Week 2: In-class Demo

Three real world datasets:
Medical Dataset : EHR (Electronic Health Records)
Financial Dataset : Generated by credit/debit card transactions
Social Media Dataset : Data collected from Twitter posts

Dipti Kothari | Neeti Jaiswal | Sanchit Lodha | Weiwei Duan | Yulong Pei

Datasets

How they are collected:
Medical : Collected and updated in real time by Hospitals, Clinics, Labs and other health providers
Financial : Credit card companies collect them during each transaction.
Social Media : Twitter collects them when there is a tweet.

Temporal and spatial coverage:
Medical : The data is collected in real time and at any given point of time a disease and its existence at various geo-locations can be populated.
Financial : Timestamp is associated with each transaction and data is generated across globe
Social Media : Timestamps on every tweet along with its location.

Structure:
Medical : Administrative and billing data, Patient demographics, Progress notes, Vital signs, Medical histories, Diagnoses, Medications, Immunization dates, Allergies, Radiology images, Lab and test results.
Financial : amount of truncation, merchant name, time, date and credit/debit card name
Social Media : id, link, retweet, text, author, images, videos

5 V’s

Volume

Medical : Health Care Industry has been generating petabytes of data. For example, Kaiser Permanete, a health network of over 9 million members has around 26.5 to 44 petabytes of data in the form of EHR (Baker, Baker, & Dworkin, 2017).

Financial : Due to the massive number of transactions that happen in a particular company, it has large sets of data to analyze. For example, Mastercard alone has 10 Petabytes of data which they aggregate and augment using 700,000 rules.

Social Media : with ~500 million tweets per day amounting to 100s of Terabytes per day, generating massive amounts of data.

Variety

Medical : demographics, pharmacy data, lab reports (structured data); medical image (unstructured data)

Financial : Structured; Credit/Debit card companies receive data regarding amount of truncation, merchant name, time, date and credit/debit card name

Social Media : basic info of tweets like id, date, link, text, author (structured data); images and videos (unstructured data)

5 V's (cont.)

Velocity
Medical : Every second real time data gets generated as EHRs are updated by labs, clinic or hospitals
Financial : transactions happen continuously, so data is being generated constantly; EX: Mastercard - generate 65 billion transaction per year worldwide; Visa handles 141 billion total transactions in a year
Social Media : very fast, ~6000 tweets per second, 500 million tweets per day

Veracity
Medical : The quality of data is high and there is little scope for dirty data
Financial : Veracity of credit/debit card transaction is high since each transaction is authenticated
Social Media : The quality/accuracy is hard to determine with misinformation; and with typos, abbreviations, slangs, the quality of data also decreases

5 V’s (cont.)

Value

Medical : EHR is a valuable source of data; easier to share patient information among various health providers making it faster for patients to receive health care; disease analysis and control

Financial : Credit card transactions are most reliable data when it comes to detecting fraudulent behavior; when combined with other data can generate complete history of a customer’s spending and earning to verify loan applications; detecting market and consumer trend

Social Media : data can help companies understand consumers’ view on products, market campaign’s impact

Big Data

Medical Dataset

It is a big data set because of the complexity of healthcare results from the diversity of health-related problems and their treatments. The medical dataset is collected from various sources and integration of these data sources causes data to be of very large size, with multiple scales and incongruences.

Finance Dataset

There are about 200 Billion transactions happening every year, distributed among various companies, customers and mode of payment. To store and manage such data we cannot use conventional methods and hence this also qualifies as Big data

Social Media Dataset

Twitter generates around 500 million tweets with 100s of Terabytes of data being shared per day. Needless to say other platforms such as Facebook generating even more every day, this definitely qualifies as big data

Unintended Consequences

Medical + Financial
One can create a customer profile based on customer's medical history and transaction patterns. This data can be misused by various Insurance companies.

Medical + Financial + Social Media
Companies can utilize purchase history from financial records or individual's health issues from medical records to target specific advertisements on individuals' social media accounts.

References

Baker, J. J., Baker, R. W., & Dworkin, N. R. (2017). *Health Care Finance.* Jones & Bartlett Learning. Retrieved January 14, 2018, from https://www.barnesandnoble.com/w/health-care-finance-judith-j-baker/1126352328?ean=9781284118216

What information does an electronic health record (EHR) contain? (n.d.). Retrieved January 16, 2018, from HealthIT.gov:https://www.healthit.gov/providers-professionals/faqs/what-information-does-electronic-health-record-ehr-contain

Twitter Usage Statistics, Retrieved January 16, 2018, from http://www.internetlivestats.com/twitter-statistics/