Big Data in Healthcare: Possibilities & Challenges

Datuk Prof Dr Awang Bulgiba Awang Mahmud
MBBS MPH MAppStat PhD FFPH FPHMM FAMM FASc

About Big Data

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

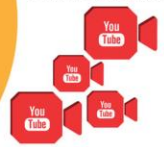
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

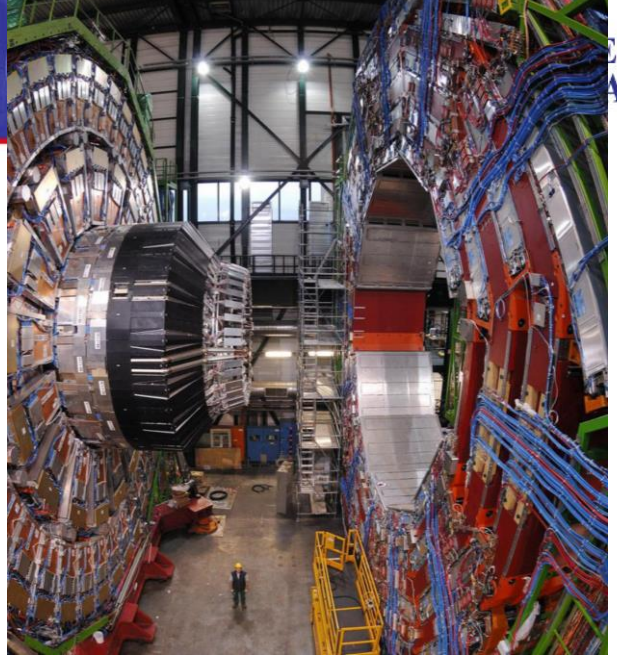
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

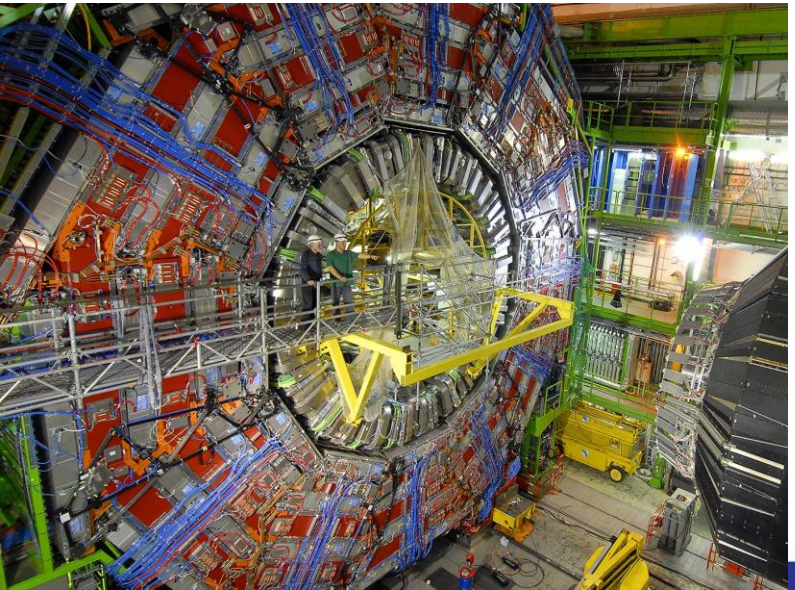
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

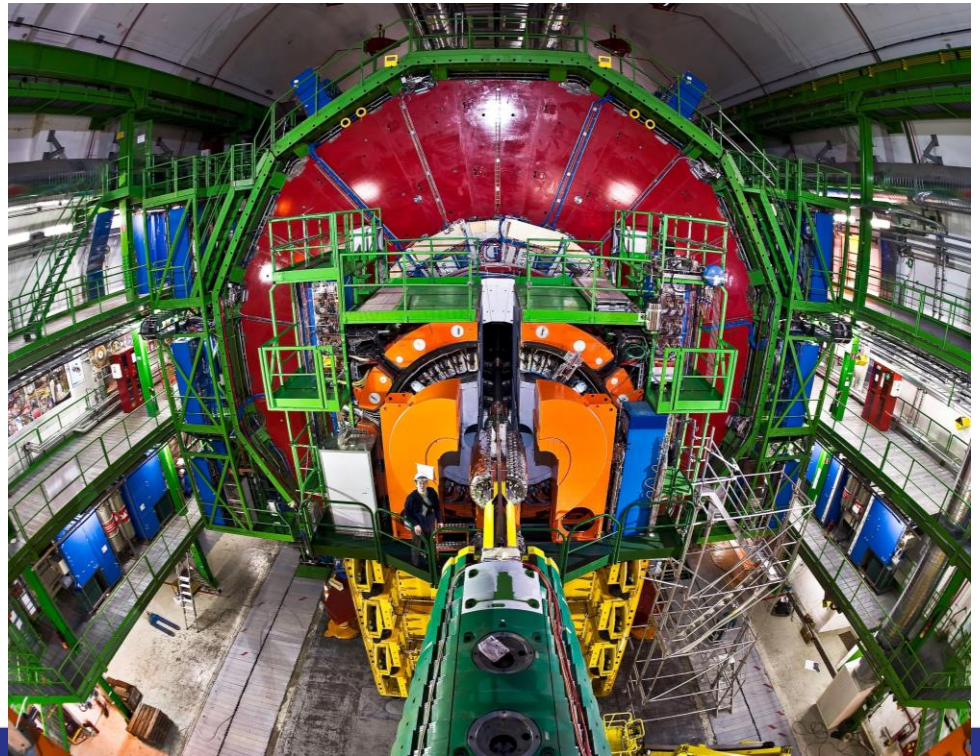
IBM

About big data

- ✓ Big data is everywhere
- ✓ Retail stores which track shoppers
- ✓ Online stores which track shoppers
- ✓ Banks which track transactions
- ✓ Immigration data
- ✓ Mobile phone data
- ✓ Disease data
- ✓ Search engines data
- ✓ Google Maps data
- ✓ E hailing data
- ✓ Social media data
- ✓ Astronomical data
- ✓ CERN data

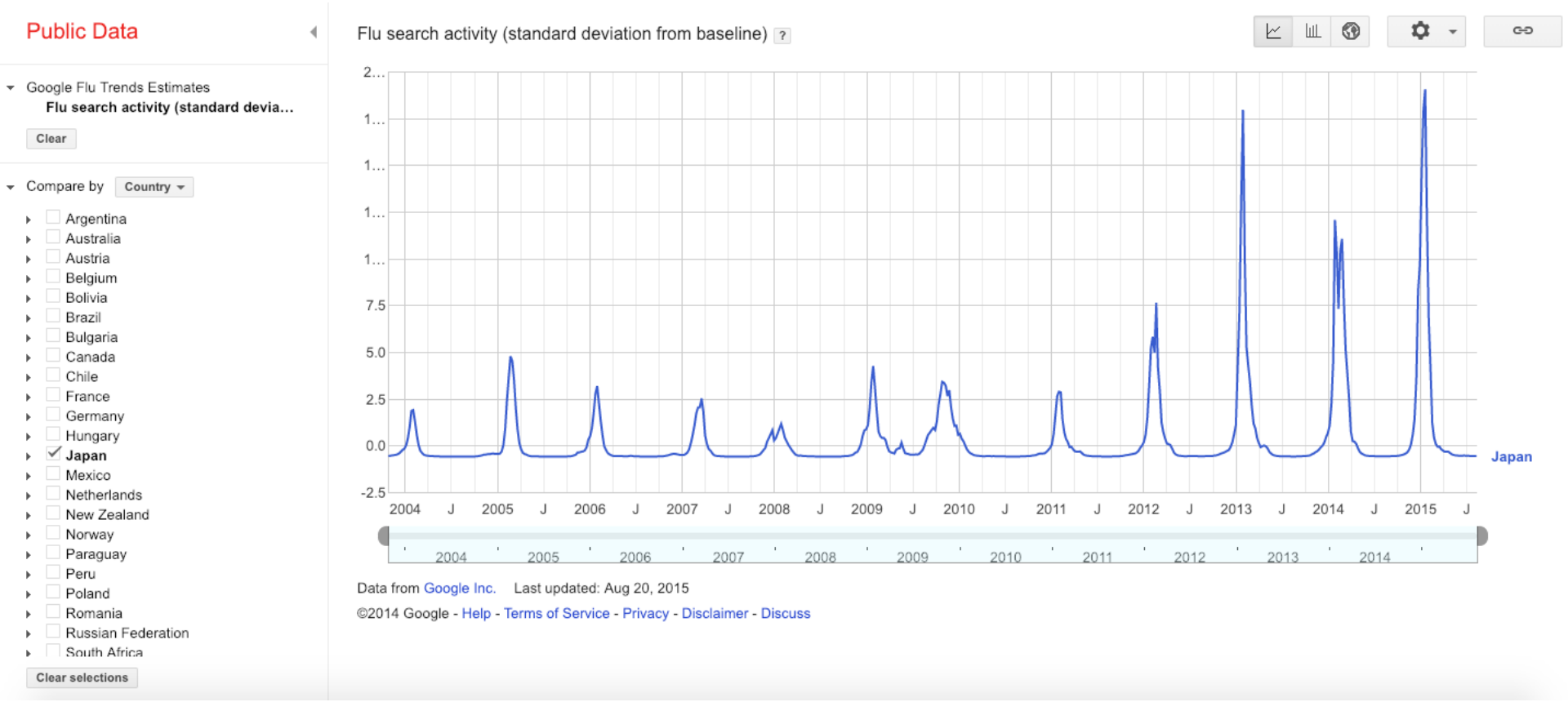






ERSITY
ALAYA

Google Flu (Japan)



Google Flu

- An old resource (discontinued in 2015)
- This resource predicts outbreaks of flu by counting the number of flu-related Internet searches using the word “flu” that occur
- Was done for some countries

How can we use big data in healthcare?

Improve value of healthcare

New York's Medicaid Redesign Team

The state of New York launched a nine-year programme to improve health outcomes, the efficiency of service delivery and value for money for the state's six million Medicare beneficiaries in 2011.[35]

The programme uses big data analytics and linking of databases across primary care, acute care and community health settings to improve the value of healthcare and shift from a traditional payment system to a values based system. (Values based means payment is provided for desired outcomes rather than for standard service delivery).

The Medicaid Redesign Team (MRT) utilised an agile analytics platform using HCI3 grouper technology that integrates a massive dataset across all care settings, covering three years of activity for millions of patients. This allows the New York State to understand the quality and total cost of care and provides a single analytics infrastructure enabling consistent reporting and facilitating sophisticated benchmarking for all.

As a result of the changes made under the MRT strategy, $17 billion in future expenditures were avoided over a five year period, of which nearly half was reinvested into a reform incentive payment for care providers.

Prioritise high value healthcare

The Medicare Benefits Schedule (MBS) Review

The MBS Review – Interim Report[40] reports on the increasing number of Medicare pathology tests annually (up from 46 per cent of the population in 2003–04 to 54 per cent in 2013–14). It further reports that the proportion of the population that has an annual diagnostic imaging service has also increased to 37 per cent from 30 per cent in 2003–04, with the number of services per capita increasing from 2.2 to 2.6 during the same time period.

http://www.health.gov.au/internet/main/publishing.nsf/Content/mbsr-interim-report

Increasing rates of testing, together with the increasingly sensitive nature of tests is concerning because of the potential for over-diagnosis and over treatment. For example, overuse of routine imaging without a strong evidence base may lead to further unnecessary investigations and may itself cause unnecessary exposure to radiation.[41]

Pharmaceutical R & D

- Electronic Health Records improve patient selection for clinical trials
- Linked multiple EHRs help create cohort studies
- Real time analysis of clinical trials and patient records
- Pharmaceutical R&D can thus be accelerated

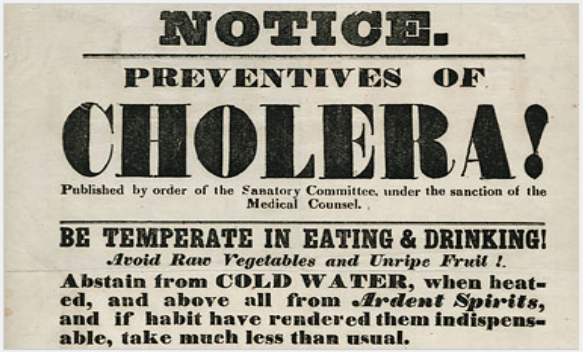
Personalised & targeted healthcare

- Dr can use patient history + behaviours + genetic data to identify personalized treatment and drugs
- Smart devices & social media can help patient record behavioural data which can be analysed in realtime to predict prognosis

Improve population health measures

“Old” approach

London's 1854 Cholera Outbreak:
Data Mapping Halts an Epidemic



NOTICE.

PREVENTIVES OF

CHOLERA!

Published by order of the Sanatory Committee, under the sanction of the Medical Counsel.

BE TEMPERATE IN EATING & DRINKING!

Avoid Raw Vegetables and Unripe Fruit !.

Abstain from COLD WATER, when heated, and above all from *Ardent Spirits*, and if habit have rendered them indispensable, take much less than usual.

Victorian medical advice was frequently off the mark by mod..

In 1854, a Cholera outbreak swept through the Soho neighborhood of London.

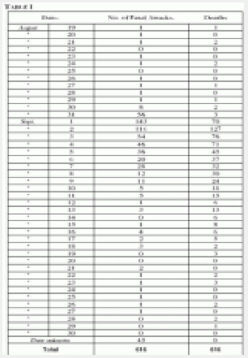
616 people died.

Physician John Snow was skeptical of existing theories of disease transmission, which often blamed "miasmas," or bad air. The germ theory of disease circulation had not yet been outlined.

https://public.tableau.com/en-us/s/gallery/mapping-1854-cholera-outbreak

London's 1854 Cholera Outbreak:
Data Mapping Halts an Epidemic

An Outbreak Begins	Collecting the Data	Mapping the Results	Snow's Analysis: Focus on Broad St.	Ending an Epidemic



This is Snow's original table showing the chronology of deaths and their total. He notes that not all deaths have recorded addresses; thus, his data set is partially incomplete.

Dr. Snow canvassed the neighborhood. He collected the addresses of those who died, noted the number of deaths at each location, and tabulated the results.

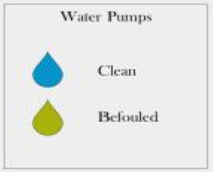
https://public.tableau.com/en-us/s/gallery/mapping-1854-cholera-outbreak

London's 1854 Cholera Outbreak:
Data Mapping Halts an Epidemic

Then, Snow mapped these data points onto a map of the Soho neighborhood.

The results were startling.

..

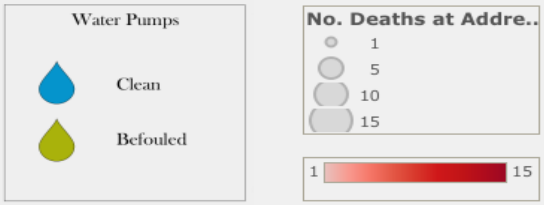


https://public.tableau.com/en-us/s/gallery/mapping-1854-cholera-outbreak

London's 1854 Cholera Outbreak:
Data Mapping Halts an Epidemic

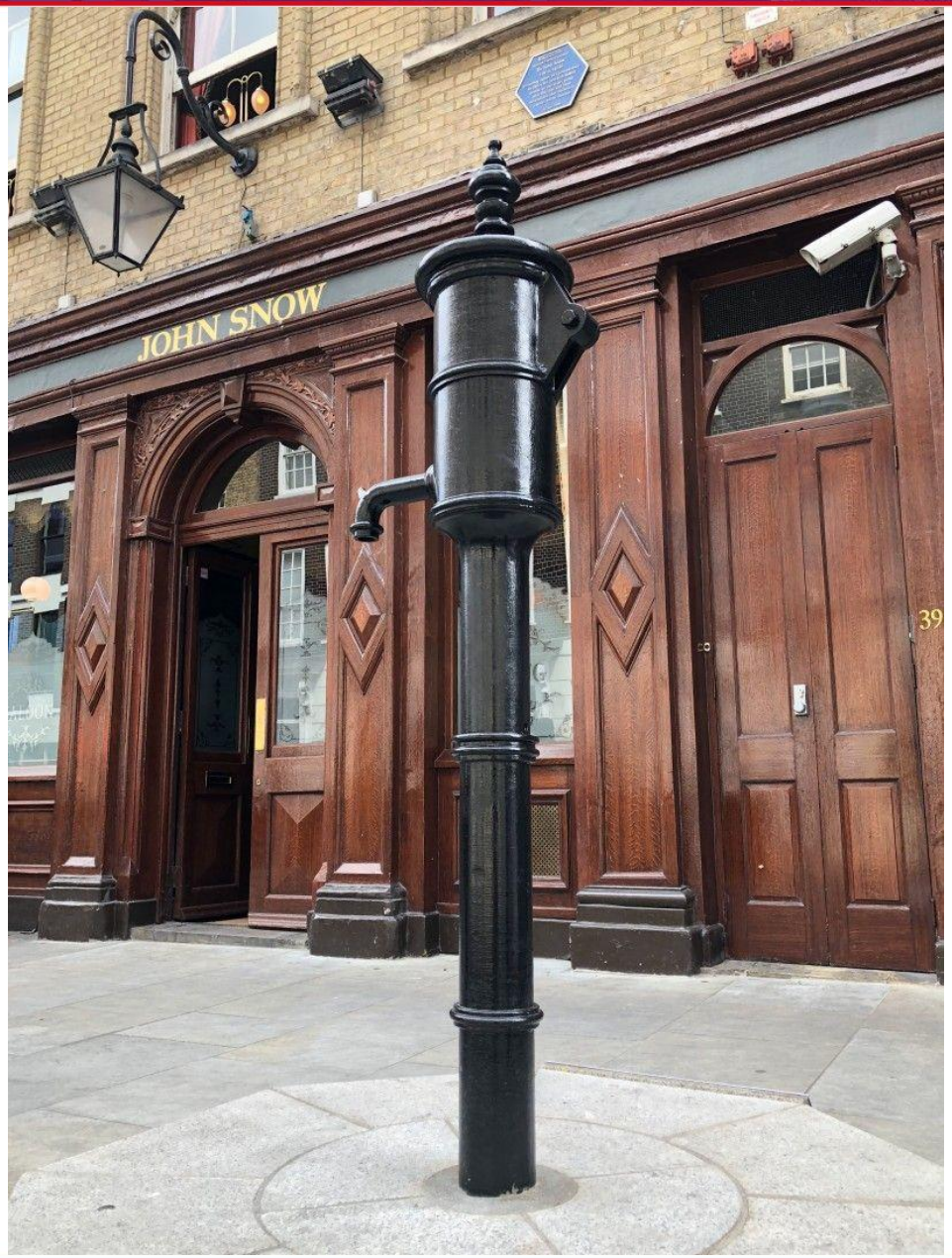
He noted that the bulk of deaths were concentrated in an area that generally used the same municipal water pump: Broad Street.

Snow realized that the Broad Street water pump--and poor water quality--was likely the source of the outbreak. ..



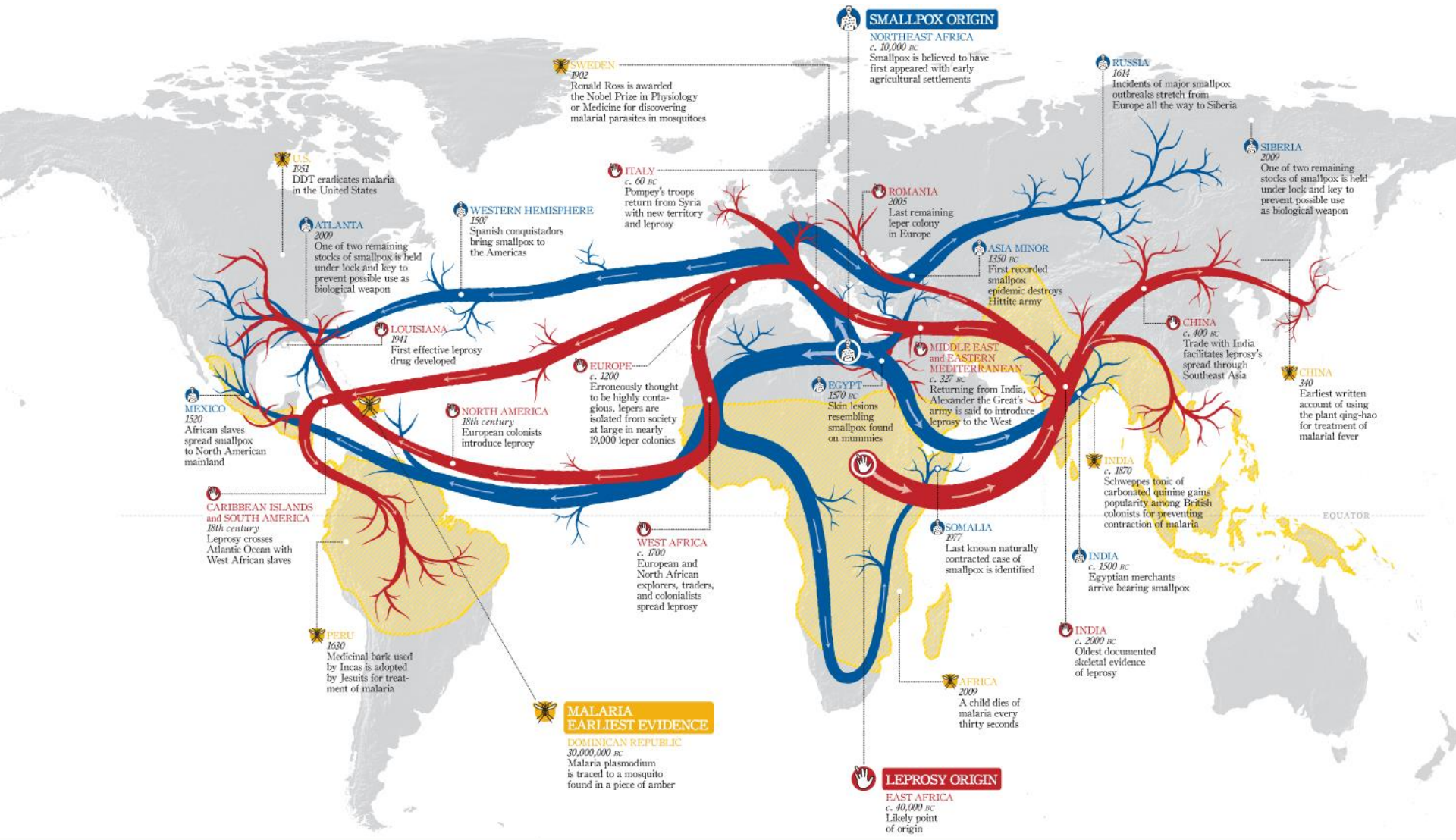
https://public.tableau.com/en-us/s/gallery/mapping-1854-cholera-outbreak

Broad Street Pump



However …

Infectious diseases spread faster than ever before



Let us take a look at how quickly

… H1N1 spread throughout the world in 2009

“New” approach

Heat-map Visualisation

Published on Apr 20, 2012

This is a heat-map visualization of the prevalence of flu in New York City, as observed through public Twitter data. The more red an area is, the more people are afflicted by flu at that location. Emergent aggregate patterns are easily discovered with a second-by-second resolution.

In order to harness big data we need to ensure …

- that healthcare data systems are not fragmented & data collection is not piecemeal.
- we have a developed infrastructure & talent to deal with big data as these take time to develop.
- issues of privacy and security are dealt with in a holistic manner and that our laws and our policies do not impede our use of big data.
- we act swiftly to maximise the benefits that we will gain from big data in healthcare.

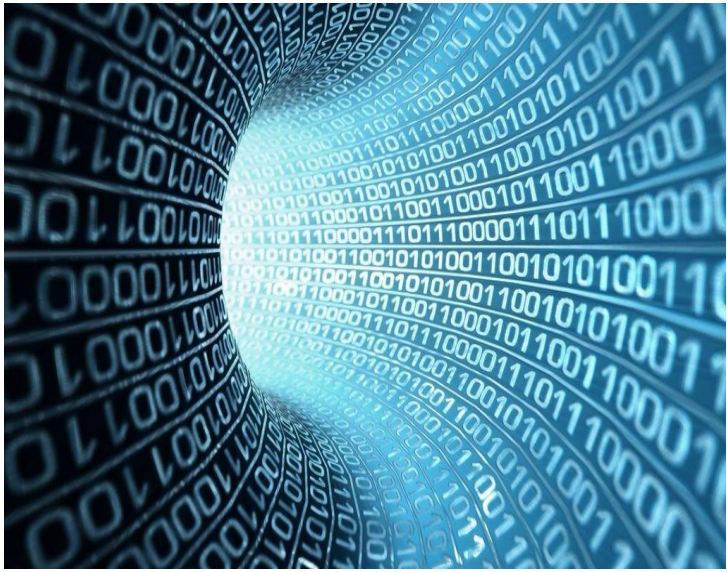
Challenges with Big Data

Volume issue with Big Data

- Volume is so large leading to very high computing & network requirements
- Rate of increase in volume is staggering
- Projecting storage requirements is difficult as every new feature in an application introduces new space requirements
- Need to backup data also means added hardware & storage requirements
- Formats of data storage change

Velocity issue with Big Data

- Data is streamed at high speeds
- Requirement for always-on connection
- Requirement for very high bandwidth



Variety issue with Big Data

- Data may not be structured in the usual format we are used to e.g. CSV, ASCII etc with proper fields and data
- Data could be in Entity-Attribute-Value (EAV) format
- Data could be in image (JPG, PNG, RAW etc.) or video format in various extensions (AVI, MPG4, MOV etc.)
- Making sense of these data will be an issue

Veracity issue with Big Data

- Much of the data may not be accurate
- Some of the data could be fake (by accident or intention)
- Some of the data are just noise
- Some data are just repetition of other people's data and not original (e.g. people forwarding a message)

How do we deal with these issues?

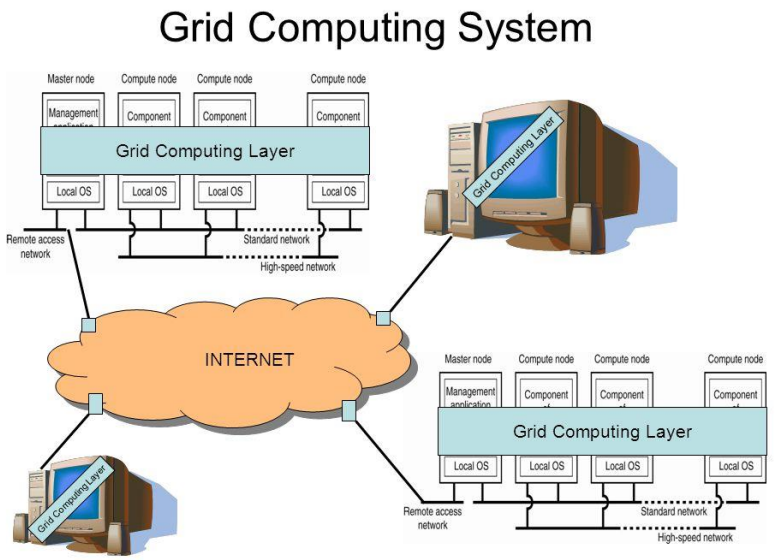
Big data management

- We cannot apply the normal approaches of managing data to big data
- Big data can be really huge (Volume)
- Big data can be streamed all the time (Velocity)
- Big data is of various types & can change (Variety)
- The authenticity and accuracy of big data can be difficult to verify (Veracity)

We need to use different approaches

Dealing with Large Volume

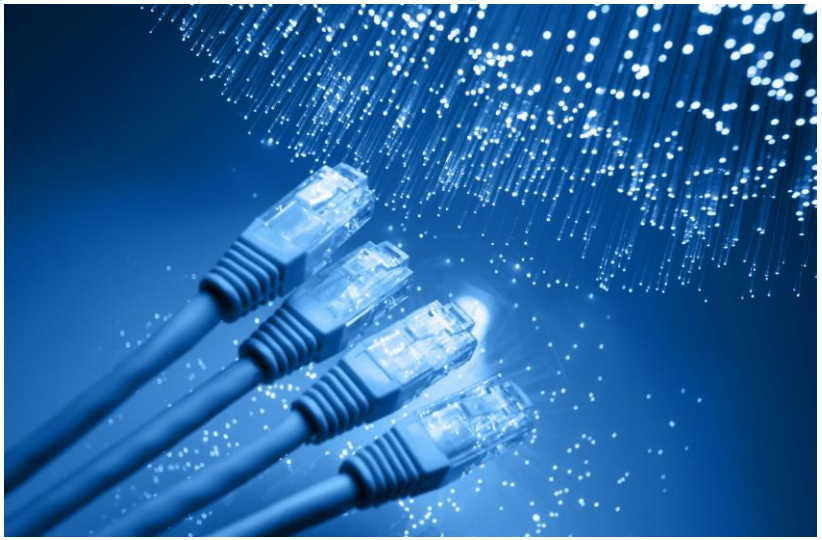
- Very high volume means usual storage & computing solutions cannot be used
- This has led to alternative solutions:
  - HPCs (High Performance Computers) which combine many Central Processing Units (CPUs) & Graphical Processing Units (GPUs)
  - Grid or distributed computing
  - Pay-per-use server computing e.g. Amazon web service
  - Quantum computing
  - Can DNA storage be a possibility?



amazon web services™

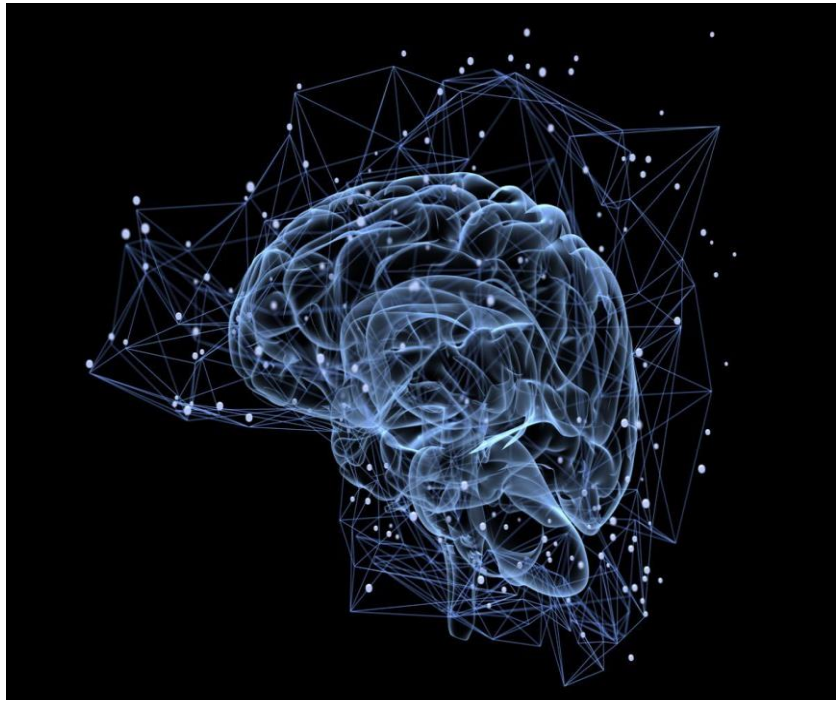
Dealing with High Velocity

- This requires better connections
- Research is ongoing
  - 4G makes it possible for 1 Gbps mobile speeds (although the best I have managed with my 4G connection is 140 Mbps)
  - 5G will make it possible for 20 Gbps mobile speeds & low latency
  - Fibre bandwidth must step up to 1000 Gbps (1 Tbps) or more
  - Freeing up wireless frequencies (why do you think the government is introducing MyTV?)
  - Tbps will be required for Internet gateways of the future within 5 years



Dealing with Great Variety

- A lot of AI is now being used to deal with the great variety of data
- AI approaches
  - Neural networks, heuristic approaches
  - Fuzzy logic
  - Genetic algorithms
  - Hybrid approaches
  - Data mining approaches
- AI is already being used and will be commonplace with data management



Dealing with Uncertain Veracity

- A lot of AI is now also being used to deal with the uncertain veracity of data
- AI approaches
  - Neural networks, heuristic approaches
  - Fuzzy logic
  - Genetic algorithms
  - Hybrid approaches
  - Data mining approaches
- AI will be a must to sift through data

A word of caution

1. “Over-fitting” the results
2. Ever-changing data & results due to continuous streaming
3. Failure to recognise changes fast enough
4. Failure to understand the approach
5. Failure to understand “black box”
6. Over-reliance on AI where data does not exist
7. Ethical concerns

We need to act

But are we prepared for big data?

- Big data is already here
  - 20 years ago, Hospital Selayang became Malaysia's first paperless hospital
  - It was talking about accumulating perhaps 3-6 TB ($6*10^4$ bytes) of data after a few years
- Now a pendrive can hold 2 TB
- Our own hospital UMMC is now talking about PB (10^{15} bytes or 1000 TB) of data per year

In summary we need …

- ✓Lots more storage space with redundancy
- ✓Much faster network connections
- ✓Faster computers to process data
- ✓Better algorithms to analyse data
- ✓But most of all we need better trained people to make use of all these big data
- ✓Better trained data scientists
- ✓Better trained health informaticians
- ✓Better informed public

And we need to make sure ..

1. Healthcare systems are not fragmented
2. Data collection is not piecemeal
3. Develop talent and infrastructure
4. Issues of privacy & security are dealt with in a holistic manner
5. Laws and policies do not impede use of big data


Thank You