan pendiri konsultan di bidang data science, saat ini pengolahan data menjadi penentu kompetisi bisnis antar perusahaan. Taufik memberi ilustrasi kompetisi Gojek versus Grab. Siapa yang mampu membuat profile pelanggannya, memilih perhitungan harga yang tepat dan promo yang tepat maka dia akan menjadi pemenangnya. Trend ini minimal sampai 5 tahun kedepan akan sama dengan saat ini. Bahkan mungkin bisa lebih intense lagi.

[3] https://www.techfor.id/pendidikan-data-science-di-indonesia/ [diakses 12 Juni 2020]

Lalu industri apa saja yang membutuhkan data science? Menurut Taufik, pada jaman sekarang semua perusahaan/industri bahkan institusi pendidikan sangat membutuhkan data science. Kalau perusahaan ritel seperti Tokopedia, Bukalapak, Blibli dan Lazada tidak mampu untuk mengolah data dengan baik maka akan collapse (tidak mampu bersaing). Demikian juga otomotif. Industri pariwisata juga membutuhkan data science.

Berdasarkan hasil survei, saat ini kebutuhan data scientist di Indonesia diperkirakan baru terpenuhi sekitar 50%4. Hasil survei yang dilakukan oleh Sharing Vision terhadap 27 perusahaan (dan dipublikasikan pada Januari 2019) menunjukkan bahwa 66% responden menilai Big Data akan booming di Indonesia pada 1-2 tahun ke depan. Selain itu, hasil survei ini juga menunjukkan bahwa 48% perusahaan sudah memasukkan pengembangan sistem Big Data ke dalam IT Strategic Plan, bahkan 33% di antaranya sudah mengoperasikan sistem tersebut dan 33% lainnya sedang mengembangkan sistem big data.

Belum terpenuhinya lowongan data scientist di Indonesia ini, sejalan dengan ini: Pada laporan *Global Skills Index 2020* yang diterbitkan oleh Coursera (penyelenggara kursus daring global dengan 65 juta peserta anggota), untuk bidang Data Science, Indonesia ditempatkan pada posisi *lagging* atau tertinggal. Dari 60 negara (di benua Amerika, Eropa, Asia, Afrika dan Australia) yang ditelaah, Indonesia berada di posisi 56. Padahal untuk bidang teknologi, Indonesia berada di posisi *emerging*, urutan ke 31(Coursera, 2020).

1.7. Informasi Bab-bab Buku

Konten buku ini dibagi menjadi dua bagian, dengan deskipsi sebagai berikut:

Bagian Pertama: berisi bab ini dan contoh aplikasi data science dan pekerjaan data scientist dalam menggali insights pada berbagai bidang spesifik. Inti konten dari tiap bab diberikan dibawah ini:

- Bab 2: Aplikasi statistik dan visualisasi terhadap data *smartwatch* yang dikenakan pada para partisipan penelitian untuk mendapatkan pola belajar dan tidur yang mendukung prestasi akademis yang bagus.
- Bab 3: Paparan sederhana tentang bagaimana rekomendasi item pada website e-commerce "dihitung" berdasar data *rating* pengunjung, dengan memanfaatkan algoritma *collaborative filtering item-based*.

[4] https://jabar.sindonews.com/berita/4305/1/sdm-data-scientist-di-indonesia-masih-minim#:~:text=%22Kebutuhan%20data%20scientist%20saat%20ini,22%2F1%2F2019)

- Bab 4: Pemanfaatan teknik *clustering* (pengelompokan) untuk menganalisis data menu dan bahan masakan yang hasilnya dapat digunakan untuk membantu kita dalam memilih menu kuliner yang sesuai selera.
- Bab 5: Dari data pengindera jauh satelit beserta data lain yang didapatkan di sawah, dapat dibuat model prediksi dengan regresi untuk memperkirakan jumlah panen padi (ketika sawah masih hijau).
- Bab 6: Contoh-contoh pemanfaatan berbagai bentuk visualisasi dari data untuk mendapatkan insights dari data dengan studi kasus data COVID-19.
- Bab 7: Informasi yang terkait dengan pola hidup sehat dapat diperoleh dari aplikasi ponsel yang mengambil data *smartwatch*. Bab ini memberikan paparan sederhana tentang teknik klasifikasi dengan Jaringan Syaraf Tiruan, yang merupakan cikal-bakal dari sistem *deep learning*. Data yang dikumpulkan dan dianalisis adalah data aktivitas pemakai *smartwatch,* sedangkan model klasifikasi yang dibuat dimanfaatkan untuk memprediksi kualitas tidur pemakainya.
- Bab 8: Pemanfaatan algoritma *user-based collaborative fitlering* dan algoritma pengelompokan Fuzzy c-Means untuk menganalisis data rating film dan hasilnya dapat dimanfaatkan untuk memberikan rekomendasi film yang cocok bagi penonton.
- Bab 9: Pemaparan tentang bagaimana para pengguna ponsel berkontribusi dalam mengumpulkan data yang dimanfaatkan Google untuk memberikan informasi tentang kepadatan trafik di peta Google. Juga tentang bagaimana kita dapat melakukan praktek kecil-kecilan untuk menganalisis data trafik dari peta tersebut bagi kepenting kita.

Bagian Kedua: berisi paparan yang lebih teknis yang terkait tentang big data dan contoh hasil penelitian dosen dan mahasiswa yang terkait dengan big data dan Data Science. Inti konten dari tiap bab adalah:

- Bab 10: Pemaparan tentang big data, mengapa sekarang populer, dan berbagai teknologi yang sudah tersedia untuk mengumpulkan, memanajemen dan menganalisis big data.
- Bab 11: Contoh pemanfaatkan teknologi big data, khususnya Spark, Hadoop, Kafka untuk mengumpulkan aliran data Twitter dan menganalisisnya dengan statistika sederhana.
- Bab 12: Pengembangan dan perbandingan algoritma pengelompokan paralel k-Means pada lingkungan sistem big data Hadoop dan Spark.
- Bab 13: Pengukuran dimensi tubuh dengan memanfaatkan data dari perangkat permainan konsol Xbox (yang memiliki sensor untuk menangkap dan mengenali gerakan dan gestur tubuh pemain). Hasilnya berpotensi untuk dimanfaatkan, misalnya, pada penentuan ukuran pada pembelian baju secara daring.
- Bab 14: Data berupa foto (citra) seringkali perlu di-praolah terlebih dahulu agar dapat dianalisis lebih lanjut, misalnya untuk keperluan pengenalan bentuk-bentuk objek pada

citra. Bab ini memaparkan segmentasi citra untuk mempraolah data citra menggunakan algoritma Particle Swarm Optimization.

Dengan beberapa variasi konten pada bab-bab di atas, diharapkan para pembaca akan mendapatkan pengetahuan awal tentang Data Science dan big data yang memadai.

Referensi

(APEC, 2017) Asia-Pacific Economic Cooperation (APEC) – Human Resource Development Working Group, *Data Science and Analytics Skills Shortage: Equipping the APEC Workforce with the Competencies Demanded by Employers*, July 2017.

(Coursera, 2020) Coursera, *Global Skills Index*, 2020

(EMC, 2015) EMC Education Services, *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, Wiley Publ., USA, 2015.

(IBM Cognitive Class-1, 2020) IBM Cognitive Class, *Introduction to Data Science*, https://cognitiveclass.ai/courses/data-science-101 (diakses 6 Juni 2020)

(IBM Cognitive Class-2, 2020) IBM Cognitive Class, *Big Data 101*, https://cognitiveclass.ai/courses/what-is-big-data (diakses 6 Juni 2020)

(Han & Kamberlin, 2012) J. Han & Kamberlin, *Data Mining Concept and Techiques 3rd Ed.*, Morgan Kauffman Publ., USA, 2012

(Linkedin-US, 2020) Linkedin, *2020 US Jobs Trends*, 2020.

(Linkedin-CA, 2020) Linkedin, *2020 Canada Emerging Jobs*, 2020.

(Linkedin-Aus, 2020) Linkedin, *2020 Emerging Jobs Report Australia*, 2020.

(Linkedin-Sing, 2020) Linkedin, *2020 Emerging Jobs Report Singapore*, 2020.

(Linkedin-Malay, 2020) Linkedin, *2020 Emerging Jobs Report Malaysia*, 2020.

(Linkedin-Indon, 2020) Linkedin, *2020 Emerging Jobs Report Indonesia*, 2020.

(Marr, 2009) B. Marr, *What is Industry 4.0? Here's A Super Easy Explanation For Anyone*, Forbes, 2 September 2018, https://www.forbes.com/sites/bernardmarr/2018/09/02/what-is-industry-4-0-heres-a-super-easy-explanation-for-anyone/#318549f99788 [diakses 13 Juni 2020]

(McKinsey, 2018) McKinsey & Co, *Analytics comes of age*, New York, NY, USA, 2018.

(WEC, 2019) World Economic Forum, *Data Science in the New Economy: A new race for talent in the Fourth Industrial Revolution*, Swiss, 2019.