BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topic 1

Business Intelligence, Data Science and Data Mining

LEARNING OUTCOMES

- Setelah mempelajari materi ini peserta kuliah diharapkan mampu mengidentifikasikan dan memahami:
  - Tujuan umum dari mata kuliah
  - Literatur Utama yang digunakan
  - Software pendukung
  - Decision support system
  - Pengertian Business Intelligence, Business Analytics, , Data science and data mining

OUTLINE MATERI :

1. Pendekatan Konseptual
2. Data Analytical Thinking
3. Pengertian Data Sains, Data Engineering, dan Pengambilan Keputusan Berbasis Data (Data-Driven Decision Making)
4. ”Data Processing” dan "Big Data"
5. Kesimpulan

PENDAHULUAN

Kuliah ini berisi seperangkat konsep dasar atau prinsip yang mendasari berbagai teknik untuk mengekstrak pengetahuan yang berguna dari data. Konsep ini berfungsi sebagai *dasar* bagi banyak algoritma data mining yang terkenal. Selain itu, konsep ini mendasari analisis masalah bisnis yang berpusat pada data, penciptaan dan evaluasi solusi data sains, dan evaluasi strategi dan usulan data sains secara umum. Dengan demikian, dalam kuliah ini akan dibahas seputar prinsip-prinsip umum dan bukan tentang algoritma tertentu.

Anda tidak diasumsikan memiliki penguasaan matematika yang canggih. Namun, terkadang tidak dapat dihindari bersinggungan dengan materi yang agak teknis – dengan tujuan untuk memberikan pemahaman yang agak mendalam tentang data sains dan data mining. Namun secara umum, pembahasan matematis diusahakan minimal dan fokus pada penyajian yang se"konseptual" mungkin.

Pendekatan Konseptual Data sains

Dalam kuliah ini diperkenalkan kumpulan konsep dasar data sains yang dipandang paling penting. Konsep tersebut mencakup proses yang dimulai dari menemukan masalah, menerapkan teknik data sains, dan implementasinya dalam pengambilan keputusan. Konsep ini juga mendasari serangkaian besar metode dan teknik analisis bisnis.

Konsep konsep tersebut dapat dikategorikan ke dalam tiga tipe umum:

1. Konsep tentang bagaimana data sains harus diselaraskan dengan organisasi dan lingkungan kompetitifnya, termasuk cara untuk menarik dan mempertahankan tim data sains; bagaimana data sains menuntun ke arah keunggulan kompetitif; dan konsep taktis dalam implementasi proyek data sains yang sukses.

2. Konsep umum berpikir analitis berbasis data. Ini membantu dalam mengidentifikasi data yang sesuai dan metode yang tepat dalam konteks bisnis tertentu. Konsepnya meliputi proses *data mining* serta berbagai metode data mining *tingkat tinggi*.
3. Konsep umum untuk mengekstrak pengetahuan dari data, yang mendasari beragam teknik data sains dan algoritmanya.

Sebagai contoh, salah satu konsep mendasar dalam data sains adalah menentukan kesamaan dua entitas berdasarkan data. Ini mendasari berbagai analisis seperti menemukan pelanggan yang serupa dengan pelanggan tertentu. Prinsip ini merupakan inti dari beberapa algoritma prediksi seperti memprediksi target value seperti probabilitas penggunaan sumber daya oleh klien atau probabilitas pelanggan menanggapi suatu tawaran. Prinsip ini juga mendasari teknik clustering; mengidentifikasi kelompok kelompok yang memiliki kesamaan fitur tanpa tujuan tertentu. Konsep kesamaan juga membentuk basis bagi *information retrieval*, di mana dokumen atau halaman web yang relevan dengan kueri penelusuran diprioritaskan dalam pencarian. Konsep kesamaan juga mendasari beberapa algoritma umum untuk rekomendasi. Buku berorientasi algoritma tradisional biasanya menampilkan masing-masing teknik dalam Sesi Sesi yang berbeda, dengan nama yang berbeda, kendati terdapat aspek umum yang tersembunyi dalam rincian algoritma atau proposisi matematikanya. Fokus kuliah ini, tertuju pada konsep umun, dengan menampilkan teknik teknik khusus dan algoritma spesifik sebagai manifestasi alami dari konsep tersebut.

Contoh lain, dalam mengevaluasi model, terdapat istilah lift – Istilah ini digunakan untuk menilai efektifitas model; yaitu variasi yang dijelaskan oleh model dibanding kejadian secara kebetulan. Konsep ini ditemukan secara luas dan berulang di dalam dunia data sains; digunakan untuk mengevaluasi model yang berbeda dalam konteks yang berbeda. Algoritma untuk penargetan iklan dievaluasi dengan menghitung lift yang didapat untuk populasi yang ditargetkan. Konsep lift digunakan dalam memberi bobot bukti terhadap suatu kesimpulan. Lift membantu menentukan apakah co-occurrence (asosiasi) dalam data patut diperhatikan, bukan sekadar konsekuensi alami dari popularitas.

Diyakini bahwa menjelaskan data sains seputar konsep dasar semacam itu tidak hanya membantu mahasiswa, namun juga memfasilitasi komunikasi antara pemangku kepentingan

bisnis dan Data saintis. Ini menyediakan kosa kata bersama dan memungkinkan kedua belah pihak saling memahami dengan lebih baik. Konsep bersama mengarah pada diskusi yang lebih dalam yang dapat menemukan masalah kritis yang tidak terjawab.

Pemikiran Data-Analitik (Data – Analytical Thinking)

Dalam lima belas tahun terakhir terlihat investasi yang meningkat dalam infrastruktur bisnis, yang pada gilirannya meningkatkan kemampuan organisasi untuk mengumpulkan data dari seluruh perusahaan. Hampir setiap aspek bisnis terbuka untuk pengumpulan data seperti data: *operations*, manufaktur, manajemen rantai pasokan, perilaku pelanggan, kinerja pemasaran, prosedur alur kerja, dan sebagainya. Pada saat bersamaan, informasi dari pihak eksternalpun tersedia secara meluas seperti tren pasar, berita industri, dan pergerakan pesaing. Ketersediaan data yang luas ini telah menyebabkan meningkatnya kebutuhan akan metode yang dapat mengekstrak informasi dan pengetahuan yang berguna dari data. Inilah ranah data sains.

Ubiquitous Data

Dengan sejumlah besar data yang tersedia, perusahaan di hampir setiap industri memberikan perhatian pada pemanfaatan data untuk keunggulan kompetitif. Di masa lalu, perusahaan dapat mempekerjakan tim statistik, pemodel, dan analis untuk mengeksplorasi dataset secara manual, namun volume dan variasi data jauh melampaui kemampuan analisis manual. Pada saat yang sama, kemampuan komputer menjadi jauh lebih tinggi, kemampuan jaringan telah berkembang, dan algoritma telah sangat maju yang dapat mengakses dataset untuk memungkinkan analisis yang lebih luas dan lebih dalam daripada sebelumnya. Konvergensi fenomena ini telah melahirkan aplikasi bisnis yang dikembangkan berdasarkan prinsip-prinsip data sains dan teknik data mining.

Aplikasi teknik data mining terbanyak ditemukan dalam bidang pemasaran seperti *targeted marketing*, iklan online , dan rekomendasi untuk penjualan silang (cross-selling).

Data mining digunakan dalam CRM secara umum yakni untuk menganalisis perilaku pelanggan agar dapat mengelola atrisi (churn, turnover) dan memaksimalkan customer value yang

diharapkan. Industri keuangan menggunakan data mining untuk penilaian kredit dan perdagangan valas, dan dalam operasi sehari harinya dalam mendeteksi kecurangan (fraud detection) dan manajemen tenaga kerja. Peritel utama seperti Walmart dan Amazon menerapkan data mining di seluruh bisnis mereka, mulai dari pemasaran hingga manajemen rantai pasokan.

Tujuan utama kuliah ini adalah untuk membantu Anda melihat masalah bisnis dari perspektif data dan memahami prinsip-prinsip penggalian pengetahuan yang berguna dari data. Ada struktur dasar pemikiran analitik data , dan prinsip dasar yang harus dipahami. Ada juga area tertentu dimana intuisi, kreativitas, akal sehat, dan pengetahuan domain harus diikutsertakan. Perspektif data akan memberi Anda struktur dan prinsip, dan ini akan memberi Anda kerangka untuk menganalisis secara sistematis masalah tersebut. Seiring Anda mendapatkan pemikiran analitik yang lebih baik, Anda akan mengembangkan intuisi mengenai bagaimana dan di mana menerapkan kreativitas dan pengetahuan domain.

Dalam dua sesi pertama kuliah, akan dibahas secara rinci berbagai topik dan teknik yang berkaitan dengan data sains dan data mining. Istilah "data sains" dan "data mining" sering digunakan secara bergantian, pada tingkat tinggi, data sains adalah seperangkat prinsip dasar yang mememberikan tuntunan bagaimana mengekstraksi pengetahuan dari data. Data mining adalah pekerjaan mengekstraksi pengetahuan dari data, melalui berbagai teknik yang didasarkan prinsip-prinsip ini. Sebagai istilah, "data sains" sering diterapkan secara lebih luas daripada "data mining", namun teknik data mining memberikan beberapa ilustrasi paling jelas tentang prinsip-prinsip data sains.

Adalah penting untuk memahami data sains meskipun Anda tidak akan pernah menerapkannya sendiri. Pemikiran analitis memungkinkan Anda mengevaluasi proposal untuk proyek data mining. Misalnya, jika seorang karyawan, konsultan, atau target investasi potensial mengusulkan untuk memperbaiki aplikasi bisnis berbasis data mining, Anda harus dapat menilai proposal secara sistematis dan memutuskan apakah layak atau tidak. Ini tidak berarti Anda mengetahui apakah itu benar-benar akan berhasil, tetapi bisa menemukan kekurangan yang nyata, asumsi yang tidak realistis, dan bagian bagian yang hilang atau terputus.

Kuliah ini menjelaskan sejumlah prinsip data sains fundamental, dan akan menghadirkan setidaknya satu teknik data mining yang mewujudkan prinsip tersebut. Untuk setiap prinsip

biasanya ada banyak teknik spesifik yang mewujudkannya, jadi dalam kuliah ini contoh yang dipilih bertujuan untuk menekankan prinsip-prinsip dasar dalam preferensi teknik tertentu. Ini menekankan, kita tidak akan mengangkat perbedaan antara data sains dan data mining, kecuali di mana ia akan memiliki efek substansial dalam memahami konsep tertentu.

Dalam bagian berikut akan dikaju dua studi kasus singkat untuk menganalisis data untuk mengekstrak pola prediktif.

Contoh: Badai Frances

Berita *New York Times* di tahun 2004:

Badai Frances sedang dalam perjalanan melintasi Karibia, mengancam langsung pantai Florida. Warga bergegas mencari perlindungan di tempat yang lebih tinggi, namun di tempat lain di salah satu sudut kota, Bento nville, Ark., Eksekutif Wal Mart melihat bahwa situasi tersebut menawarkan kesempatan besar bagi salah satu senjata berbasis data terbaru mereka yaitu teknologi prediktif.

Seminggu menjelang diterjang badai, Linda M. Dillman, CIO WalMart, mendesak stafnya untuk membuat prakiraan berdasarkan kejadian yang terjadi ketika Badai Charley menerjang beberapa minggu sebelumnya. Didukung oleh triliunan bytes data history pembelian yang disimpan di data warehouse Wal-Mart, dia merasa bahwa perusahaan itu 'dapat meramalkan apa yang akan terjadi, alih-alih menunggu hal itu terjadi,' (Hays, 2004)

Perhatikan mengapa prediksi berbasis data mungkin berguna dalam skenario ini. Mungkin diprediksi bahwa pihak yang tinggal di jalur badai akan membeli lebih banyak air kemasan. Mungkin, tapi pemikiran ini mungkin bukan hal baru, dan mengapa kita membutuhkan ilmu data untuk menemukannya?

Akan lebih bermanfaat untuk menemukan pola yang tidak biasa. Untuk melakukan ini, analis dapat memeriksa sejumlah besar volume data Wal Mart dari situasi sebelumnya yang serupa (seperti Hurricane Charley). Dari pola seperti itu, perusahaan mungkin bisa mengantisipasi permintaan yang tidak biasa dan produk yang diserbu pada situasi serupa.

Memang, itulah yang terjadi. The New York Times (Hays, 2004) melaporkan bahwa: "... para ahli data mining menemukan bahwa toko tersebut memang memerlukan produk tertentu - dan bukan hanya barang seperti senter biasa. "Kami belum tahu di masa lalu bahwa penjualan Tart-Pop stroberi meningkat, tujuh kali tingkat penjualan normal menjelang badai," kata Dillman dalam sebuah wawancara . "Dan barang terlaris yang terjual menjelang badai adalah bir."

Contoh: Memprediksi Customer Churn

Bagaimana analisis data semacam itu dilakukan? Pertimbangkan skenario bisnis kedua yang lebih khas dan bagaimana hal itu bisa ditangani dari perspektif data. Masalah ini akan menjadi contoh yang akan dijadikan ilustrasi bagi banyak masalah yang diangkat dalam kuliah ini dan memberikan kerangka acuan yang sama.

MegaTelCo, salah satu perusahaan telekomunikasi terbesar di Amerika Serikat, menghadapi masalah besar dengan retensi pelanggan dalam bisnis nirkabel mereka. Di wilayah mid-Atlantic, 20% pelanggan ponsel beralih ke provider lain saat kontrak mereka berakhir, dan semakin sulit untuk mendapatkan pelanggan baru. Karena pasar ponsel sekarang jenuh, pertumbuhan besar di pasar nirkabel sulit diraih. Perusahaan komunikasi terlibat dalam pertempuran untuk menarik pelanggan kompetitor sambil mempertahankan pelanggan sendiri. Pelanggan beralih dari satu perusahaan ke perusahaan lain disebut *churn*. Apabila ini terjadi, maka akan mendatangkan kerugian yang besar. Suatu perusahaan harus mengeluarkan insentif untuk menarik pelanggan baru sementara perusahaan lain kehilangan pendapatan saat pelanggan tersebut beralih.

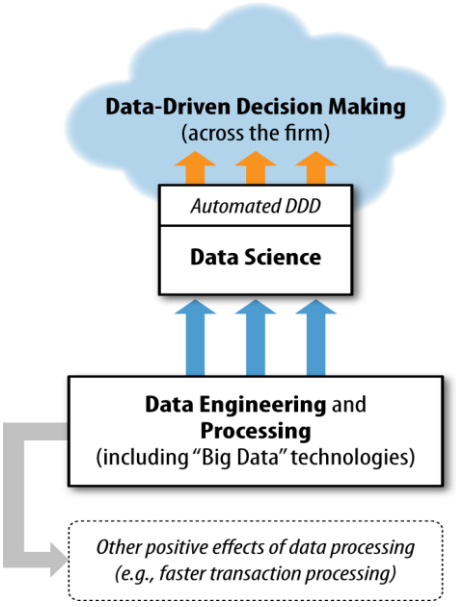
Menarik pelanggan baru jauh lebih mahal daripada mempertahankan yang sudah ada, jadi anggaran pemasaran sebaiknya dialokasikan untuk mencegah churn. Bagian pemasaran telah merancang penawaran khusus untuk program retensi. Tugas Anda dan tim data sains adalah bagaimana menggunakan sumber data MegaTelCo yang sangat besar untuk menentukan pelanggan mana yang harus ditawarkan program khusus tersebut sebelum kontrak mereka berakhir.

Data apa yang mungkin Anda gunakan dan bagaimana menggunakannya? Secara khusus, bagaimana menentukan pelanggan yang menjadi target tawaran khusus tersebut agar mengurangi

pengeluaran dana insentif? Menjawab pertanyaan seperti ini jauh lebih rumit daripada yang terpikir sebelumnya. Masalah ini akan dihadirkan berulang kali dalam kuliah ini, dan kecanggihan solusi akan ditingkatkan secara bertahap saat mendiskusikan konsep konsep fundamental data sains. Retensi pelanggan telah menjadi fokus utama teknologi data mining di bisnis telekomunikasi dan keuangan.

Pengertian Data Sains, Data Engineering, dan Pengambilan Keputusan Berbasis Data (Data-Driven Decision Making)

Data sains melibatkan prinsip, proses, dan teknik untuk memahami fenomena melalui analisis data secara otomatis. Tujuan akhir dari data sains adalah pendukung bagi pengambilan keputusan, yang menjadi tantangan bagi pelaku bisnis.



Gambar 1-1. Data sains dalam konteks berbagai proses yang berkaitan dengan data dalam organisasi.

Gambar 1-1 memperlihatkan kedudukan data sains dalam konteks berbagai proses terkait data dalam organisasi. Data sains perlu dibedakan dari pekerjaan pengolahan data lainnya yang semakin mendapat perhatian dalam dunia bisnis.

Data driven decision making (DDD- Pengambilan keputusan berbasis data) mengacu pada praktik yang mendasarkan keputusan pada analisis data, bukan pada intuisi semata. Misalnya, seorang penjual bisa merancang iklan berdasarkan pengalamannya yang panjang di dalam bisnis dan pengamatannya yang peka akan faktor faktor penyebab keberhasilan. Ia mendasarkan pilihannya pada analisis data mengenai bagaimana konsumen bereaksi terhadap iklan yang berbeda. Ia menggunakan kombinasi dalam pendekatan ini. Level pemanfaatan DDD dalam pengambilan keputusan berbeda beda.

Manfaat pengambilan keputusan berbasis data telah ditunjukkan secara meyakinkan. Ekonom Erik Brynjolfsson dan rekan-rekannya dari MIT dan Penn's Wharton School melakukan studi tentang bagaimana DDD mempengaruhi kinerja perusahaan (Brynjolfsson , Hitt , & Kim, 2011). Mereka mengembangkan ukuran terhadap DDD yang menilai perusahaan mengenai seberapa instens data digunakan dalam membuat keputusan. Mereka menunjukkan secara statistik, semakin tinggi penggunaan data dalam pengambilan keputusan, semakin produktif perusahaan tersebut; dan perbedaannya tidak kecil. Satu standar deviasi lebih tinggi pada skala DDD dikaitkan dengan peningkatan produktivitas sebesar 4% sampai 6%. DDD juga berkorelasi dengan tingkat pengembalian aset, return on equity, utilisasi aset, dan nilai pasar yang lebih tinggi, dan hubungan tersebut nampaknya bersifat kausal.

Jenis keputusan yang diamati dalam kuliah ini berkaitan dengan dua jenis: (1) keputusan terkait "penemuan" sesuatu yang baru yang digali dari data, dan (2) keputusan yang berulang, terutama dalam skala besar. Contoh Walmart menggambarkan masalah tipe 1: Linda Dillman ingin menemukan pengetahuan yang akan membantu Walmart mempersiapkan kedatangan Badai Frances.

Pada tahun 2012, Target, salah satu pesaing Walmart menjadi fokus pemberitaan terkait pengambilan keputusan berbasis data dengan kasusnya sendiri, yang juga merupakan contoh masalah tipe 1 (Duhigg , 2012). Seperti kebanyakan pengecer, Target peduli dengan kebiasaan belanja konsumen, apa yang mendorongnya, dan factor apa yang dapat mempengaruhinya.

Konsumen cenderung memiliki inersia dalam kebiasaan berbelanja mereka yang sulit berubah. Pengambil keputusan di Target memahami bahwa kehadiran seorang bayi yang baru lahir dalam suatu keluarga adalah satu titik di mana orang hampir pasti mengubah kebiasaan berbelanja mereka secara signifikan. Kata analis Target, "Begitu mereka membeli popok dari kita, mereka juga akan mulai membeli barang-barang lainnya." Sebagian besar pengecer mengetahui hal ini dan mereka saling bersaing untuk menjual produk terkait bayi kepada keluarga baru. Karena kebanyakan catatan kelahiran bersifat publik, pengecer memperoleh informasi tentang kelahiran dan mengirimkan penawaran khusus kepada keluarga baru.

Namun, Target ingin terdepan di dalam kompetisi ini. Mereka tertarik untuk mengetahui apakah seseorang tengah menunggu kelahiran bayi. Jika ini bisa dilakukan, mereka akan mendapatkan keuntungan dengan melakukan penawaran sebelum pesaing mereka. Dengan menggunakan teknik data sains, Target menganalisis data historis pada pelanggan yang kemudian disimpulkan bahwa pelanggan itu dalam keadaan hamil, dengan cara mengekstrak informasidari histori tersebut. Misalnya, ibu hamil sering mengubah diet mereka, membeli suplemen, dan sebagainya. Indikator ini dapat digali dari data historis, dirangkai menjadi model prediktif, dan kemudian digunakan dalam kampanye pemasaran.

Model prediktif merepresentasikan sebagian kompleksitas dunia nyata, memusatkan perhatian pada seperangkat indikator tertentu yang berkorelasi dengan kuantitas atau karakter tertentu (siapa yang akan beralih, pembeli potensial, siapa yang sedang hamil, dll). Baik dalam contoh Walmart dan Target, analisis data tidak dilakukan hanya untuk menguji hipotesis sederhana. Data dieksplorasi dengan harapan ada sesuatu yang berguna yang akan ditemukan.

Contoh churn menggambarkan masalah DDD tipe 2. MegaTelCo memiliki ratusan juta pelanggan, masing-masing kandidat berpotensi untuk beralih. Puluhan juta pelanggan memiliki kontrak yang akan berakhir setiap bulannya, sehingga memiliki kemungkinan peningkatan jumlah churn dalam waktu dekat. Jika terhadap setiap pelanggan dapat diprediksi tingkat kemungkinan churn maka akan mendatangkan keuntungan besar dengan menerapkan kemampuan ini kepada jutaan pelanggan dalam populasi. Logika yang sama berlaku untuk banyak area seperti: pemasaran langsung (direct selling), periklanan online, penilaian kredit, perdagangan keuangan, manajemen help desk, deteksi penipuan, peringkat pencarian, rekomendasi produk, dan seterusnya.

Pada era 1990-an, pengambilan keputusan secara otomatis mengubah industri perbankan dan konsumen secara dramatis. Bank dan perusahaan telekomunikasi juga menerapkan sistem skala besar untuk mengelola keputusan identifikasi kecurangan berbasis data. Karena sistem ritel semakin terkomputerisasi, keputusan terkait merchandising juga terotomatisasi. Contoh terkenal adalah sistem rekomendasi otomatis Amazon dan Netflix. Saat ini kita melihat sebuah revolusi dalam periklanan, yang dipicu oleh peningkatan jumlah konsumen yang berbelanja secara online. Sebagai konsekwensi kemampuan ini memungkinkan keputusan pemilihan iklan secara real time.

Pengolahan Data (Data Processing) dan "Big Data"

Tidak semua aktivitas data processing merupakan data sains. Data engineering dan data processing sangat penting untuk mendukung data sains. Namun data sains memiliki pengertian yang lebih umum. Belakangan ini banyak keterampilan pemrosesan data, sistem, dan teknologi sering kali salah dikategorikan sebagai data sains. Data sains membutuhkan akses terhadap data dan mendapat manfaat dari rekayasa data yang canggih. Namun teknologi ini bukan termasuk teknologi data sains. Data engineering mendukung data sains, seperti yang ditunjukkan pada Gambar 1-1 , namun berguna untuk lebih banyak hal seperti pemrosesan transaksi yang efisien, pemrosesan sistem web modern, dan pengelolaan kampanye iklan online.

Teknologi "big data" (seperti Hadoop, HBase , dan MongoDB) mendapat banyak perhatian media baru-baru ini. Big data pada dasarnya adalah dataset yang terlalu besar untuk sistem pengolahan data tradisional, dan oleh karena itu memerlukan teknologi pemrosesan baru. Seperti teknologi tradisional, teknologi big data digunakan untuk banyak tugas, termasuk rekayasa data. Teknologi big data digunakan terutama untuk mendukung data mining dan aktivitas data sains lainnya, seperti yang ditunjukkan pada Gambar 1-1 .

Sebuah studi dilakukan oleh ekonom Prasanna Tambe dari NYU's Stern School, yang meneliti sejauh mana teknologi big data dapat membantu perusahaan (Tambe , 2012). Ia menemukan bahwa, penggunaan teknologi big data meningkatkan produktivitas yang signifikan. Secara khusus, satu standar deviasi pemanfaatan yang lebih tinggi dari teknologi big data diasosiasikan

dengan peningkatan produktivitas 1% sampai 3% lebih tinggi daripada rata-rata perusahaan; satu standar deviasi yang lebih rendah dalam hal penggunaan big data berakibat pada penurunan produktivitas sebesar 1% sampai 3% . Hal ini menyebabkan perbedaan produktivitas yang sangat besar pada perusahaan-perusahaan besar.

Dari Big Data 1.0 Ke Big Data 2.0

Salah satu cara untuk membandingkan teknologi big data adalah dengan mengambil analogi dengan adopsi teknologi internet dalam bisnis. Dengan Web 1.0, organisasi bisnis fokus pada penerapan teknologi internet dasar, sehingga mereka dapat memperlihatkan eksistensinya di web, membangun kemampuan perdagangan elektronik, dan meningkatkan efisiensi operasi mereka. Big Data 1.0 dapat dianalogikan sebagai, investasi perusahaan dalam membangun kemampuan big data, terutama untuk mendukung operasi , dan meningkatkan efisiensi.

Begitu perusahaan telah mengadopsi teknologi Web 1.0 secara menyeluruh (dan dalam prosesnya telah menurunkan harga teknologi ini), mereka mulai melihat lebih jauh. Mereka mulai bertanya apa yang Web bisa lakukan untuk mereka, dan tanpa disadari memasuki era Web 2.0, di mana sistem baru dan perusahaan mulai memanfaatkan sifat interaktif Web. Perubahan yang diakibatkan oleh pergeseran pemikiran ini sangat luas; di antaranya komponen keterlibatan jejaring sosial , dan bangkitnya "suara" konsumen.

Fase Big Data 2.0 diharapkan mengikuti Big Data 1.0. Begitu perusahaan telah mampu mengolah data masif dengan cara yang fleksibel, pertanyaannya adalah: "Apa yang saat ini bisa dilakukan, yang tidak dapat dilakukan sebelumnya, atau dapat melakukan dengan lebih baik dibanding sebelumnya? " Prinsip dan teknik yang akan diperkenalkan dalam kuliah ini akan dapat diterapkan jauh lebih luas dan mendalam daripada yang dapat dilihat saat ini.

Penting untuk dicatat bahwa di era Web 1.0 beberapa perusahaan telah menerapkan gagasan Web 2.0 jauh di depan. Amazon adalah contoh utama, menggabungkan "suara" konsumen sejak awal, dalam penilaian produk, dalam ulasan produk (dan lebih dalam, dalam penilaian ulasan produk). Amazon adalah perusahaan terdepan, yang memberikan rekomendasi berdasarkan data yang masif. Pengiklan online harus memproses data dalam jumlah sangat besar (miliaran tayangan iklan per hari) dan mempertahankan throughput yang sangat tinggi (sistem penawaran waktu nyatanya harus membuat keputusan dalam puluhan milidetik). Industri ini dan industri sejenis menjadi petunjuk kemajuan data dan big data yang kemudian akan diadopsi oleh industri lain.

Data dan Kemampuan Data Sains sebagai Aset Strategis

Bagian sebelumnya mengangkat salah satu prinsip dasar data sains: data, dan kemampuan mengekstrak pengetahuan yang berguna dari data, harus dipandang sebagai aset strategis. Terlalu banyak bisnis yang menganggap analisis data berkaitan terutama untuk mewujudkan nilai dari data yang ada, dan seringkali tanpa memperhatikan apakah perusahaan memiliki sumber daya manusia yang memadai dalam melakukan analisis. Data dan sumber daya manusia yang berkemampuan analitis haruslah saling melengkapi. Tim data saintis terbaik bisa saja menghasilkan nilai yang tidak signifikan tanpa data yang sesuai; data yang tepat seringkali tidak dapat secara substansial memperbaiki keputusan tanpa data saintis yang sesuai. Sebagai aset, kedua hal ini memerlukan investasi. Studi kasus berikut memperlihatkan bahwa berinvestasi dalam aset data dapat terbayarkan.

Cerita klasik tentang Bank Signet dari tahun 1990an menyediakan contoh kasus yang menarik. Sebelumnya, di tahun 1980an, data sains telah mentransformasikan bisnis *consumer credit*. Pemodelan probabilitas terjadinya kecurangan telah mengubah industri dari penilaian secara personal terhadap kemungkinan gagal bayar ke strategi skala besar, yang memungkinkan terealisasinya *economies of scale*. Pada saat itu, kartu kredit pada dasarnya memiliki harga yang seragam, karena dua alasan: (1) perusahaan tidak memiliki sistem informasi yang memadai untuk menangani penetapan harga yang berbeda dalam skala besar, dan (2) manajemen bank percaya bahwa pelanggan tidak dapat menerima adanya diskriminasi dalam pemberian kredit. Sekitar tahun 1990, dua orang visioner (Richard Fairbanks dan Nigel Morris) menyadari bahwa teknologi informasi cukup mampu sehingga mereka dapat melakukan pemodelan prediktif yang lebih canggih dengan menggunakan teknik yang akan dibahas di sepanjang kuliah ini. Beberapa istilah akan digunakan seperti , batas kredit, transfer saldo awal yang rendah, cash back, loyalitas poin, dan sebagainya). Ironisnya, kedua pria ini tidak berhasil membujuk bank-bank besar untuk menerima mereka sebagai konsultan dan memberika kesempatan kepada mereka untuk mencoba. Akhirnya, mereka berhasil meyakinkan bank regional kecil di wilayah Virginia: Bank Signet. Manajemen Bank Signet yakin bahwa pemodelan profitabilitas, bukan hanya probabilitas default, adalah strategi yang tepat. Mereka tahu bahwa hanya sebagian kecil pelanggan yang benar-benar berkontribusi pada lebih dari 100% keuntungan bank dari kartu kredit (karena

sisanya macet atau merugi). Jika mereka berhasil dengan model profitabilitas, mereka bahkan bisa membuat penawaran yang lebih baik kepada pelanggan terbaik dari bank-bank besar.

Namun Bank Signet memiliki satu masalah yang sangat besar dalam menerapkan strategi ini. Mereka tidak memiliki data yang sesuai untuk model profitabilitas dengan tujuan memberikan penawaran yang berbeda untuk pelanggan yang berbeda. Belum pernah ada yang melakukan ini.

Apa yang bisa Bank Signet lakukan? Mereka terbawa ke dalam salah satu strategi pokok data sains yaitu: memperoleh data yang diperlukan dengan konsekwensi biaya. Data adalah aset bisnis, yang perlu diinvestasi. Dalam kasus Signet, data dapat dihasilkan pada profitabilitas pelanggan diberikan persyaratan kredit yang berbeda melalui eksperimen. Karena perusahaan melihat kerugian ini sebagai investasi di data, mereka tetap bertahan meskipun keluhan dari para pemangku kepentingan menghantui. Akhirnya, operasi kartu kredit Signet berbalik dan menjadi sangat menguntungkan dan memisahkan diri dari operasi bank yang lain, yang kini berada di bawah bayang bayang keberhasilan kredit konsumen.

Sumber: https://www.capitalone.com/credit-cards/

Fairbanks dan Morris menjadi Chairman dan CEO dan Presiden dan COO, dan terus menerapkan prinsip prinsip data sains di seluruh bisnis bukan hanya pada akuisisi pelanggan tetapi juga retensi.

Gagasan data sebagai aset strategis tentu tidak terbatas pada Capital One, atau industri perbankan lainnya. Amazon mampu mengumpulkan data awal pelanggan online, dan dapat mempertahankan pelanggan lebih mudah, Jelas bahwa Facebook memiliki aset data yang luar

biasa; apakah mereka memiliki hak strategi data sains untuk mengambil keuntungan penuh dari itu adalah pertanyaan terbuka.

Berpikir Analytic berbasis Data

Menganalisis studi kasus seperti masalah churn membantu untuk memahami konsep "Data-analitis." Ketika dihadapkan dengan masalah bisnis, Anda harus dapat menilai apakah dan bagaimana data dapat meningkatkan kinerja. Akan diperkenalkan konsep dasar dan prinsip-prinsip yang memfasilitasi pola berpikir data sains. Dalam kuliah ini akan didiskusikan kerangka kerja agar analisis dapat dilakukan secara sistematis.

Perusahaan di banyak industri tradisional mengeksploitasi sumber daya baru berdasar data untuk keunggulan kompetitif. Mereka memanfaatkan tim data sains yang membawa teknologi canggih untuk meningkatkan pendapatan dan mengurangi biaya. Selain itu, banyak perusahaan baru yang sedang berkembang dengan data mining sebagai komponen strategis. Facebook dan Twitter, bersama dengan banyak perusahaan "Digital 100" (Business Insider , 2012), memiliki valuasi yang tinggi terutama disebabkan aset data yang mereka miliki.

Sebagai tambahan jika ada konsultan mengajukan proposal untuk proyek data mining untuk meningkatkan bisnis Anda, Anda harus dapat menilai apakah usulan tersebut masuk akal. Dengan pemahaman tentang dasar-dasar data sains anda harus dapat mengajukan beberapa pertanyaan untuk menentukan apakah argumen valuasi masuk akal.

Ini memerlukan interaksi yang dekat antara data saintis dan pelaku bisnis yang bertanggung jawab untuk pengambilan keputusan. Perusahaan di mana pelaku bisnis kurang memahami apa yang para Data saintis lakukan akan mengalami kerugian yang besar, karena mereka membuang-buang waktu dan usaha atau lebih buruk lagi, karena mereka akhirnya membuat keputusan yang salah.

Kuliah ini berkonsentrasi pada dasar-dasar data sains dan data mining, yang berisi serangkaian prinsip, konsep, dan teknik serta struktur pemikiran dan analisis. Hal hal tersebut memungkinkan Anda untuk memahami proses data sains tanpa perlu fokus secara mendalam pada sejumlah besar algoritma data mining tertentu.

Data mining dan Data sains

Kuliah ini mencurahkan perhatian pada ekstraksi pola pola yang berguna dari data (Fayyad, Piatetsky -Shapiro, & Smyth, 1996), dan prinsip-prinsip dasar yang mendasari data sains. Dalam contoh prediksi churn, pola perilaku, yang berguna yang dapat membantu kita untuk memprediksi para pelanggan yang mungkin beralih, atau bahwa dapat membantu kita untuk merancang layanan yang lebih baik.

Konsep dasar data sains diambil dari berbagai bidang analisis data. Beberapa konsep dasar didiskusikan berikut ini:

1. Ekstrak pengetahuan yang berguna dari data untuk memecahkan masalah bisnis dapat dilakukan secara sistematis dengan mengikuti proses dengan tahapan yang telah didefinisikan secara baik. The Cross Industry Standard Process for Data Mining, disingkat CRISP-DM (CRISP-DM Project, 2000), adalah salah satu kodifikasi dari proses ini.
2. Dengan jumlah data yang masif, teknologi informasi dapat digunakan untuk menemukan atribut deskriptif informative tentang entitas yang menjadi perhatian. Contoh churn, pelanggan menjadi entitas yang menarik, dan sikap setiap pelanggan mungkin dijelaskan oleh sejumlah besar atribut, seperti penggunaan, sejarah layanan pelanggan, dan berbagai faktor lainnya. Atribut atribut ini memberikan informasi tentang kemungkinan pelanggan meninggalkan perusahaan ketika kontraknya berakhir. Berapa banyak informasi? Kadang-kadang proses ini secara kasar disebut sebagai menemukan variabel yang "berkorelasi" dengan churn (konsep ini akan dibahas lebih jauh).
3. Jika Anda mengamati data secara mendalam, Anda akan menemukan sesuatu, tapi mungkin tidak dapat digeneralisasi di luar data yang Anda amati. Hal ini disebut overfitting.
4. Merumuskan solusi data mining dan mengevaluasi hasil dilakukan secara berhati-hati dan harus dikaitkan dengan konteks di mana mereka akan digunakan.

Inilah empat konsep dasar data sains yang akan dieksplorasi. Disamping itu akan dibahas konsep dasar tersebut secara lebih detil, dan bagaimana mereka membantu kita untuk menyusun pemikiran data analitik dan memahami teknik data mining dan algoritma, serta aplikasi data sains secara umum.

KESIMPULAN

Kuliah ini adalah tentang ekstraksi informasi yang berguna dan pengetahuan dari volume data yang besar, dalam rangka meningkatkan kualitas pengambilan keputusan bisnis. Ini dipicu oleh tersedianya data dalam jumlah besar dan menyebar di hampir semua sektor industri dan unit bisnis, sehingga memiliki peluang untuk data mining.

Keberhasilan bisnis yang berorientasi data saat ini mengharuskan kita untuk memikirkan bagaimana konsep dasar ini diterapkan pada masalah bisnis tertentu untuk memikirkan data-secara analitis. Misalnya, dalam sesi ini kita membahas prinsip bahwa data harus dianggap sebagai aset bisnis, dan begitu kita berpikir ke arah ini, kita mulai bertanya apakah (dan berapa banyak) yang harus kita investasikan pada data. Dengan demikian, pemahaman tentang konsep dasar ini penting tidak hanya untuk ilmuwan data itu sendiri, namun bagi siapa saja yang bekerja dengan ilmuwan data, menggunakan ilmuwan data, berinvestasi pada data, atau mengarahkan penerapan analisis dalam sebuah organisasi.

Berpikir secara data-analitis dibantu oleh kerangka konseptual yang dibahas di sepanjang kuliah ini. Misalnya, melakukan ekstraksi secara otomatis pola dari data adalah proses dengan tahap yang didefinisikan dengan baik, yang merupakan subjek sesi berikutnya. Memahami proses dan tahapan membantu menyusun pemikiran data analitik, dan membuatnya lebih sistematis sehingga kurang rentan terhadap kesalahan dan kelalaian.

DAFTAR PUSTAKA

1. Foster Provost & Tom Fawcett (2013) Data Science for Business: What you need to know about data mining and data analytic thinking, O'Reilly, ISBN: 978-1-449-36132-7.

2. Sharda, R., Delen, D., Turban, E., (2018). Business intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.