Big data challenges and opportunities in healthcare: application to detecting faint signals

Dr. Greg Slabaugh
City University London
School of Informatics
Data, data, and more data

- According to IBM, *90% of the data in the world today was created in the past two years.* [1]
- According to International Data Corporation, the total amount of global data is expected to grow to 2.7 zettabytes during 2012. [2]
- The data is growing exponentially (43% growth rate) and is estimated to be 7.9 zettabytes by 2015. [3]

<table>
<thead>
<tr>
<th>Term</th>
<th>SI prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>kilobyte (KB)</td>
<td>$10^3$</td>
</tr>
<tr>
<td>megabyte (MB)</td>
<td>$10^6$</td>
</tr>
<tr>
<td>gigabyte (GB)</td>
<td>$10^9$</td>
</tr>
<tr>
<td>terabyte (TB)</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>petabyte (PB)</td>
<td>$10^{15}$</td>
</tr>
<tr>
<td>exabyte (EB)</td>
<td>$10^{18}$</td>
</tr>
<tr>
<td>zettabyte (ZB)</td>
<td>$10^{21}$</td>
</tr>
<tr>
<td>yottabyte (YB)</td>
<td>$10^{24}$</td>
</tr>
</tbody>
</table>
US Library of Congress

- The world’s largest library
  - 151.8 million items on **838 miles of bookshelves**
  - 34.5 million books and other print materials
  - 13.4 million photographs
  - 5.4 million maps
  - 6.5 million pieces of sheet music
  - 66.6 million manuscripts

⇒ By one estimate\(^4\), the entire print collection is roughly 10 petabytes of data
7.9 zettabytes (2015) is >700,000 Libraries of Congress
Sources

Images from [4]
Not limited to internet media

- Large datasets are impacting nearly all areas of business
  - Transactional data
    - Walmart: >1M transactions / hour
    - Visa: >12.5M transactions / hour[^5]
  - Networked sensors
    - Surveillance
    - Automobiles
  - Product development and manufacturing
    - Integration of data from R&D, engineering, manufacturing
    - RFID
  - Healthcare
    - Patient records
    - Diagnostic tests
    - Imaging
Why does big data matter?

• Big data is not just about storing large datasets
• Rather, it is about *leveraging* datasets
  ◦ Mining datasets to find new meaning
  ◦ Combining datasets that have never been combined before
  ◦ Making more informed decisions
  ◦ Offering new products and services

⇒ Data is a vital asset, and analytics are the key to unlocking its potential

“We don’t have better algorithms than anyone else, we just have more data.” [7]
- Peter Norvig, Director of Research, Google, spoken in 2010
Recognising the value

- Oracle
  - Integration of R (statistical programming language) into database software
- IBM
  - InfoSphere BigInsights
- Google
  - BigQuery (preview – invitation only): web service that enables interactive analysis on massive datasets – billions of rows
- Opera solutions
  - Big data analytics software based on machine learning
- Explorys
  - Explore and compare populations of patients based on medical data records
- Apache: Hadoop
Big data in healthcare

• The past
  ◦ Data is by product of providing healthcare services
  ◦ Data sets filed and never seen again; some datasets discarded; hardcopies
  ◦ Electronic health records (EHRs): primary value is that they make life easier for doctors and bring down storage costs

• The future
  ◦ Data is a central asset
  ◦ New analytics to mine data and help extract meaning
  ◦ Datasets integrated and cross-referenced – personalised medicine
  ◦ Digital records aggregated across patients – population studies

• Potential applications[8]
  ◦ Spotting unwanted drug interactions
  ◦ Identifying the most effective treatments
  ◦ Predicting onset of disease before symptoms emerge
  ◦ Analysing of disease patterns
Challenges in healthcare

- IT infrastructure
- Access to data
- Analytics
- Data privacy
- Legal issues
- Data integrity

Slide adapted from [9]
Opportunities in healthcare

- $165B Clinical
- $9B Public health
- $108B R&D
- $5B Business model
- $47B Accounts

Slide adapted from [9]
A case study: colorectal disease

- Colorectal cancer **second** most prevalent cancer in Western countries[^10]
- *940,000* cases occur annually
- *655,000* deaths annually
- **if** detected early, *90%* of patients live at least ten years
Colorectal cancer

- Pathophysiology
  - Adenomatous polyps
    - Benign tumours of a glandular organ (colon is a mucosal organ)
    - Common, particularly in patients 50+
    - Greater than 10 mm: higher likelihood of developing into cancer
  - Cancer
    - Can invade below colon surface and spread to other organs

⇒ Progression is slow (5+ years to polyp, 5+ more to cancer)
Screening methodologies

- Faecal occult blood test (FOBT)
- Optical colonoscopy (OC)

+ Effectiveness, cost, can remove polyps during procedure
- Invasive, sedation, risk of perforation, occlusions, physical limitations, patient compliance
Screening methodologies

• CT Colonography (CTC), or “Virtual Colonoscopy”
  ◦ Examination of the colon using CT imaging
  ◦ Patient given laxatives to clear the colon
  ◦ Patient consumes a faecal tagging solution designed to coat any residual stools or liquid
  ◦ Thin tube inserted into rectum to inflate colon with gas (CO₂)
  ◦ Images taken with patient in prone and supine positions
  ◦ Images are analysed by radiologist for colorectal lesions
Data overload

• Optical colonoscopy
  ◦ Approx. 20 minutes of HD video per patient (often not recorded)
  ◦ 30,000,000 procedures performed annually worldwide
  ◦ Roughly 15 petabytes of data per year (and growing exponentially)

• CT colonography
  ◦ Each CT series (prone, supine) has roughly 500 images of size 512x512
  ◦ 1,000,000 procedures performed annually worldwide
  ◦ Roughly 0.5 petabytes of data per year (and growing exponentially)

⇒ All this data must be reviewed by a physician (gastroenterologist or radiologist)
CT Colonography images

- This patient has a polyp in their colon. Did you see it?

⇒ Polyps can be very subtle and difficult to detect, even for expert radiologists
Computer-aided detection (CAD)

- CAD consists of image processing and pattern recognition algorithms designed to detect polyps that may be of interest to a physician.
- CAD draws the radiologist’s attention to regions that may have otherwise been overlooked.
- “Spell-checker” for medical images.
- Characterised by:
  - Sensitivity (percentage of polyps found)
  - Number of false positives
- CAD is designed to be complementary – it is *not* a replacement for physician.
How does CAD work?

- Image pre-processing
- Organ segmentation
- Candidate generation
- Feature calculation
- Classifier

Results

Performance

- Largest ever colon CAD clinical study (3000+ patients)
  - Dr. Perry Pickhardt (U. Wisconsin)
  - Published in *Radiology* 2010\[^{12}\]; picked up in radiology press
  - 4.7 false positives per series

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity ≥ 6 mm</th>
<th>Sensitivity ≥ 10 mm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>By Polyp</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All polyps</td>
<td>90.1% (547/607)</td>
<td>95.9% (168/175)</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>94.6% (265/280)</td>
<td>96.8% (119/123)</td>
</tr>
<tr>
<td>Adv. neoplasms</td>
<td>97.0% (128/132)</td>
<td>96.8% (119/123)</td>
</tr>
<tr>
<td>Cancer</td>
<td>100% (13/13)</td>
<td>100% (13/13)</td>
</tr>
<tr>
<td>Sessile</td>
<td>90.5% (370/409)</td>
<td>100% (52/52)</td>
</tr>
<tr>
<td>Pedunculated</td>
<td>93.8% (105/113)</td>
<td>97.2% (70/72)</td>
</tr>
<tr>
<td>Flat</td>
<td>77.8% (49/63)</td>
<td>79.3% (23/29)</td>
</tr>
<tr>
<td><strong>By Patient</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All polyps</td>
<td>93.8% (350/373)</td>
<td>95.5% (137/142)</td>
</tr>
</tbody>
</table>

⇒ CAD identified 15 polyps that were missed by *expert* radiologists
Advanced analytics

Manifold learning\cite{13}
Advanced analytics

Population regression\textsuperscript{[14]}
Multi-modality, multi-scale and heterogeneous
Other “dimensions” to consider
References

[1] IBM quote, microscope.co.uk
[2] International Data Corporation 2012 prediction, IDC website
[3] CenturyLink 2015 prediction, ReadWriteWeb website
[7] Peter Norvig quote, CNET website
[8] Economist report on Big Data
[9] Wipro infographic