Big Data in Healthcare: Myth, Hype, and Hope

Woojin Kim, MD
Disclosure

• Co-founder/Shareholder – Montage Healthcare Solutions, Inc
• Consultant – Infiniti Medical, LLC
• Advisory board – Zebra Medical Vision Ltd
BIG DATA

“No other book offers such a clear and balanced tour of the opportunities and downsides of our infatuation with data.”

—WALL STREET JOURNAL

Viktor Mayer-Schönberger
and
Kenneth Cukier
Research in big data analytics working to save lives of premature babies

By Nicole Bogart
Tech Reporter  Global News

Objectives

What is Big Data?
What is Big Data?

Popular term for any collection of data sets (both structured and unstructured)
What is Big Data?

Popular term for any collection of data sets (both structured and unstructured)

so BIG and complex
What is Big Data?

Popular term for any collection of data sets (both structured and unstructured)

so BIG and complex

that it is difficult to process using traditional database processing techniques
"3D Data Management: Controlling Data Volume, Velocity, and Variety" by Doug Laney, Feb 2001
Volume

Velocity

Variety

Variability

Value

Veracity

http://en.wikipedia.org/wiki/Big_data
Google processes over 40K search queries/sec, over 3.5 billion in a single day
THE AVERAGE SMARTPHONE HAS

8 GIGABYTES
OF STORAGE

SO IT WOULD TAKE

128 SMARTPHONES
TO EQUAL

1 TERABYTE
AND

131,072 SMARTPHONES TO EQUAL 1 PETABYTE

302,080 PETABYTES OF DATA WORLDWIDE
THE AVERAGE PERSON STORES APPROXIMATELY 450MB OF DATA EACH MONTH WHICH MEANS IT WOULD TAKE YOU 198,841 YEARS TO SAVE A PETABYTE OF DATA
PETABYTE USAGE BY COMPANIES

- Bluehost: 90 petabytes
- Facebook: 180 petabytes
- Google Street View: 20 petabytes
- Mozy: 90 petabytes
- Bing: 300 petabytes
- eBay: 104 petabytes

http://mozy.com/blog/wp-uploads/2013/06/interactive_infographic/?int_cid=us-blog-txt-petabyte
What about radiology?
12 million reports from 1988 to 2015
12 million reports from 1988 to 2015
22 GB
12 million reports from 1988 to 2015
22 GB
12 million reports from 1988 to 2015
22 GB

PACS archive?
12 million reports from 1988 to 2015

22 GB

PACS archive?

250 TB
There are a million gigabytes in a petabyte.

1024 megabytes = 1 gigabyte

1024 gigabytes = 1 terabyte

1024 terabytes = 1 petabyte
Your eyes take about **400 ms** to blink once
Your eyes take about 400 ms to blink once.

Programs that use trading algorithms can make a trade based on fast-moving market data in about 0.7 ms.
Big Data

Fast Data

Actionable Data

Relevant Data

Smart Data
This is where we started to draw taller bars to please you guys.
Adapted from Gartner (http://timoelliott.com/blog/2013/02/gartnerbi-emea-2013-part-1-analytics-moves-to-the-core.html)
What happened?  
Why?  
What will happen?  
How?  

Descriptive Analytics  
Diagnostic Analytics  
Predictive Analytics  
Prescriptive Analytics

Value  
Difficulty

Foresight  
Insight  
Hindsight

Adapted from Gartner (http://timoelliott.com/blog/2013/02/gartnerbi-emea-2013-part-1-analytics-moves-to-the-core.html)
Adapted from Gartner (http://timoelliott.com/blog/2013/02/gartnerbi-emea-2013-part-1-analytics-moves-to-the-core.html)
What is Analytics 3.0?
big data @ work

Dispelling the Myths, Uncovering the Opportunities

THOMAS H. DAVENPORT

From the bestselling author of Competing on Analytics
## Three eras of analytics

<table>
<thead>
<tr>
<th></th>
<th>Analytics 1.0</th>
<th>Analytics 2.0</th>
<th>Analytics 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Type</strong></td>
<td>Small, structured</td>
<td>Large, unstructured</td>
<td>All types combined</td>
</tr>
<tr>
<td><strong>Primary Analytics Type</strong></td>
<td>Descriptive</td>
<td>Descriptive, predictive</td>
<td>Prescriptive</td>
</tr>
</tbody>
</table>

*Davenport, TH. Big Data @ Work*
## Three eras of analytics

<table>
<thead>
<tr>
<th></th>
<th>Analytics 1.0</th>
<th>Analytics 2.0</th>
<th>Analytics 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Type</strong></td>
<td>Small, structured</td>
<td>Large, unstructured</td>
<td>All types combined</td>
</tr>
<tr>
<td><strong>Primary Analytics Type</strong></td>
<td>Descriptive</td>
<td>Descriptive, predictive</td>
<td>Prescriptive</td>
</tr>
</tbody>
</table>

*Davenport, TH. Big Data @ Work*
## Three eras of analytics

<table>
<thead>
<tr>
<th></th>
<th>Analytics 1.0</th>
<th>Analytics 2.0</th>
<th>Analytics 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Type</strong></td>
<td>Small, structured</td>
<td>Large, unstructured</td>
<td>All types combined</td>
</tr>
<tr>
<td><strong>Primary Analytics Type</strong></td>
<td>Descriptive</td>
<td>Descriptive, predictive</td>
<td>Prescriptive</td>
</tr>
</tbody>
</table>
### Three eras of analytics

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Analytics 1.0</th>
<th>Analytics 2.0</th>
<th>Analytics 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary Analytics Type</strong></td>
<td>Descriptive</td>
<td>Descriptive, predictive</td>
<td>Prescriptive</td>
</tr>
<tr>
<td><strong>Data Type</strong></td>
<td>Small, structured</td>
<td>Large, unstructured</td>
<td>All types combined</td>
</tr>
<tr>
<td>Level 0 – Fragmented Point Solutions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1 – Enterprise Data Warehouse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2 – Standardized Vocabulary &amp; Patient Registries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3 – Automated Internal Reporting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 4 – Automated External Reporting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 5 – Waste &amp; Care Variability Reduction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 6 – Population Health Management and Suggestive Analytics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 7 – Clinical Risk Intervention &amp; Predictive Analytics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 8 – Personalized Medicine &amp; Prescriptive Analytics</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://www.healthcatalyst.com/healthcare-analytics-adoption-model/
Saving 1 min per driver/day

$14.5$ million/year
Saving 1 mile per driver/day

$30 million/year
What about radiology?
What about healthcare?
Predictive Analytics Seen as Future of Personalized Healthcare

While health analytics can help improve quality outcomes, increase patient satisfaction and reduce costs, the value in long-term health management is becoming more apparent. RELATED

- Press Ganey Introduces Quantifiable Patient Suffering Index
Predicting Hospital Readmissions

Year

Journal publications


https://www.healthcatalyst.com/predictive-analytics-healthcare-lessons
@woojinrad
THANK YOU
Big Data

David Piraino, M.D.
Conflict of Interest

- AGFA advisory Board
- Royalties on Stereotactic Patents
- I’m an “old guy” and don’t like change
Yes, we have it in imaging
Yes, we can discover knowledge from it
No, we don’t have to tools to discover knowledge
Big Data Definition

• data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy.

• an accumulation of data that is too large and complex for processing by traditional database management tools.
Volume Cleveland Clinic

• Definition

• The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name ‘Big Data’ itself contains a term which is related to size and hence the characteristic.

New Data Per DAY Cleveland Clinic Radiology

• 50 GB Databases
• 1.5 TB images
• 1+ TB log files
• 0.5 TB other
• ? Sensor data
• Total 3-4 TB
• That is pretty BIG

http://en.wikipedia.org/wiki/Big_data
Volume US

• Definition

• The quantity of data that is generated is very important in this context. It is the **size** of the data which determines the **value** and potential of the data under consideration and whether it can actually be considered Big Data or not. The name ‘Big Data’ itself contains a term which is related to size and hence the characteristic.

New Data Per DAY US (Projecting from Cleveland Clinic data)

• 35 TB Databases
• 1+ PB images
• 800 TB log files
• 400 TB other
• ? Sensor data
• Total 2-3 PB per day

• That is **BIG**
• How **BIG** is the rest of imaging?

PACS data per year estimate of 0.6 to 1.2 Exabytes per year as 50% of total medical data.

http://en.wikipedia.org/wiki/Big_data
Variety

• Definition

This means that the category to which Big Data belongs is also a very essential fact that needs to be **known by the data analysts**. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

Images, structured text, unstructured text, wave forms, scanned documents, video etc

• Categories
  - Image
    - CT
    - MG
    - Etc.
  - Text
    - Reports
  - Structured data
    - Breast
    - Lung screening
  - Log files
  - Video
  - Wave forms
  - Etc.

http://en.wikipedia.org/wiki/Big_data
Velocity

• Definition

The term ‘velocity’ in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

It’s getting faster

• We can generate image data and log data quickly
• Turn around times are decreasing
  – ED under 30 minutes
  – Average within a few hours
• As we provide more real time interpretation, we need more real time analysis.
• From EMR data can we predict what resources we need in the next few minutes, hours, and days
• From population data can we predict resources over months to years.
• From outcome data, personal health information and genomic data can we predict the “best” exam for each patient

http://en.wikipedia.org/wiki/Big_data
Variability

• **Definition**
  
  This is a factor which can be a **problem** for those who **analyse** the data. This refers to the **inconsistency** which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Variable and unstructured text and images

• Unstructured reports
• Structured reports
• Unstructured ordering and clinical information
• Etc

http://en.wikipedia.org/wiki/Big_data
Veracity

• Definition

• The **quality** of the data being captured can vary greatly. Accuracy of **analysis** depends on the veracity of the source data.

Radiologists are always right

• So no problem here
• Well we can work on this
• There are definitions of “truth” like surgery and pathology as well as standard outcomes that can be used

http://en.wikipedia.org/wiki/Big_data
Complexity

• Definition

  Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be **linked, connected and correlated** in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the ‘complexity’ of Big Data.

  Correlation is possible but not easy

  • Cross document and cross patient linkages are available in most cases
  • Correlation is possible but not easy
  • Prediction is much harder

http://en.wikipedia.org/wiki/Big_data
We should approach as a Data Science problem

Data analysis hierarchy

<table>
<thead>
<tr>
<th>Analytic Style</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>Summarize without further interpretation</td>
</tr>
<tr>
<td>Exploratory</td>
<td>Search for trends, correlations, or relationships to generate new ideas</td>
</tr>
<tr>
<td>Inference</td>
<td>Does the exploratory analysis hold true for other similar data but not how or why</td>
</tr>
<tr>
<td>Predictive</td>
<td>Subset of data is used to predict an outcome on a single person or unit</td>
</tr>
<tr>
<td>Causal</td>
<td>Identifies magnitude of causal effect. Usually only possible with randomized trials</td>
</tr>
<tr>
<td>Mechanistic</td>
<td>One measurement always leads to a specific behavior like gravity</td>
</tr>
</tbody>
</table>
Correlation: % CT Utilization to Mean Temperature (0.81)