Big Data in Healthcare

Dr. Refael Barkan, M.D. Ph.D.
Head of RDE at HIT
Bar-Ilan University, May 2016
What is Big Data?

• Quite nebulous, in the same way that the term “cloud” covers diverse technologies
• To clarify matters, the four Vs: **Volume**, **Velocity**, **Variety** and **Veracity** are commonly used to characterize different aspects of.
Big Data – Exponential Growth
Big Data - Volume

• Refers to the vast amounts of data generated and shared every second: emails, twitter messages, photos, video, sensor data etc.

• **From 2013 to 2020, the digital universe will grow by a factor of 10 – from 4.4 trillion gigabytes to 44 trillion.**

• It more than doubles every two years (~Moore’s Law).

• Nowadays, in a single day we process amounts equivalent to decades and centuries – from Megabytes ($2^{10}$B) to Exabytes ($2^{60}$B), Zettabytes ($2^{70}$B) and Yottabytes ($2^{80}$B)
**Big Data - Velocity**

- Refers to the speed at which:
  - new data is generated
  - data moves around
  - data is required to be processed

- Examples:
  - Need to check credit card transactions for fraudulent activities as soon as possible
  - Competition among trading systems in their ability to quickly analyze social media networks to pick up signals that trigger decisions to buy or sell shares
Big Data - Variety

• Refers to the different types of data we can now use

• In the past we focused on structured data that neatly fits into tables or relational databases - easy to manipulate, store and analyze.

• Today, however, the majority of the world’s data is unstructured data (e.g. photos, video sequences, blogs or social media updates), and cannot be easily converted into structured data.
Big Data - Veracity

- Refers to the messiness or trustworthiness of the data, biases, noise and abnormality in data.
- With many forms of Big Data, quality and accuracy are less controllable.
The Data Revolution & Evolution

<table>
<thead>
<tr>
<th>FROM</th>
<th>TO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data as a fact of life</td>
<td>Data as a factor of production</td>
</tr>
<tr>
<td>Collecting</td>
<td>Connecting</td>
</tr>
<tr>
<td>Analyzing</td>
<td>Predicting</td>
</tr>
<tr>
<td>Information</td>
<td>Insights</td>
</tr>
<tr>
<td>“One version of the truth”</td>
<td>Multiple sources and perspectives</td>
</tr>
<tr>
<td>Structured</td>
<td>Unstructured</td>
</tr>
<tr>
<td>Relational</td>
<td>Non-relational</td>
</tr>
<tr>
<td>Centralized processing</td>
<td>Distributed parallel processing</td>
</tr>
<tr>
<td>Terabytes</td>
<td>Petabytes, exabytes, ...</td>
</tr>
<tr>
<td>Analytics as niche</td>
<td>Analytics for everyone</td>
</tr>
<tr>
<td>Limited participation</td>
<td>An era of experimentation and innovation</td>
</tr>
</tbody>
</table>
Healthcare’s (Big) Data

Big Data in healthcare can be broadly summarized into three categories:

• **Traditional medical data** - primarily originated from electronic medical records (EMRs)
• **“Omics” data** - refer to large-scale datasets in the biological and molecular fields
• **Patient Generated Data (PGD)**
“Omics” Data

- Include: genomics data, transcriptomics (the set of all RNA molecules), proteomics data, metabolomics, protein-DNA interactions, protein-protein interactions, phenomics (biomarkers), epigenetics, microbiomics, lipidomics etc.

- Analyzing this data yields understanding of the mechanisms of diseases and acceleration of the individualization of medical treatments (e.g. “precision medicine”).
Epigenetics

• The study of changes in gene expression or cellular phenotypes, caused by mechanisms other than changes in the underlying DNA sequence

• Some of these changes have been shown to be heritable.

• For example, prions (infectious proteins) are clearly epigenetic, perpetuating themselves through altered folding states.
Microbiomics

• The **microbiome** is the collective existence of non-human cells in the human body.

• **An estimated 90% of cells found in the human body are not human at all.**

• Though considerably smaller in size, these approximately **100 trillion cells** add up to a mass of almost **1–2 kg** in an adult individual.
Microbiomics

• Recently, increasingly viewed as a new human organ.
• The microbiome has been associated with both health and a growing number of human diseases and conditions.
• Examples: irritable bowel syndrome, obesity, diabetes, asthma, arthritis, cardiovascular disease, and autism
PGD – Patient Generated Data
PGD Examples

• Google has developed a smart contact lens diabetics can use to read blood sugar levels through the tears in their eyes, providing a new glucose reading as often as once a second.

• Ralph Lauren has unveiled recently the Polo Tech, a compression shirt outfitted with sensors. The biometric shirt operates like a Fitbit, providing information on heart rates and stress levels.
Information Chaos

• Healthcare providers, the physicians and the patients have more data than ever before:
  ○ Physicians through the EMRs
  ○ Patients through a ePHRs, mHealth apps, PGD (e.g. sensors)
• And we keep talking about patient empowerment…
The Physician Point of View

- I see 20 patients per day, spend 10 minutes with each, don’t have enough time.
- My patients are becoming multi-morbid, more and more data…
- I hate the EMR, I spend too much time on it and it’s hard to find information.
- I can not really track my patient compliance.
- Patients demand a lot, they may sue me (and win…)
- Don’t even think about additional information from sensors, genetics, etc. It’s already too much for me…
The Patient Point of View

- I feel that the physician looks at the computer and does not really listen to me.
- I am lost and confused. For every disease there is a specialist and they don’t seem to talk with each other.
- The physician does not have enough time to explain to me what should I do to improve my condition.
- The Internet and patients group provide me with a lot of information, but what is the best treatment for ME?
Information Chaos
Barriers for Patient Empowerment

- Patient-Physician Information Gap
- Lack of Patient Proactive Role
- Lack of Support for Multi-morbidities
- Integration of patient’s platforms (e.g. mHealth) with the EMR
Patient-Physician Information Gap

• Physician faces with a formidable task of integrating the clinical story from many sources in a short time.
• Patients have limited access to medical data.
• Can patient transform the data into an easy to use knowledge?
Patient-Physician Information Gap

• Do doctors provide patients with an overall (and concise) view of disease progression? Multi-morbidities?
• Displaying key indicators in cool graphs is not enough!
• Patients are left with a partial picture and lack the context (“why-do-I-have-to-do-this?”), resulting in
  o Lack of motivation
  o Compliance problems
Patient-Physician Information Gap

- Can the physician and the patient really talk?
- Do they have a common language?
- Can the patient literally “take home” the message?
Lack of Patient Proactive Role

- Patients are left with the task of managing themselves.
- But are they involved in crafting the treatment plan?
- Do they understand:
  - the reasoning behind the treatment plan?
  - the effect a successful implementation of the plan will have on their health?
- Reducing the instructions to a series of tasks and reminders is sufficient?
Lack of Support for Multi-morbidities

- Multi-morbidity is the new norm!
- Make the complex treatment plan even more complex
  - Conflicting goals
  - Conflicting treatments
- Applications support multi-morbidity?
Integration of mHealth and EMR

- Data gathered by mHealth applications is often not integrated with the patient's EMR medical history.
- Physician cannot access it if not integrated.
Data Use and Health Analytics

• Increase in life expectancy and in the prevalence of chronic diseases & multi-morbidities means increasing number of interactions with healthcare providers and increase of data generated.

• EHR data of a single patient include Petabytes of data (2^{50} B).

• Personalized medicine and specifically the relatively new omics-data may result in Exabytes (2^{60} B) of data (related to a single patient).

• Evidence-based medicine requires analysis of information generated from as many patients as possible. If we consider only 1/10 of humanity (roughly 10^9 people), the amount of data to be processed is in the order of thousands of Yottabytes (10^{24} B).
Data Use and Health Analytics

• Until recently, algorithms and infrastructure required to process these amounts of unstructured data were not in-par with this formidable task.

• The increase in processing power coupled with developments in AI (e.g. NLP and DL) make the vision of machine-generated personalized guidance, as well as population level analysis, feasible and real.

• IBM Watson can serve as an illustration to the new capabilities that stem from the improved hardware infrastructure and the new computational methods and algorithms.
Thank You

Questions