Challenges in Managing Health-Linked Big Data

Moderator

Hassan Khachfe, Lebanese International University, Lebanon

Panelists

Sung-Bae Cho, Yonsei University, Korea

Sandjai Bhulai, VU University Amsterdam, The Netherlands

Thomas Klemas, MIT Lincoln Laboratory, USA

Alexander Ponomarenko, Nizhny Novgorod, Russia

- Big Data is the "new oil"
  - $100B annually in sensitive devices
  - One firm may pay more than $1.5M/yr for storage
- Data is increasing exponentially ➔ Need for more potent storage media
  - 5¼ disks ➔ 750 KB
  - 3½ disks ➔ 1 MB
  - CD's
  - DVD's
  - BR's
  - HD, shares drives, clouds, etc.

Data is generated from all types of applications

- simple
  - Office files
  - Forms
  - news feeds
  - Etc.

- Complex
  - Media sharing
  - social networks
  - Space telescopes
  - etc.

In Healthcare, data needs special treatment

- Privacy
  - Patients' records
  - Insurance Claims
  - Etc.
- security / financial
  - Government files
  - Research repositories
  - Etc.
- Complexity
  - Imaging files
  - Telemedicine
  - Simulations

- Opportunities
  - Economic growth
  - Technology enhancement
  - Technology transfer
  - Etc.

- Challenges
  - Technology
  - Cost
  - Privacy
  - Legislations
  - Culture
  - Etc.

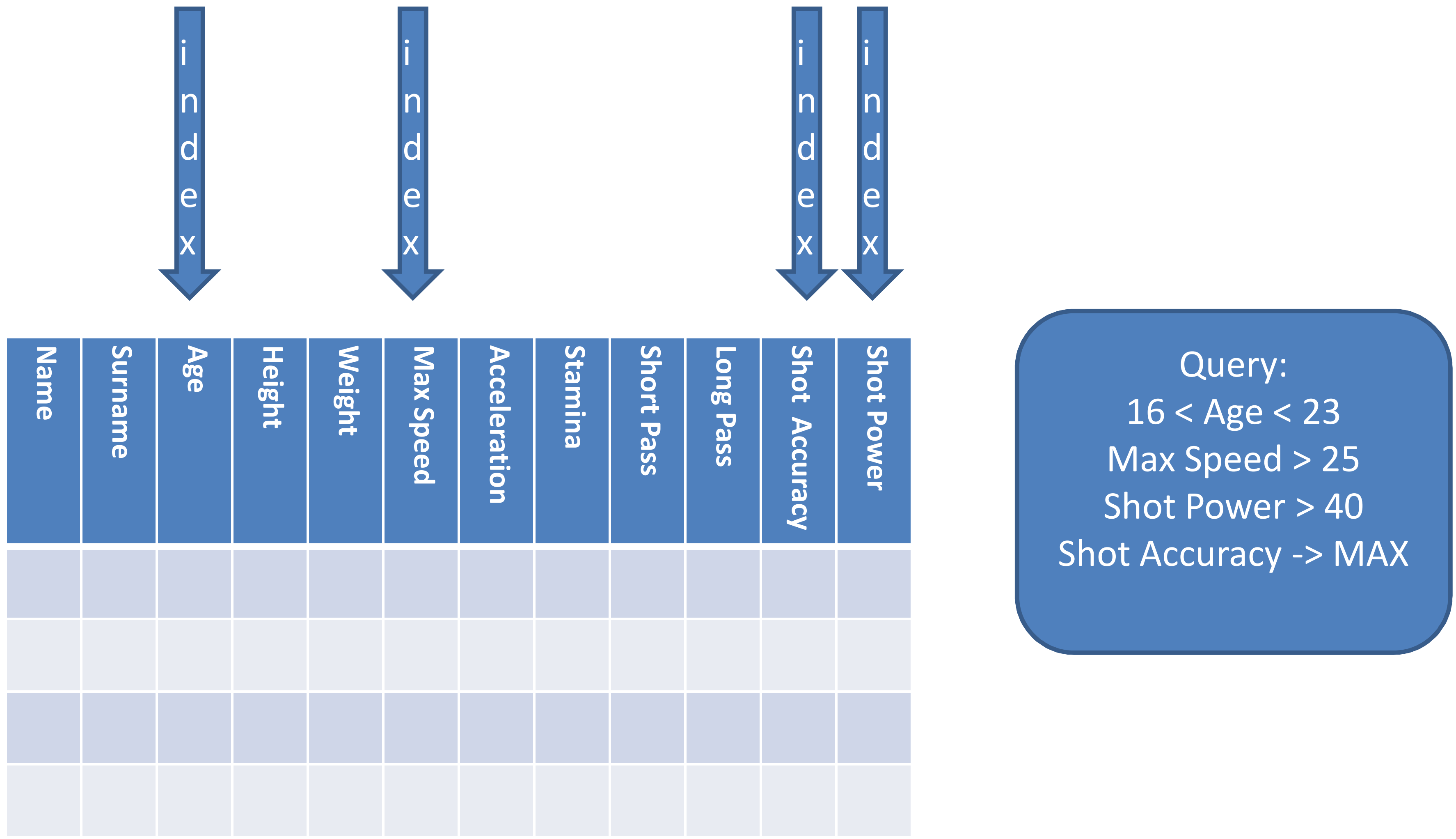
A new way to organize and to search the data

Data analytics 2014, Rome, Italy

Alexander Ponomarenko,
National Research University Higher School of Economics, LATNA Laboratory

Name
Surname
Age
Height
Weight
Max Speed
Acceleration
Stamina
Short Pass
Long Pass
Shot Accuracy
Shot Power

Name	Surname	Age	Height	Weight	Max Speed	Acceleration	Stamina	Short Pass	Long Pass	Shot Accuracy	Shot Power



index
index
index
index
Name
Surname
Age
Height
Weight
Max Speed
Acceleration
Stamina
Short Pass
Long Pass
Shot Accuracy
Shot Power
Query:
16 < Age < 23
Max Speed > 25
Shot Power > 40
Shot Accuracy -> MAX

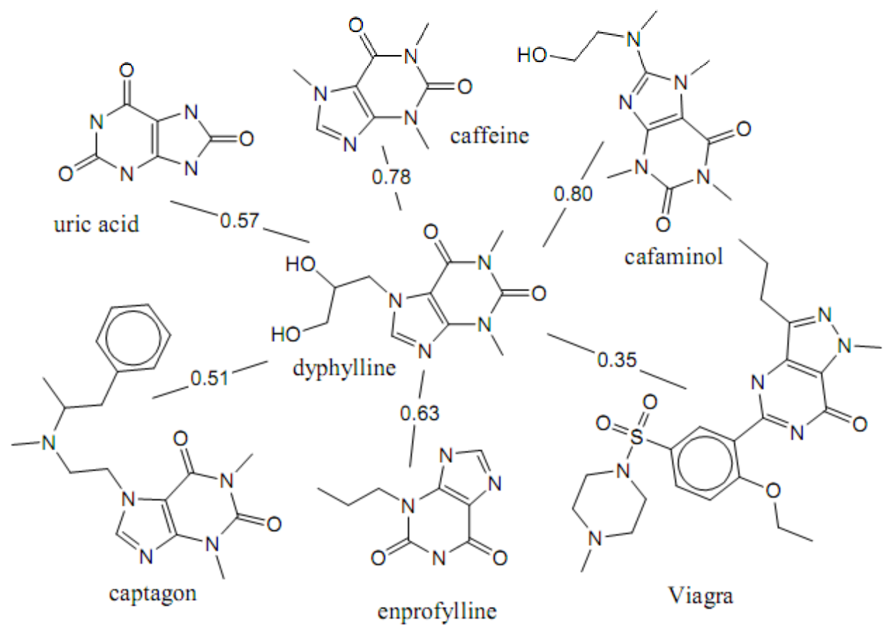
Why is similarity?

- Any event in the history of organism is, in a sense, **unique**.
- *Recognition*, *learning*, and *judgment* presuppose an ability to categorize stimuli and classify situations by **similarity**
- Similarity (*proximity*, *resemblance*, *communality*, *representativeness*, *psychological distance*, etc.) is **fundamental** to theories of *perception*, *learning*, *judgment*, etc.

Max Common Subgraph Similarity

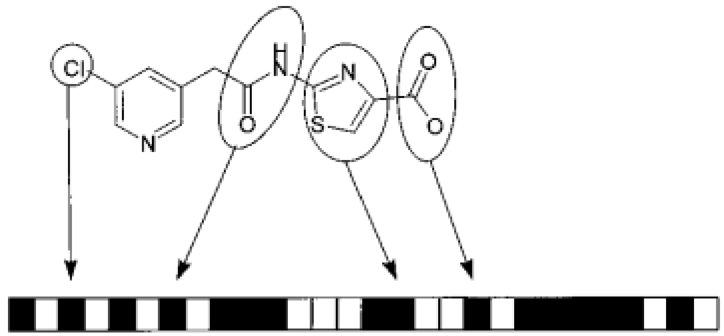
$$sim(G1,G2) = \frac{(|V(G1,G2)| + |E(G1,G2)|)^2}{(|V(G1)| + |E(G1)|)\cdot(|V(G2)| + |E(G2)|)}$$

$$d(G1,G2) = 1 - sim(G1,G2)$$



Tanimoto metric

$$tanimoto = \frac{c}{a+b-c}$$



a – number of non zero bits in first molecule fingerprint
b – number of non zero bits in second molecule fingerprint
c – number of common non zero bits

Fingerprint is array of bit where every bit corresponds to particular molecular feature

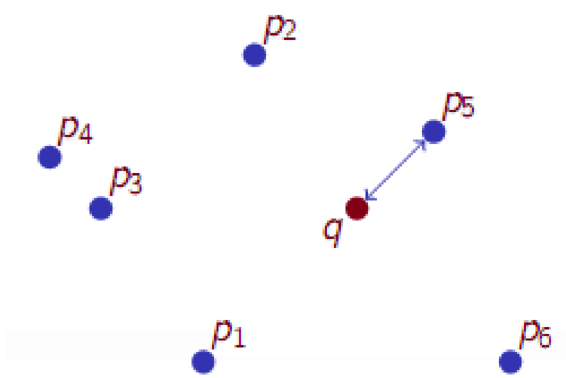
Nearest Neighbor Search

Let D – domain

$d: D \times D \to R_{[0;+\infty)}$ - distance function which satisfies properties:

- strict positiveness: d(x, y) > 0 ⇔ x ≠ y,
- symmetry: d(x, y) = d(y, x),
- reflexivity: d(x, x) = 0,
- triangle inequality: d(x, y)+d(y, z) ≥ d(x, z).

Given a finite set $X = \{p_1, \ldots, p_n\}$ of n points in some metric space *(D, d),* need to build a data structure on X so that for a given query point $q \in D$ one can find a point $p \in X$ which minimizes *d(p, q) with as few distance computations as possible*



Examples of Distance Functions

- L_p **Minkovski distance** (for vectors)
  - L_1 – city-block distance
  - L_2 – Euclidean distance
  - L_∞ – infinity

$$L_1(x, y) = \sum_{i=1}^{n} | x_i - y_i |$$

$$L_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

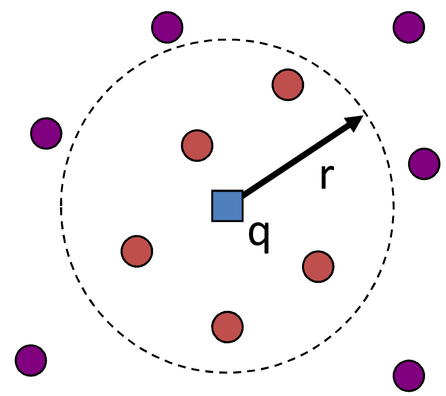
$$L_\infty(x, y) = \max_{i=1}^{n} |x_i - y_i|$$

- **Edit distance** (for strings)
  - minimal number of insertions, deletions and substitutions
  - d('application', 'applet') = 6

- **Jaccard's coefficient** (for sets A,B)

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Range Query



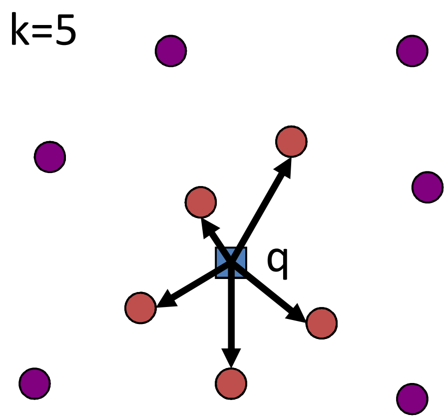
- range query
  - $R(q,r) = \{ x \in X \mid d(q,x) \leq r \}$

… all museums up to 2km from my hotel …

Nearest Neighbor Query

- the nearest neighbor query
  - $NN(q) = x$
  - $x \in X, \forall y \in X, d(q,x) \leq d(q,y)$

- k-nearest neighbor query
  - $k\text{-}NN(q,k) = A$
  - $A \subseteq X, |A| = k$
  - $\forall x \in A, y \in X - A, d(q,x) \leq d(q,y)$



… five closest museums to my hotel …

Our requirements for structure

Scalability

Distributed Architecture

- There should not be any central element (like p2p system)
- Any element of data structure should be able to perform search
- Any element of data structure should be able to start Adding process of new data

Scalable with number of elements

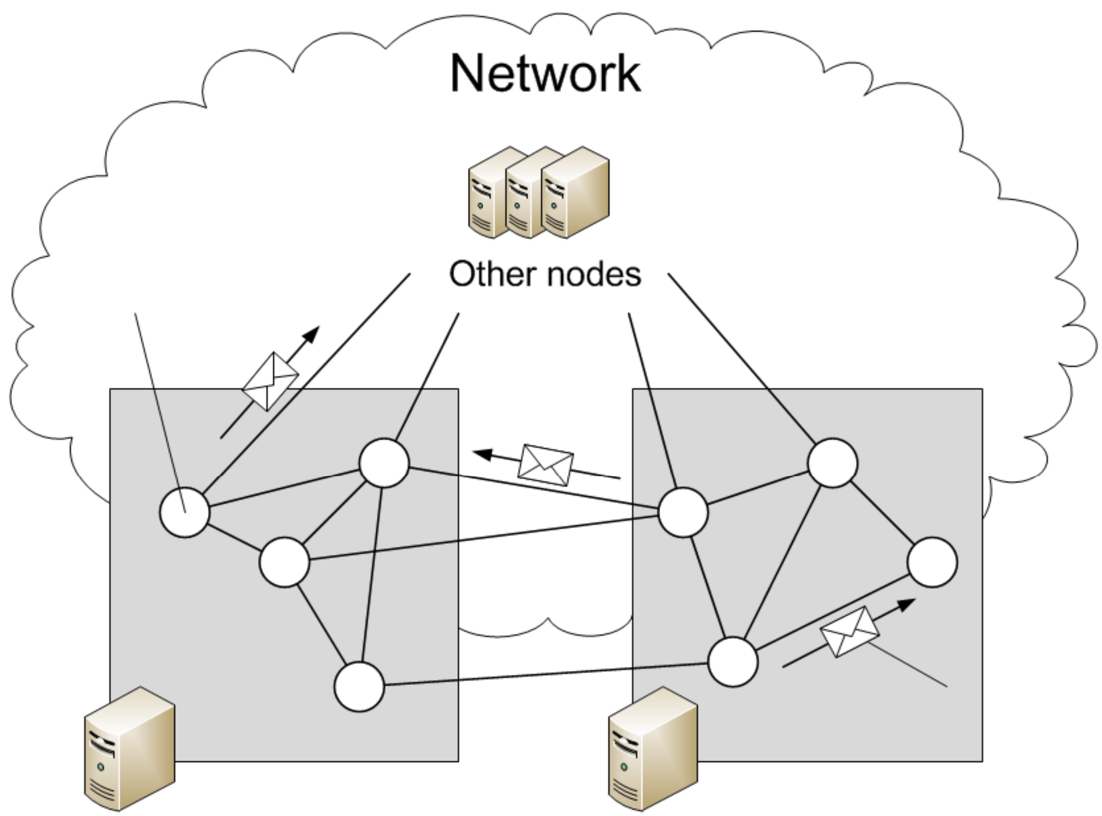
Insert and Search ~ log(n)

Universality

Abstract Metric Space

Scalable on number of dimension

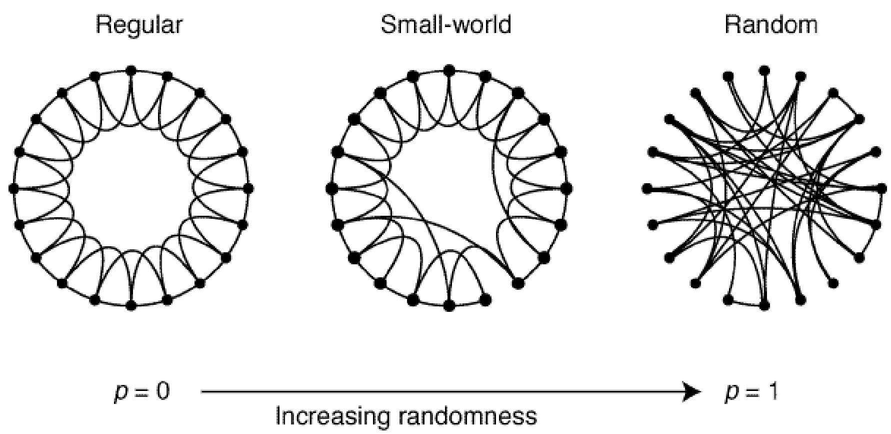
Relaxation to Approximate Nearest Neighbor (ANN)

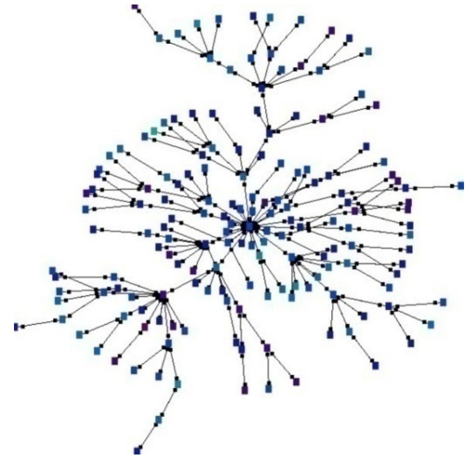


- Connect all data into the global overlay network on the level of data
- Use metric space search instead of sequence of exact searches

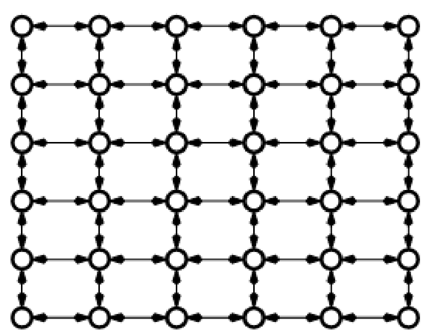
The Small World Networks

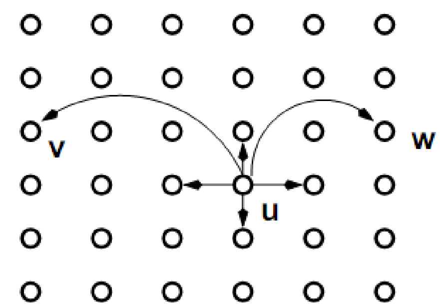
Two famous “Blind” models: “Watts-Strogatz” and “Barabási–Albert”



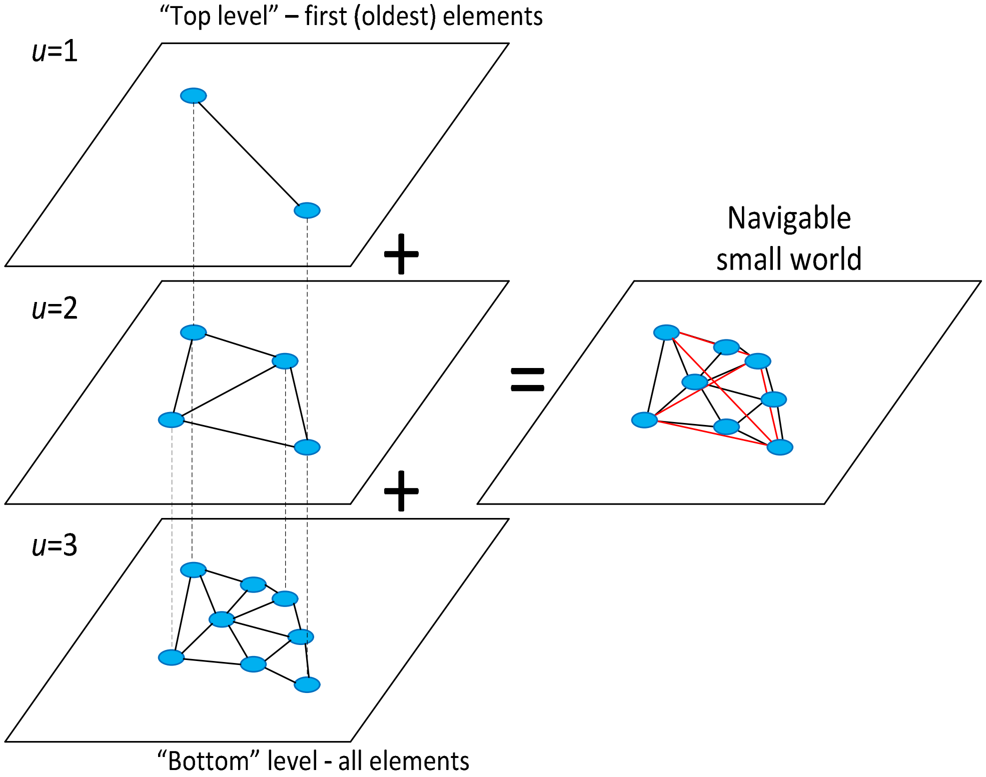


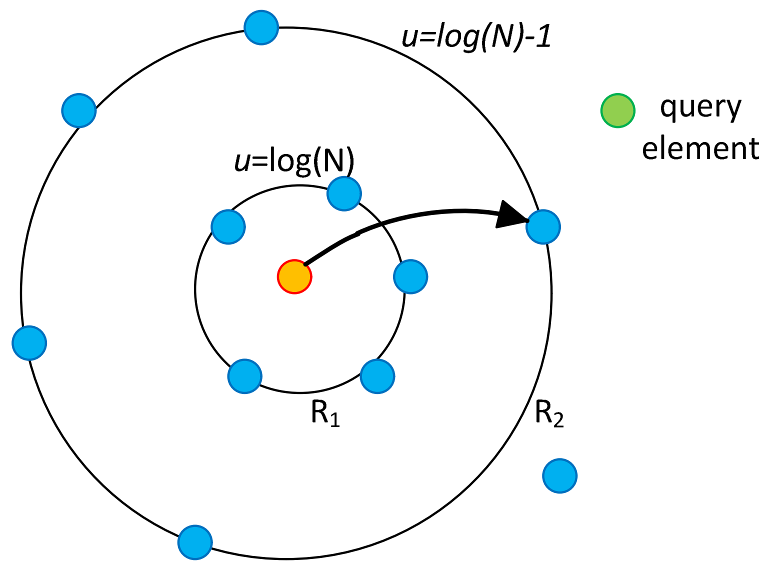
Navigable small world model of Klienberg





Navigable Small World





Wikipedia dataset

Vector Space Model

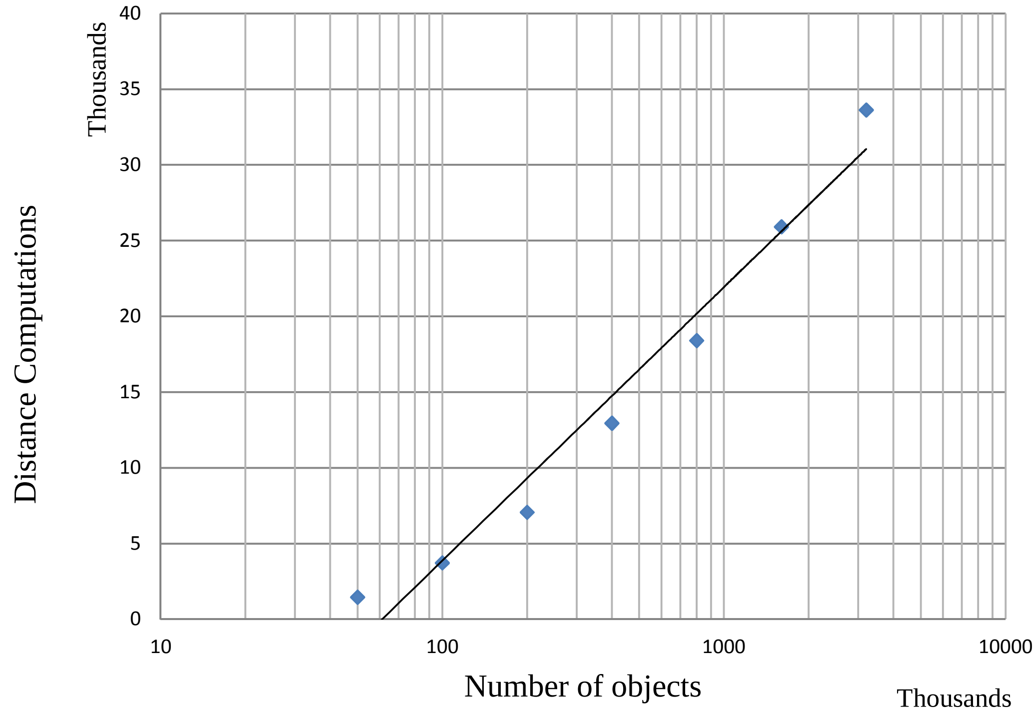
$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{n,j})$$

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$$

$$sim(d_j, q) = \frac{d_j \cdot q}{\| d_j \| \cdot \| q \|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

Wikipedia (cosine similarity): is a data set that contains 3.2 million vectors represented in a sparse format.
This set has an extremely high dimensionality (more than 100 thousand elements). Yet, the vectors are sparse: On average only about 600 elements are non-zero.

Scaling of Navigable Small World data structure



Non-Metric Space Library

Leonid Boytsov and Bilegsaikhan Naidan. "Engineering Efficient and Effective Non-metric Space Library." Similarity Search and Applications. Springer Berlin Heidelberg, 2013. 280-293.

Available at: https://github.com/searchivarius/NonMetricSpaceLib

Challenges in Managing Health Linked Big Data

Data Analytics 2014

Sandjai Bhulai
Associate Professor

VU VRIJE UNIVERSITEIT AMSTERDAM
IS VERDER KIJKEN

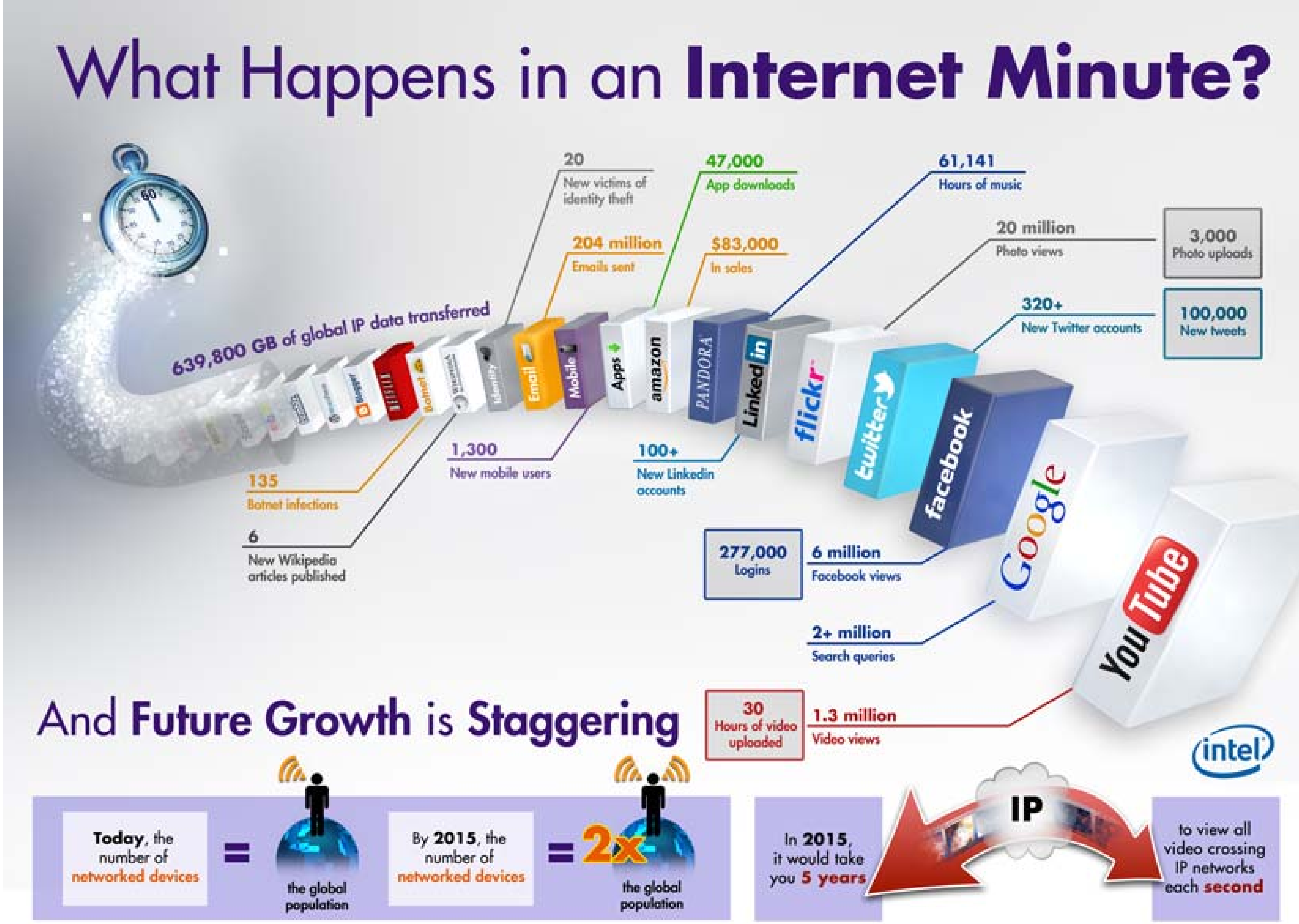
$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending


$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain
€250 billion
potential annual value to Europe's public sector administration—more than GDP of Greece
$600 billion
potential annual consumer surplus from using personal location data globally


What Happens in an Internet Minute?
639,800 GB of global IP data transferred
20
New victims of identity theft
204 million
Emails sent
47,000
App downloads
$83,000
In sales
61,141
Hours of music
20 million
Photo views
3,000
Photo uploads
320+
New Twitter accounts
100,000
New tweets
135
Botnet infections
6
New Wikipedia articles published
1,300
New mobile users
100+
New LinkedIn accounts
277,000
Logins
6 million
Facebook views
2+ million
Search queries
30
Hours of video uploaded
1.3 million
Video views
Identity
Email
Mobile
Apps
amazon
PANDORA
LinkedIn
flickr
twitter
facebook
Google
YouTube
And Future Growth is Staggering
Today, the number of networked devices
=
the global population
By 2015, the number of networked devices
= 2x
the global population
In 2015, it would take you 5 years
IP
to view all video crossing IP networks each second
intel

Social listening:

SOCIAL LISTENING

10 Top Apps For Eating Healthy

An a day keeps the doctor away….!

An APP a day keeps the doctor away….!

SOCIAL LISTENING

SOCIAL LISTENING

Panel on DATA ANALYTICS / GLOBAL HEALTH

Challenges in Managing Health Linked Big Data

August 27, 2014

Sung-Bae Cho
Department of Computer Science, Yonsei University
Seoul 120-749, Republic of Korea
http://sclab.yonsei.ac.kr

Outline

- Big data in healthcare
- The 3 "Vs" of big data in healthcare
- An architecture of big data analytics
- Challenges

Big Data in Healthcare

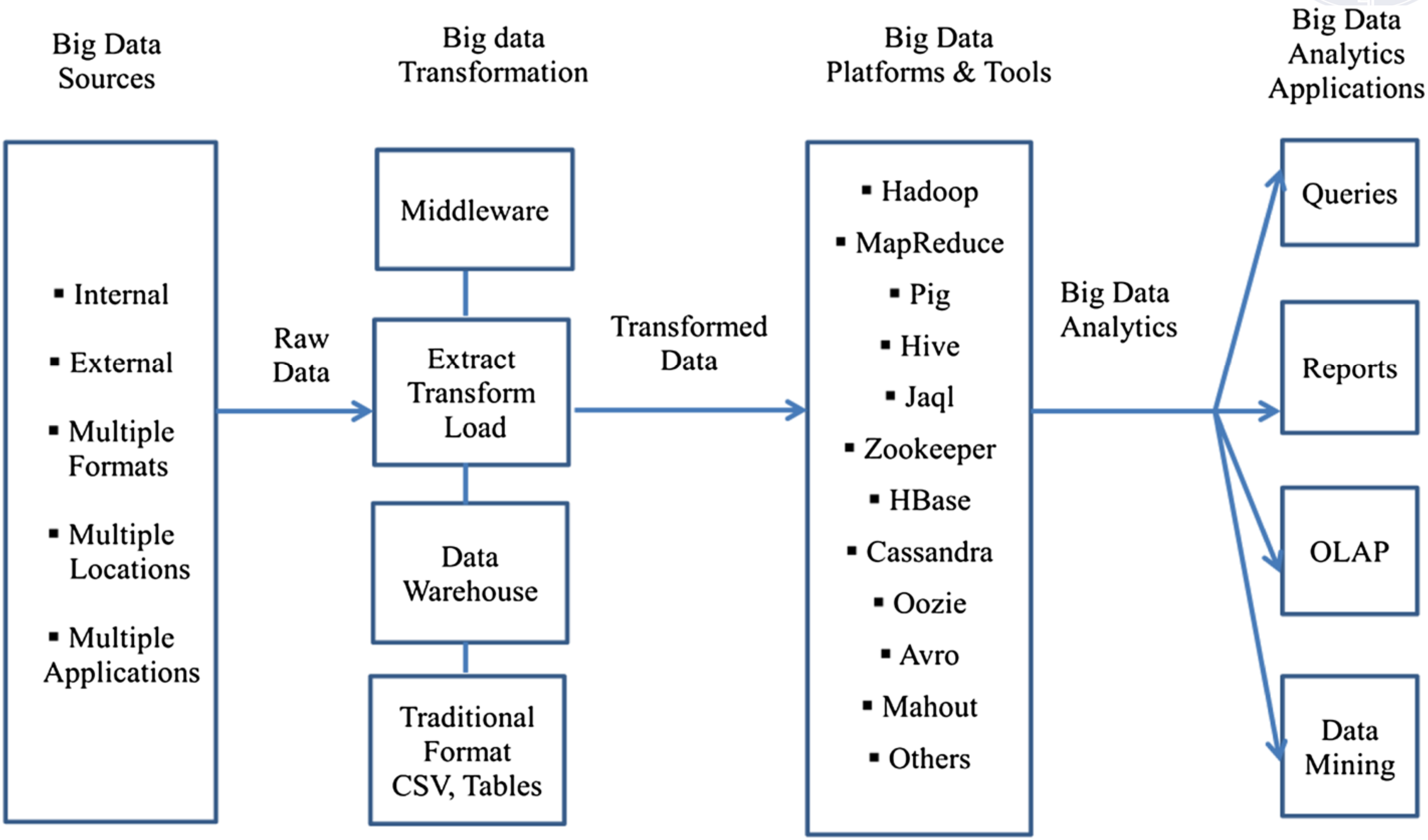
- Problem of big data in healthcare
  - Electronic health data sets are so large and complex that they are difficult to manage with traditional software and hardware; nor can they be easily managed with traditional or common data management tools and methods

- "Big data" in the healthcare industry
  - Clinical data from clinical decision support systems: physician's written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data
  - Patient data in electronic patient records
  - Machine generated/sensor data, such as from monitoring vital signs
  - Social media posts, including tweets, blogs, status updates on Facebook and other platforms, and web pages
  - Less patient-specific information, including emergency care data, news feeds, and articles in medical journals

The 3 “Vs” of Big Data in Healthcare

- Volume
  - Personal medical records, radiology images, clinical trial data FDA submissions, human genetics and population data genomic sequences
  - Newer forms of big data, such as 3D imaging, genomics and biometric sensor readings
- Velocity
  - Velocity of mounting data increases with data that represents regular monitoring, such as multiple daily diabetic glucose measurements, and blood pressure readings
- Variety
  - Structured and semi-structured data: instrument readings and data generated by the ongoing conversion of paper records to electronic health and medical records
  - Unstructured data: office medical records, handwritten nurse and doctor notes, hospital admission and discharge records, paper prescriptions, radiograph films, MRI, CT and other images

An Architecture of Big Data Analytics (Raghupathi & Raghupathi, 2014)



Challenges

- Legislation and governance challenges
  - Managerial issues of ownership, governance and standards have to be considered; Healthcare data is rarely standardized, often fragmented, or generated in legacy IT systems with incompatible formats
- Technical challenges
  - Data extraction and linkage
  - Data quality and accuracy
- Big data analytics challenges
  - Many architectures and platforms, and the dominance of the open source paradigm in the availability of tools
    - Lack of technical support and minimal security
    - Require a great deal of programming skills
  - Challenge of developing methodologies and the need for user-friendly interfaces
    - Big data analytics in healthcare needs to be packaged so it is menu-driven, user-friendly and transparent


Q&A

Challenges of Health Linked Big Data

Data Analytics / Global Health Panel Presentation

27 August 2014

Dr. Thomas J. Klemas
Senior Fellow
Sensemaking/PACOM Fellowship
Network Science Research Center
Swansea University, Swansea, Wales

Introduction to Big Data

Vast, complex data sets are a resource of great potential but require new approaches to analyze

Technology is enabler: storage capacity doubling every 40 months

Data is being generated at a dizzying rate which is also rapidly accelerating

- By 2003, 5 exabytes of data were generated by human kind
- Today, this much data is generated in a few days
- Now, total human data is measured in Zetabytes (ZB) and is expected to double every 2 years

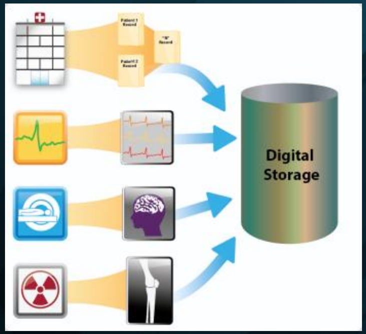
Digital format enables application of Data Science methods

- By 2000, 25% of data stored digitally.
- Today more than 98% of data is stored digitally.

2012: Big Data Research and Development Initiative ($200M)

Characteristics of Big Data

Related Challenges



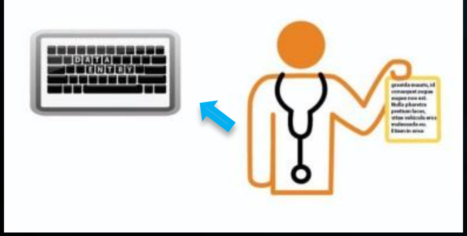
Size: vast amounts of historical data

Velocity: streams of real-time, very recent data

Diversity: many components of data that can be stored in variety of forms

Uncertainty: data and metadata can be incorrect or incomplete

Accessibility: availability of data to those that need it



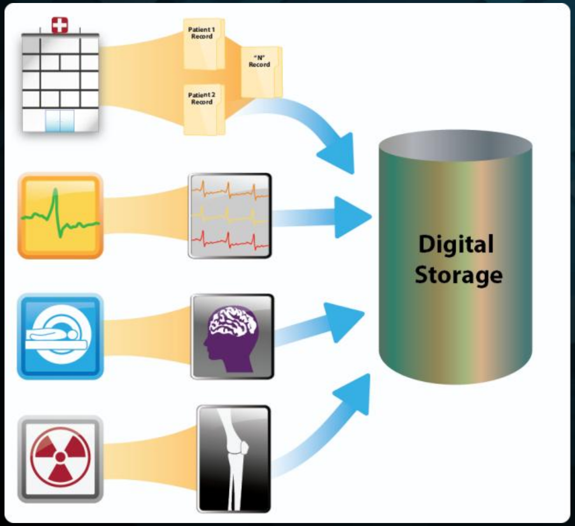
Goal: Maximize value – Keep important data, Use best sources, Integrate appropriate metadata and linkage to other enabling data

Big Data in Global Health Care

Healthcare accounted for 500 PB in 2012 but expected to reach 25000 PB by 2020

Conversion of existing data to electronic form

Generation of new data: images, sensor readings, genomics



Traditional Global Health Data Sources

Clinical data records: diagnoses, prescriptions, notes, images, …

Genetic data, biometrics from sensors

Self-reported data, billing and cost data

Statistics from medical facilities, clinical trials

Additional sources to enable big data: Social & economic data, environmental factors

Significant potential to aid medical professionals with Analytics and Data Science improvements

Example: Colorectal Cancer

Electronic Health Records (EHR)

Adoption Rate and Related Details

2005: 30% medical facilities use EHR in USA*
2008: 17% of doctor offices
2009
- 9% of hospitals
- Health Information Technology for Economic and Clinical Health (HITECH) Act

2011
- 50% medical facilities use EMR in USA*
- 75% hospitals use EMR in USA*

2012
- 40% of doctor's offices
- 44% of hospitals

2013: Nearly $16.6B HITECH funding disbursed to more than 4100 hospitals and 320K medical professionals

* Different sources offered different values for EHR adoption. Discrepancy may be due to differing assessment or definition of "EHR"

Improving Global Health

Big Data Potential

Medical care contributes an estimated 10% to an individual's health*
Other factors are more significant to health

- Behavior
- Genetics
- Social and Economic status
- Environment

Big Data for Healthcare will require integrated analysis of medical data merged with associated external data describing behavior, environment, social and economic factors, etc.

* Assessment from "A Policy Forum on the Use of Big Data in Health Care", Health Program, *Health Innovation Initiative, Meeting Proceedings, June 25, 2013.*

Motivations for Big Data

Potential Benefits

Detect and identify
- Highest Risk Individuals
- Public health threats

Monitor patients and communities

Support Decision Making of Health Professionals and Patients
- Aid diagnosis
- Help focus treatment and evaluate efficacy.
- Predict impact of threats. Aid targeting of resources to combat threats
- Provide valuable statistics and related information
- Improve allocation of and access to resources

Support Discoveries and Innovation
- Recognize patterns that are key factors towards health
- Aid discovery of new relationships or links
- Example: Identify cause and discover cures for disease

Achieve new Paradigm: results-oriented vs service-oriented medical care
Reduce cost

Challenges to Healthcare Big Data

- Quality of data decreases usefulness
  - Frequently incomplete or biased by errors and omissions
    - Primarily due to human entry of clinical data
    - Examples: Description of symptoms, notes
  - Lack of context due to incomplete or inaccurate metadata
- Limited access to data
  - Privacy and security concerns
  - Proprietary value of data
  - Additional challenge to link with external data sources
- Incompatibility between data sets
  - Multiple data sets cannot be easily compared due to parameter differences
  - Formatting or lack of standards hinder use of data

Overcoming Big Data Challenges

Requires Health Data Enhancements

- Integration of data
  - Claims and Costs
  - Clinical records and medical images
  - Pharmaceutical RD data, clinical trials
  - Patient behavior and attitude/feelings/opinions data
  - Other
- Improved Data quality
  - Data standards
  - Better data entry
    - Automation: Standardized computerized forms, Improved Natural Language Processing, Data Structuring and Grouping tools
- Increased Sharing, Security, and Availability of Analytical, Visualization Tools
- Innovative Data Science Techniques to optimize impact of data and to achieve transformation to Outcome Driven Medicine

THANK YOU!

References

1) Groves, Kayyali, Knott, Van Kuiken, “The Big Data Revolution in Healthcare Accelerating Value and Innovation”, Center for US Health System Reform Business Technology Office, McKinsey&Company, Jan 2013
2) “A Policy Forum on the Use of Big Data in Health Care”, Health Program, Health Innovation Initiative, Meeting Proceedings, June 25, 2013.
3) Slabaugh, “Big Data Challenges and opportunities in healthcare: application to detecting faint signals”, City University London, School of Informatics.
4) Lawrence, Pickhardt, Kim, Robbins, “Colorectal polyps: stand-alone performance of computer-aided detection in a large asymptomatic screening population”, In Radiology, 255, 791-798, 2010
5) Patil, Seshadri, “Big Data security and privacy issues in healthcare”, Nanthealth
6) Schultz, “Turning Healthcare Challenges into Big Data Opportunities: A Use-Case Review Across the Pharmaceutical Development Lifecycle”, Bulletin of the Association for Information Science and Technology, Vol 35 No 5, June – July 2013
7) Hamilton, “Big Data is the Future of Healthcare”, Cognizant Business Consulting, 20-20 Insights, Sept 2012.
8) Reddy, Sun, “Big Data Analytics for Healthcare”, Tutorial for SIAM International Conference on Data Mining, 2013
9) Adler-Milstein, Jha, “Healthcare's Big Data Challenge”, http://www.ajmc.com/publications/issue/2013/2013-1-vol19-n7/Healthcares-Big-Data-Challenge, July 16, 2013

References (continued)

10) Orlanes, https://www.linkedin.com/today/post/article/20140628172719-309345724-big-data-challenges-and-opportunities-of-diagnostic-imaging-technologies-in-the-provider-patient-care-delivery-setting, LinkedIn
11) http://obssr.od.nih.gov/scientific_areas/methodology/mhealth/KDD2014.aspx, Office of Behavioral and Social Sciences Research website, National Institute of Health
12) http://www-03.ibm.com/software/products/en/category/bigdata, IBM website
13) Riskin, http://www.healthcareitnews.com/news/big-data-opportunity-and-challenge, Healthcare IT News website