Kevin Chen

1/14/17

Notes

Define Data Science

* Brief history
  + Role of CS: Empowering statistic by solving practical problems by providing number crunching and massive storage.
  + The invention of the database management system (DBMS) and Relational DBMS accelerate the pace of marriage between statistics and computer science
  + Early 1990s: big data begins to explode.
  + Late 1990s: data science harnesses the predictive power of data.
  + Mid 2000s: the word “analytics” emphasized statistical and quantitative analysis as well as “predictive” modeling.
* Concepts and scope
  + Big data refers to a data set whose nature defies conventional ways of processing. It is a relative term, constantly changing.
  + Machine learning frees human and depends on algorithms.
  + Data mining: process of discovering a pattern. At the beginning of a mining process, you don’t know what you are looking for. Use various algorithms. Thus, use big data.
  + Data analytics is driven by hypothesis and intentionality.
* Analytics
  + Big data analytics make use of distributed computing technologies and data analytics techniques to process big data sets.
  + Distributed computing is a approach to break down a task into smaller pieces. (Divide and conquer.) Extreme multithreading.
  + Cloud computing: provides a platform to support distribute computing. It provides computer hardware in data centers. A software solution is also necessary (e.g. Hadoop and NoSQL databases)
  + Once you get the hardware and software infrastructures, you’re ready to run data analytics programs to ask specific questions concerning big data sets.
* Enabling technologies (that make data science a reality)
  + Data infrastructures: support how data is shared, processed, and consumed. Example: distributed computing and cloud computing. Virtualization and distributed file sharing (e.g. RAID and HDFS) enable cloud computing.
  + Data management technologies: handled by database management systems (DBMS). Requires scalable, reliable, and efficient ways to manage data. Majority of business-related data is unstructured, but still a lot of data is stored in conventional relational DBMS in a structured format.
  + Visualization technologies: effective communication with nonexperts. Intuitive delivery.

Marketplace

* Fraud detection
  + Machine learning and big data analytics reduce the number of “false positives” in fraud detection (where normal stuff is happening but the system marks the activity as suspicious).
  + Machine learning brings its ability to learn on its own what is the best way to detect a fraud through numerous trial and error.
* Social media analytics
  + Social media generates a lot of data.
  + Also contains data about data, such as location and timestamps.
  + Personality insights by IBM gives companies more info about their customers.
  + Technologies used: text mining and parsing through an application programming interface (API). Manipulation of data for analysis purposes (such as the use of JSON)
* Disease control
  + Social media analytics may be helpful for disease control.
  + Emotional factors (as shown through people’s posts) are linked to heart disease. Emotional stress -> heart disease.
  + Study methods: linguistic analysis (e.g. use of the word “hate” or “wonderful”) and big data analytics.
* Dating services
  + Conventional Match Making techniques: match percentage (how similar two individuals are). Compatibility predictive model
  + Big data analytics are being used to enhance the quality of these services.
  + Keep track of candidates’ online behavior.
* Simulations and Data Science
  + Types of simulations: 1. pure mathematical (e.g. physics engine for video game). But a purely model-driven simulation is not always accurate. 2. Data Driven (feeds real life data into a simulation model, improving the accuracy of a simulation and its predictive power).
  + Useful for climatology (weather forecasting), etc.
* Climate Research and Data Science
  + The Madingley Model project (sponsored by MSFT) simulates the entire ecosystem of Earth. Analyzes the impact of climate changes on all lifeforms on Earth.
  + Data fed into the predictive models: reporting environmental data through social media, sensor readings from various Internet of Things (IoT) devices, and conventional climate data
* Network Security
  + There is a movement toward leveraging machine learning and big data analytics.
  + The next frontier: logging and monitoring
  + Microsoft Azure offers a machine learning service on which users can build their own intrusion detection solutions or use services provided by MSFT

Skills

* Data mining (and analytics)
  + Involve a myriad of data manipulation
  + Text Retrieval: builds on National Language Processing (NLP)
  + Classification: constructs a model labelling a group of data objects into a specific category. Represented in discrete sets.
  + Prediction: builds a model that produces continuous or ordered values that form a trend.
  + Clustering: groups similar data objects into a class. It reveals features distinguishing one class of data objects from the other -> new discoveries
  + Classification vs. clustering: classification starts with predefined labels while the labels are created after the fact for clustering
* Machine learning
  + Elements: model -> trial and error -> improved model -> produces insight
  + Two types of learning: supervised and unsupervised
  + Supervised: reinforced by feedback in the form of training data. Accumulative classification.
  + Unsupervised: no training data or external feedback.
* Natural Language Processing (NLP)
  + Refers to a collection of different ways for a computer to make sense out of its interaction with a human being through a natural language.
  + Areas: CS, computer linguistics, AI, Human Computer Interaction (HCI)
  + Aspects of NLP: Tokenization and parsing isolate each symbol from a text and conduct a grammatical analysis. Sentence segmentation separates one sentence from another in text. Named entity recognition (NER) identifies which text symbol maps to what types of proper names.
  + A significant portion of data is unstructured (data not extracted from a database).
* Statistics
  + Data science stems out of statistics. Data scientists must know statistics.
  + Probability, correlation, variables/distributions/regressions, null hypothesis significance tests (NHST), confidence intervals/t-tests/ANOVA/chi-square
  + Software tools: R, Excel, and SAS.
  + Models commonly used in data: logistic regression, support vector machines (SVMs) and Bayesian methods
* Visualization
  + Helps overcome the challenge of effectively communicating the results of data analytics to a lay audience.
  + Characteristics of effective visualization: display data at multiple levels of details, avoid distorting the data.
  + Tableau is one of the most popular visualization tools.

Roles

* Data scientist or engineer
  + The role of a data scientist: generalist instead of a specialist. Acts as a liaison between the leadership and data science specialists.
  + Entrance barrier to a data science job is relatively low. Passion for data.
* Business intelligence (BI) architect
  + BI: collecting, managing, and processing corporate data to provide actionable information for the leadership and employees of a company.
  + Characteristics of BI: technology and software driven.
  + Architects are usually senior and in the pinnacle of a technical career.
  + Responsibilities: design/implement system architectures to maximize the potential of a company’s assets. Needs to be able to link IT systems throughout a company.
* Machine Learning scientists
  + Program computers to learn on their own.
  + Job requirements: creative and independent, discipline, attention to details and quality.
  + Technical skills required: math (esp. statistics and probability), modeling, computing and programming.
  + Practical IT Skills: proficiency in programming languages (e.g. Python, C++, Java, and R), text mining (awk, grep, find, and sort), distributed computing (e.g. Hadoop and cloud computing).
  + Demand will only grow.
* Business analytics specialist
  + Make things happen: Implement BI architectures according to the direction and supervision of the BI architect.
  + The service of business analytics is being commoditized (e.g. Amazon QuickSight)
* Data visualization developer
  + Industry neutral: data visualization developers can work in any industry.
  + Examples: R&D organizations, media companies, data analytics groups
  + Collaborations: data scientists, BI architects, machine learning specialists, business analytics specialists. Collaborate with goal of identifying the best means to visually express data to develop new insight and make critical business decisions.
  + Required skills: programming (esp. web dev and GUI platforms), database systems, query languages, data visualization software

Certifications

* MCSE Business Intelligence
  + Demonstrates proficiency with MSFT SQL server products.
  + Expected skills: make queries, administer databases, implement data warehouses, develop data models and reports, and design BI solutions.
* Cloudera Certified Professional (CCP)
  + Four big data related certifications: CCP Data Scientist, Cloudera Certified Developer for Apache Hadoop (CCDH), Cloudera Certified Administrator for Apache Hadoop (CCAH), Cloudera Certified Specialist in Apache HBASE (CCSHB)
  + Expected skills: descriptive and inferential statistics on big data, advanced analytical techniques on big data, machine learning at scale
  + Scenario based exams.
* EMC Data Science Associate (EMCDSA)
  + EMCDSA Certification: tests one’s knowledge of virtualization, data analytics, and the role of a data scientist
  + Other EMC certifications: Data Center Architect (EMCDCA), Cloud Architect (EMCCA)
* Oracle Certification
  + Oracle Business Intelligence (OBI) certificate: tested on OBI Applications for Customer Relationship Management (CRM), OBI Applications for Enterprise Resource Planning, installing/building/configuring BI dashboards, making queries and managing systems.

Future of Big Data

* Emerging technologies
  + Cloud computing, big data analytics, and machine learning.
  + Various online retail data services are available in the cloud through various vendors. Makes it cheaper to use data science techniques to solve business problems. Scalability and ease of use of cloud-based data science services.
  + Machine learning: “Deep” learning is growing
* Emerging careers
  + Huge growth in data science careers.
  + How to prepare yourself for these careers: develop a passion for data science, commit yourself, identify a degree program/additional online courses, seek for guidance
  + Data scientist, BI architect, machine learning specialist, business analytics specialist, data visualization developer
* Ethics
  + Existing threats: insider threats, disgruntled employees, industrial spies, yourself (temptation for eavesdropping)
  + The ethical integrity is important. Build intentional and proactive security within a data science product.
  + Code of ethics are still being formed. Evolving profession.
* Professional development
  + Need to be aware of new developments in the entire IT world.
  + One emerging data science technology is in-memory analytics, which means running entire data analytics operations in the main memory of a computer instead of reading some of the data back and forth from secondary memory.
  + Consider the implications of new developments.
  + Considerations: certification and recertification, professional development units, conferences, workshops, networking.

Conclusion

* Next steps
  + Be sure to take: Excel -> Python -> R -> Hadoop Fundamentals -> Data Visualization Fundamentals -> Data Mining -> Big Data Analytics