Mining Health Big Data
- Opportunities and Challenges

Alex Kuo (Ph.D)\textsuperscript{1,2}

\textsuperscript{1} School of Health Information Science
University of Victoria, BC, Canada.
\textsuperscript{2} IEEE Big Data Initiatives (Chair, Education Track)
Outline

- Introduction
- Real World Use Cases
- Big Data Analytics (BDA) Challenges
- Discussion & Future Research
- Concluding Remarks
1. Introduction
Background

- The patient health data ocean...
  - Today, a variety of modern health information systems such as EHRs, CPOE, PACS, CDSS, and lab-systems have generated an unimaginably huge volume of patient data.
  - Worldwide digital healthcare data was estimated to be equal to 500 petabytes ($10^{15}$ bytes), and is expected to reach 25 exabytes ($10^{18}$ bytes) in 2020. Hughes has also predicted that the global growth in healthcare data will be between 1.2 and 2.4 exabytes a year.

What is Big Data in Healthcare?

- **Healthcare Big Data** is different from Big Data in other disciplines in that it includes structured data (e.g. EHR), semi-structured data (e.g. HL7 messages), unstructured data (e.g. clinical notes, medical images), and other types of data (e.g. genetic, public health/behavior data). Huge volume and very heterogeneous of raw data are generated by a variety of health information systems. Those systems are used in many distributed healthcare settings, such as hospitals, clinics, laboratories and physician offices.

**Health Big data**
- Structured EHR Data
- Unstructured Clinical Notes
- Medical Imaging Data
- Genetic Data
- Other Data (Epidemiology & Behavioral)

**Other type Big data**
- Transitional/Lab data
The human body as a source of Big Data

- The diagram at the right shows what makes up that mass amount of health data.

- The amount of healthcare data that exists is growing faster than ever and poses both an opportunity and challenge to the healthcare sectors.

What Big Data can do in healthcare?

- A McKinsey Global Institute study suggests that Big Data has the potential to create more than $300 billion in value/year for US healthcare sector.
- Big Data in healthcare can do many things that previous can’t be done. For example,
  - Provide fresh insights into life science,
  - Design interventions for disease treatment,
  - Enable personalized medicine/ healthcare,
  - Identify healthcare trends,
  - Reduce medical expenditures,
  - Prevent disease outbreaks, and
  - Improve government healthcare policy makings.
2. Real World Use Cases
Case 1: Social Media Big Data Sense Public Health

- Social media Big Data can predict infectious disease outbreaks.

Google Flu Trends correlated with Influenza outbreak


Twitter messaging correlated with cholera outbreak

Case 2: Big Data and Imaging Analysis Yields High-Res Brain Map

- An NIH-funded research team has begun to bring this map of the human brain into much sharper focus.
- The researchers were able to subdivide the cerebral cortex, the brain’s outer layer, into 180 specific areas in each hemisphere. Remarkably, 97% of those areas had never before been described.

Case 3: Intermountain Healthcare's Big Data Solutions Try To Reshape Healthcare

- Intermountain Healthcare, a Utah-based system of 22 hospitals, 185 health groups and an affiliated insurer used more than 90 million of EHRs to perform health outcomes analysis.
  - OutcomesMiner helps organizations better understand how certain factors contribute to patients' outcomes; this helps researchers see how empirical evidence supports or refutes a hypothesis.
  - PopulationMiner helps users study the relationship between treatments and outcomes, with the aim of developing new medications or improving existing ones.

Reference: Brian Eastwood, Big Data Analytics Use Cases for Healthcare IT.
Case 4: A Provincial Healthcare Authority in Canada Developed a BDA Framework for Supporting Healthcare Research

Objective: To discover the association of ADT and DAD
Objective

- To discover the association of VIHA ADT and CIHI DAD
  - Example: Clinicians suspect that frequent movement of patients within the hospital can worsen outcomes. This is especially true in those who are prone to confusion due to changes in their environment (i.e. elderly). Furthermore, frequent room changes can exacerbate the confusion. If this occurs, a physician may attribute their confusion to the possible onset of an infection, a disease or a mental illness, resulting in unnecessary laboratory tests and medications.

Data source & Analytic Tools

- The VIHA clinical data warehouse has 100 fact and 581 dimension tables that in total hosts 10 billion records.
- Hadoop/MapReduce as resource/workflow allocator; HBase as the non-relational, distributed database (NoSQL), and Apache Phoenix, Spark, Drill as analytic tools.
- UVic WestGrid (4412 cores computer cluster) as the HPC platform.
Other use cases

- New York City's Kalvi HUMAN project (http://kavlihumanproject.org) used Big Data in answering how everything—from biology to behavior and environment—affects the human condition.

- Express Scripts is a pharmacy benefit management organization in the USA. Patients filed more than 1.5 billion prescriptions annually with it. The company used these transactions to drive both behavior change and process improvement. The predictive analytics can identify patients at the greatest risk of skipping or missing doses, and then proactively intervene.

- IBM Watson is a supercomputer. IBM Watson Health using the supercomputer’s natural language, hypothesis generation, and evidence-based learning capabilities allow it to function as a clinical decision support system for use by medical professionals.
Health Big Data projects in the world

- United Nations Global Pulse: Big Data For Public Health
  - Analyzing Attitudes Towards Contraception & Teenage Pregnancy Using Social Data
  - Understanding Public Perceptions of Immunisation Using Social Media
  - Analyzing Social Media Conversations to Understand Public Perceptions of Sanitation

- The European “Big Data for Better Outcomes” medicines initiative expects to invest more than $5 billion to apply big data techniques to speed up clinical drug trials while developing a sustainable healthcare delivery system [1].

- US NIH-sponsored Big Data to Knowledge (BD2K) initiative, with a budget of $656 million, funds research in biomedical informatics with a view to facilitate discovery and support new knowledge, and to maximize community engagement [2].

- Government of Canada invests in new genomics “big data” research projects aimed at real-world challenges.
China's State Council has established the broad goals for a Big Data project that will start to take shape in 2020. The program, will be using a mixture of public and private resources to promote public health. The National Health and Family Planning Commission, the ministry in charge of the program [3].

In January 2014, Korea Ministry of Science, ICT and Future Planning and the National Information Society Agency released a Medical Information Consulting program. The program will provide big data services which can help diagnose and customize treatment for patients, and help promote public health and streamline the management of medical facilities [4].

2. https://datascience.nih.gov/bd2k
3. http://www.chinabiotoday.com/articles/20160620_1
3. BDA Challenges
- Big Data are not usable until they can be aggregated and integrated into a manner that computer can process to generate knowledge.
- The Big Data Analytics (BDA) Process:
1. Data Aggregation Challenges

- Big Data research projects usually involve multiple organizations, different geographic locations, large numbers of researchers, and variety of data type.

- The challenges are
  - Where do I find the data I need?
  - How do I retrieve the data?
  - What is the data structure/format?
  - What are the data access (privacy/security) policies?
  - etc.
2. Data Maintenance Challenges

- Since Big Data involves large collections of datasets, it is very difficult to efficiently store and maintain the data in a single hard drive using traditional data management systems such as relational databases.

- Data privacy preservation is a major issue in BDA.
  - For example, the Health Insurance Portability and Accountability Act (HIPAA) requires the removal of 18 types of identifiers, including any residual information that could identify individual patients.
3. Data Integration Challenges

- Unstructured data integration issues
  - Big data are very heterogeneous
    - Unstructured data (e.g. Images, videos and audios)
    - Semi-structured data (e.g. Clinical notes)

- Structured data integration issues
  - Integrating unstructured data is a major challenge for BDA. With structured data integration there are also many issues (e.g. metadata integration).
4. Data Analysis Challenges

- Complexity of the analysis problem
  - For some analysis algorithms, the computing time increases dramatically even with small amounts of data growth (e.g. Bayesian Network, SQL-GROUP BY)

- Scale of the data
  - Even for simple data analysis, it could take several days, even months, to obtain the result when data is very large (e.g. zettabytes, $10^{21}$ bytes).
Distributing tasks over computers.

10,000 computers

(Only takes 1.16 days (32 years/10,000) for the ZB query question)
5. Pattern Interpretation Challenges

- It is challenging to represent and interpret results in a form that is comprehensible to non-experts.
- Tools of Big Data science do not protect us from skews, gaps, and faulty assumptions.
- With large datasets, it is all too easy to unveil significant value by making information transparent.
  
  Example: Schadt et al. study demonstrates the ability to use non-DNA-based information to infer a DNA-based barcode that is sufficiently specific to resolve an individual’s identity in a collection of hundreds of millions of individual genotypic profiles obtained in a completely different context.
4. Discussion & Future Research
Research Topic 1: The need of a Big Metadata Standard

How the gene ‘myosin light chain 2’ was associated with chamber type hypertrophic cardiomyopathy? The similarity relative to a subset of the genes’ features?

Where do I find the relevant data?
- Retrieve data from well known databases such as NCBI Gene and PubMed.
- Other data sources?
- Can we use a single search engine to retrieve all collections?

“The healthcare industry isn’t lacking data. The challenge is to bring together masses of data from disparate sources and synthesize it into actionable information in real time.”

- Metadata integration challenges
  - What is metadata?
    - Metadata is "data about data". It is descriptive information about a particular dataset, object or resource, including how it is formatted, and when and by whom it is collected.
  - Metadata examples
Two metadata examples

<table>
<thead>
<tr>
<th>ID</th>
<th>FirstName</th>
<th>LastName</th>
<th>Sex</th>
<th>Age</th>
<th>Symptom</th>
</tr>
</thead>
<tbody>
<tr>
<td>P45</td>
<td>Li</td>
<td>Lo</td>
<td>Male</td>
<td>47</td>
<td>S87</td>
</tr>
</tbody>
</table>

#,ID,title,gsm,series_id,gpl,status,submission_date,last_update_date,type,source_name_ch1,organism_ch1,characteristics_ch1,molecule_ch1,label_ch1,treatment_protocol_ch1,extract_protocol_ch1,label_protocol_ch1,organism_ch2,characteristics_ch2,molecule_ch2,label_ch2,treatment_protocol_ch2,extract_protocol_ch2,label_protocol_ch2,hyb_protocol,description,data_processing,contact,supplementary_file,data_row_count,channel_count


XXX.CSV
Each system uses its own metadata to describe data.

**System X (Oracle)**

<table>
<thead>
<tr>
<th>PatientName</th>
<th>Sex</th>
<th>Birthday</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex Kuo</td>
<td>1</td>
<td>2-9-1976</td>
<td>301706005</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

**System Y (MySQL)**

<table>
<thead>
<tr>
<th>FirstName</th>
<th>LastName</th>
<th>Gender</th>
<th>DOB</th>
<th>Symptom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Kuo</td>
<td>M</td>
<td>1976-9-2</td>
<td>L02.4</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

Metadata integration issue

Data structure integration issue

Instance integration issue

SNOMED CT (Abscess of foot)

ICD-10 Code
Standardized Big Data annotations

- Standardized annotations can help organize electronic resources, facilitate legacy resource integration, and support archiving and preservation.

- Some important standards
  - Dublin Core (DC),
  - Metadata Encoding and Transmission Standard (METS),
  - IEEE Learning Object Metadata (LOM)
  - ISO/IEC 11179
  - Encoded Archival Description (EAD)
  - Machine Readable Cataloguing 21 (MARC 21)
Different metadata standards serve distinct needs and communities.

Table 1. Example of Metadata Crosswalk Mapping

<table>
<thead>
<tr>
<th></th>
<th>Dublin Core</th>
<th>EAD</th>
<th>MARC 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Element</td>
<td>Title</td>
<td>&lt;titleproper&gt;</td>
<td>245 008a (Title Statement/Title proper)</td>
</tr>
<tr>
<td>Author Element</td>
<td>Creator</td>
<td>&lt;author&gt;</td>
<td>700 1#$a (Added Entry--Personal Name) (with $e=author)</td>
</tr>
<tr>
<td>Date Created Element</td>
<td>Date.Created</td>
<td>&lt;unitdate&gt;</td>
<td>720 8$a (Added Entry--Uncontrolled Name/Name) (with $e=author)</td>
</tr>
</tbody>
</table>

A real word project – Stanford CEDAR

The objective of CEDAR is to create a unified framework that researchers in all scientific disciplines can use to create consistent, easily searchable metadata.
IEEE BDI is forming a research group

- The IEEE Big Data Initiatives (BDI) is forming a research group to study on where there is a need and opportunity for developing a metadata standard for Big Data management.
- The uniform metadata (UM)
  - It is a set of elements to comprehensively and expressively describe data sources.
  - A simple example
### Table 1. Example of Metadata Crosswalk Mapping

<table>
<thead>
<tr>
<th>Element</th>
<th>Dublin Core</th>
<th>EAD</th>
<th>MARC 21</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title Element</strong></td>
<td>Title</td>
<td>&lt;titleproper&gt;</td>
<td>245 00$a (Title Statement/Title proper)</td>
</tr>
<tr>
<td><strong>Author Element</strong></td>
<td>Creator</td>
<td>&lt;author&gt;</td>
<td>700 1##a (Added Entry—Personal Name) (with $e=author)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>720$a (Added Entry—Uncontrolled Name/Name) (with $e=author)</td>
</tr>
<tr>
<td><strong>Date Created Element</strong></td>
<td>Date:Created</td>
<td>&lt;unitdate&gt;</td>
<td>260 ###$c (Date of publication, distribution, etc.)</td>
</tr>
</tbody>
</table>

5. Concluding Remarks
Today, a variety of modern health information systems have generated an unimaginably huge volume of health data—so called "Health Big Data".

Health managers and experts believe that with the data, researchers can easily reveal important information/knowledge to better health policies, improve patient treatments, and eliminate redundancies and unnecessary costs.

Big Data Analytics (BDA) is the process of extracting knowledge from Big Data. It can be considered as a processing pipeline that involves multiple distinct stages including data aggregation, maintenance, integration, analysis, and interpretation. Each stage faces several specific challenges.
Final thoughts

- Due to the broad nature of this presentation topic, the primary emphasis is on discussing the attributes of Big Data in healthcare, real world use cases, and the challenges and potential solutions for health Big Data Analytics. I do not focus on describing the details of any particular techniques or solutions.

- However, my hope is that this presentation will contribute to democratizing Big Data knowledge and advancing Big Data analytics research in healthcare.
Questions?
How Big is 2.4 exabytes?

2.4 exabytes = $2 \times 10^{12}$ phone books (4cm/book)

= 80,000,000 km

= 105 rounds from Earth to Moon

(Distances to the Moon = 384,400 km)
Study Question: List all diabetic patients with *Congestive heart failure* complication who are younger than the average diabetic patient of the patient's home country in the world.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Data Type</th>
<th>Description</th>
<th>Data Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHN</td>
<td>NUMBER(10)</td>
<td>Patient's ID number</td>
<td>1-7000000000</td>
</tr>
<tr>
<td>PatientName</td>
<td>VARCHAR2(30)</td>
<td>Patient's name</td>
<td>PatientName</td>
</tr>
<tr>
<td>Birthday</td>
<td>DATE</td>
<td>Patient's birthday</td>
<td>1903-2013 (DD-MM-YY)</td>
</tr>
<tr>
<td>Sex</td>
<td>NUMBER(1)</td>
<td>Patient's gender</td>
<td>0-1</td>
</tr>
<tr>
<td>Race</td>
<td>NUMBER(1)</td>
<td>Patient's race</td>
<td>1-6</td>
</tr>
<tr>
<td>Phone</td>
<td>NUMBER(10)</td>
<td>Home phone number</td>
<td>2501234567</td>
</tr>
<tr>
<td>Country</td>
<td>NUMBER(3)</td>
<td>Patient's country</td>
<td>1-196</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>VARCHAR2(50)</td>
<td>code</td>
<td>25 sample codes</td>
</tr>
</tbody>
</table>

```sql
SELECT *
FROM patient p0
WHERE
  (INSTR(p0.Diagnosis,'430678008'>0 OR INSTR(p0.Diagnosis,'430679000')>0) AND
  (INSTR(p0.Diagnosis,'42343007'>0)
  AND
  (2013-EXTRACT(YEAR FROM p0.Birthday)) <
  (SELECT AVG(2013-EXTRACT(YEAR FROM p1.Birthday))
    FROM patient p1
    WHERE
    p1.Country=p0.Country AND
    (INSTR(p1.Diagnosis,'430678008'>0 OR INSTR(p1.Diagnosis,'430679000')>0)
    GROUP BY Country)
ORDER BY Country;
```
Computing time for the SQL to query table with different number of rows.

- Using a Dell Inspiron 580 computer with 3.2 GHz CPU, it took 9.26 sec to get the result.
- It is hard to process this query quickly when the table contains 10 billion of rows without indexing.
Study Question: The average age of patients and percentage of females and males.

Suppose it took 1 sec for a computer to obtain the report for 1 Terabytes ($10^{12}$ bytes) records. It will take $\sim$32 years to get the result of 1 Zettabytes ($10^{21}$ bytes) records.