Big Data in medicine – today and tomorrow

Thomas D. Szucs

July 2016
Paul Gauguin (1897)

Where do we come from?
Who are we?
Where are we going?
Three questions

- Where do we come from? – Evolution of the concept of data
- Who are we? – The current meaning of Big Data in medicine
- Where are we going? – Challenges for the future
Three questions

- Where do we come from? – Evolution of the concept of data
- Who are we? – The current meaning of Big Data in medicine
- Where are we going? – Challenges for the future
What is data?

Data is a “collection of symbols that have been combined for processing purposes that represent information based on known or implicit understandings (i.e. information about situations and events)” (Gabler).
Forecast: volume of data generated around the world annually

Volume of data in exabytes

- 2005: 130 exabytes
- 2010: 1.227 exabytes
- 2012: 2.837 exabytes
- 2015: 8.591 exabytes
- 2020: 40.026 exabytes
# How much data is there now?

<table>
<thead>
<tr>
<th>Level</th>
<th>Binary/decimal</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB</td>
<td>$10^9$</td>
<td>0.8 GB human genome</td>
</tr>
<tr>
<td>TB</td>
<td>$10^{12}$</td>
<td>1 TB printed paper from 50,000 trees</td>
</tr>
<tr>
<td>PB</td>
<td>$10^{15}$</td>
<td>1 PB memory for the film Avatar and the game “World of Warcraft”</td>
</tr>
<tr>
<td>EB</td>
<td>$10^{18}$</td>
<td>5 EB every word ever spoken (text)</td>
</tr>
<tr>
<td>ZB</td>
<td>$10^{21}$</td>
<td>?</td>
</tr>
<tr>
<td>YB</td>
<td>$10^{24}$</td>
<td>Has not yet occurred in practice (expected in 2025)</td>
</tr>
</tbody>
</table>
How large is 1 zettabyte?

A pile of 295 billion USB sticks with 5 GB each

Wrapped around the earth 1.157 times
Data creation

We create as much data in 10 minutes as the first 10,000 generations of humans generated
Data waste as a future data treasure trove

650 terabytes
## Program code

<table>
<thead>
<tr>
<th>Program system</th>
<th>Lines of code (MLOC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space Shuttle</td>
<td>3</td>
</tr>
<tr>
<td>Google Android</td>
<td>11</td>
</tr>
<tr>
<td>Windows Vista</td>
<td>50</td>
</tr>
<tr>
<td>Mac OS X</td>
<td>86</td>
</tr>
<tr>
<td>Luxury vehicle</td>
<td>100</td>
</tr>
<tr>
<td>Airbus 380</td>
<td>1,000</td>
</tr>
<tr>
<td>Total of all iPhone applications</td>
<td>35,000</td>
</tr>
</tbody>
</table>

10 MLOC = stack of paper 15 meters high

MLOC = million lines of code
Distribution of the digital universe
Intelligibility of Big Data

“No human mind is capable of grasping in its entirety the meaning of any considerable quantity of data.”

Sir Ronald A Fisher
How is Big Data used?

- Multiple data sources
- Prediction and performance models
- Organisational transformation
Impact of quantitative analyses on decisions in the real world
Hollywood uses sophisticated analyses as well

“Script to $”
The Wisdom of Crowds


Francis Galton (test, 1906)
Synthesis approach: www.kaggle.com
The Dunnhumby Shopper Challenge at Kaggle
Dunnhumby Shopper Challenge: Task

Data

CUST #14345
17/06/2011
Total ...... £90

Data

CUST #14345
20/07/2011
Total ...... £84

Prediction

CUST #14345
??/??/20??
Total ...... £??
## Digital and social media dominate in 3 areas

<table>
<thead>
<tr>
<th></th>
<th>Examples</th>
</tr>
</thead>
</table>
| **Digital healthcare channel** | - Email consultations  
|                           | - Online access to laboratory data  
|                           | - Access to imaging                         |
| **Digital customer innovation** | - PatientsLikeMe                     |
| **Digital initiatives for social change** | - Let’s move campaign                              |
Three questions

- Where do we come from? – Evolution of the concept of data
- Who are we? – The current meaning of Big Data in medicine
- Where are we going? – Challenges for the future
Human body as a source of data

- Human genome: 3 GB
- 3D MRI: 150 MB
- X-ray: 30 MB
- Mammography: 120 MB
- 3D CT: 1 GB
- Average hospital: 665 TB

Annual increase in required storage for imaging: 20-40%

80% of all data is unstructured (images, videos, email)

23
Healthcare data overload (per lifetime)

- **60% exogenous data (environment, socio-economic, behaviour)**
  - 1,100 TB

- **30% genomic data**
  - 6 TB

- **10% clinical data**
  - 0.4 TB
Objectives of Big Data analysis in healthcare

- Electronic health files
- Behaviour
- Genomic
- Public health

Evidence + findings

Better outcomes
Better cost effectiveness
Four major data pools in healthcare

- Pharmaceutical research data
- Clinical data
- Patient behaviour
- Billing and cost data
Daten@Helsana

- **Policyholders**: 1.9 million
- **Benefits paid**: approx. CHF 5.2 bn. per year, approx. CHF 2.5 mn. per work hour
- **Customer calls**: approx. 2 mn. per year
- **Internet data traffic**: approx. 2 TB per month
- **Documents processed**: approx. 13.2 mn. per year
- **Digitalised invoices**: approx. 6 mn. per year, approx. 27,000 per day
- **Online transactions**: 1.7 mn. per day
- **Printing**: 6.2 tons of paper/week
- **Workstations**: 4,600 PCs / laptops
- **Saved data**: 320 TB; if saved on CDs, it would be equivalent to a 4.6-km high tower
- **Size of DWH**: approx. 30 TB
- **Largest table in the DWH**: 14 billion lines
Example: PatientsLikeMe

- 150,000 satisfied customers
- 75% understand their problems better
- 18% of epileptics had fewer emergency room visits
- 22% of patients with mood disorders were hospitalised less often
- HIV patients: 41% reduced their risky behaviours
Example: Proteus Digital Health®

“Adherence”
Self-measurement ("quantified self")
Sensors and healthcare research

Asthma hotspots and participant enrollment

- Participants enrolled/zip code
  - 0 - 5
  - 6 - 10
  - 11 - 20
- Primary roads

- Avg. wind direction
  - SW (210 degrees)
Recognising patterns in the 19th century: cholera in London
Recognising patterns today: predicting which patients will be unable to pay

The actual phenomenon (real historical data)

How Traditional stats sees it

How Machine Learning sees it

Health insurers can use geographic information systems (GIS)

Before

After
Identification of undesired effects in the web: Tamoxifen

Blogs

I saw the liver specialist – was able to get in earlier and he said my fatty liver is probably from taking Tamoxifen. I am still very scared, because Tamoxifen can cause uterine cancer. I had a stroke recently, after finishing 5 yrs of Tamoxifen. This sealed my decision. I was one of the few that developed a pulmonary embolism while on Tamoxifen – lucky me.
Incidence of the flu in Switzerland

Blue: Google
Orange: FOPH
Digital epidemiology using social media and information (2010 cholera outbreak in Haiti)
Diversification in healthcare

THE WALL STREET JOURNAL
October 22, 2007, 6:47 PM ET
Zagat Gets Into Doctor Ratings
By Venessa Fuhrman

If getting an appointment with Dr. Smith “can take weeks” but his great bedside manner is “worth the wait,” wouldn’t you like to know?

Diners are well-acquainted with the Zagat Survey approach to rating restaurants, hotels and bars. But now the consumer guide and rating company is teaming up with WellPoint to tackle the world of physicians.

Starting in January, members in some of WellPoint’s Blue Cross and Blue Shield plans will be able to go online and review and rate their experiences with doctors. The health insurer, which has 35 million members nationwide, first plans to make the rating tool available to 1 million, though where exactly remains to be disclosed.

Patients using the online tool will get to apply the same 30-point scale that has helped make or break plenty of restaurant reputations. In place of familiar food, décor, service and cost categories, though, doctors will be rated on trust, communication, availability and cost. WellPoint says each entry will display contact information, the 30-point scale ratings in each category, plus the percentage of members who recommend the physician.
“Customers” rate the healthcare system in the UK as well
Crowdfunding
Three questions

- Where do we come from? – Evolution of the concept of data
- Who are we? – The current meaning of Big Data in medicine
- Where are we going? – Challenges for the future
Technology convergence

Medicine

Biology

IT
Dimensions

Size of the genome of a person (\(3.27 \times 10^9\) bp)

Size of a gene in a person (30,000 bp)

1 cm\(^3\) transcriptome of a tumour
Physical biobanks
Digital biobanks

National Super Computer Center
Guangzhou, China

33.8 petaflops/s computing power
12.4 petabytes of storage

(Oak Ridge Laboratory, Titan Supercomputer: 17 PFLOPS/s, 20-30 PBytes)
Limits of IT: Clock speed and cooling
DNA – the miracle molecule
1 g DNA can store at least 5.5 petabytes (= 700 TByte)*

George Church stored:

- Martin Luther King, Jr.’s “I Have a Dream” speech
- A photo
- A copy of the paper “Double Helix” (1953) by Francis Crick and James Watson
- All 154 Shakespeare sonnets

All data was read back with a precision of 99.99%!

*3.7 million DVDs
Important objective

Big Data

Smart Data
“It requires a very unusual mind to undertake the analysis of the obvious.”

Alfred North Whitehead
British mathematician and philosopher
1861-1947
Data Scientist: The Sexiest Job of the 21st Century
Does Big Data lead to higher performance?

- Yes
- Companies in the upper third of those that are characterised as “data driven”
  - 5% higher productivity
  - 6% higher profitability
The flip side

- Digital dementia (Spitzer)
- Social media exhaustion
- Addiction and dependence
- Validation of data
- Control: Drive to generate data
- Some data manipulation unknown (e.g. LIBOR)
Digital dementia
Digital (R)-EVOLUTION
Manfred Spitzer

**MANFRED SPITZER**

**DIGITALE DEMENZ**

Wie wir uns und unsere Kinder um den Verstand bringen

DROEMER
Are gamers better surgeons?

Review of 142 studies

Yes!

Video games with psycho-motor components = skills relevant for laparoscopic surgery
Data protection
YOU’RE NOT THE CUSTOMER, YOU ARE THE PRODUCT
“We take the information patients like you share about your experience with the disease and sell it to our partners (i.e. companies that are developing or selling products to patients).

These products may include drugs, devices, equipment, insurance and medical services ...

You should expect that every piece of information you submit (even if it is not commonly displayed) may be shared.”
You think hackers are evil, meet the brokers

- Data surveillance industry
  - Acxiom, Epsilon, Datalogix, Reed Elsevier, BlueKai, Sopeo, Flurry
- Annual revenue: $156 bn. p.a. (= 2x the NSA budget)
- Data brokers obtain information from: ISPs, credit card companies, mobile phone companies, banks, pharmacies, motor vehicles, retail shops, online activities
- Acxiom: 23,000 servers; 50 trillion transactions
- Acxiom goals: predictive targeting, behavioural targeting, premium proprietary behavioural insights
Hospital privacy?

Hospital caregivers made 18 comments deemed to compromise a patient’s confidentiality on 13 out of 113 lift journeys.

Privacy of patients’ information in hospital lifts: observational study
Simone N Vigod, Chaim M Bell, John M A Bohnen

BMJ VOLUME 327 1 NOVEMBER 2003 bmj.com

Vigod SN 2001; Ubel PA 1995; Hasman A 1997
Goodman’s Law

“The more data that is produced and saved, the more organised crime is willing to consume it.”

MARC GOODMAN
FUTURE CRIMES
Inside the Digital Underground and the Battle for Our Connected World

NEW YORK TIMES BESTSELLER

“"The more data that is produced and saved, the more organised crime is willing to consume it.”

MARC GOODMAN
FUTURE CRIMES
Inside the Digital Underground and the Battle for Our Connected World

NEW YORK TIMES BESTSELLER
Facebook attacks

- 600,000 accounts per day
- 1 account every 14 msec
- Purpose: identity theft, criminal impersonation, tax fraud, insurance fraud etc.
IN SCREEN WE TRUST!
Data market – some prices

- 4,000 Twitter followers: $5
- 100 Facebook fans: $1,500
- 1 million Instagram friends: $3,700
Hacking the robot

Cave: Internet of Things
Thought experiment
A drug does not necessarily have to target somatic cells only.

Personalised (targeted) treatment

Personalised Bio-weapon
Data sharing in the healthcare system: block chain

Traditional vs Block chain
Data sharing in the Estonian healthcare system = gold standard
Our big opportunity

Elite controllers
Super heroes
Exceptional responders
Mining the genomes of exceptional responders

David K. Chang¹⁻⁵, Sean M. Grimmond¹,⁶, T. R. Jeffry Evans¹ and Andrew V. Biankin¹⁻⁵

Abstract | The National Cancer Institute of the United States recently announced a major new initiative in understanding the genomes or, more broadly, the molecular phenotypes of exceptional responders. What can we expect to learn from exceptional responders? What are the potential benefits, and how do we approach studying them?
Analysis of 589,306 genomes: identified individuals are resistant to severe paediatric illnesses
Mendelian disorders
Web searches and pancreatic cancer
The imagination knows no boundaries!

“There are more things in heaven and earth than are dreamt of in your philosophy.”

- William Shakespeare, Hamlet
Predictions

- “There is no reason anyone would want a computer in their home.”
  *(Ken Olsen, President of DEC, 1977)*

- “A rocket will never be able to leave the earth’s atmosphere.”
  *(New York Times, 1936)*
The train has left the station