BIG DATA MEET THE EYES OF THE LIBRARIAN

Presented by: Muhammad Akmal Ahmat

in *"Seminar Staf Ikhtisas Perpustakaan Sains Malaysia"*

19/12/2013

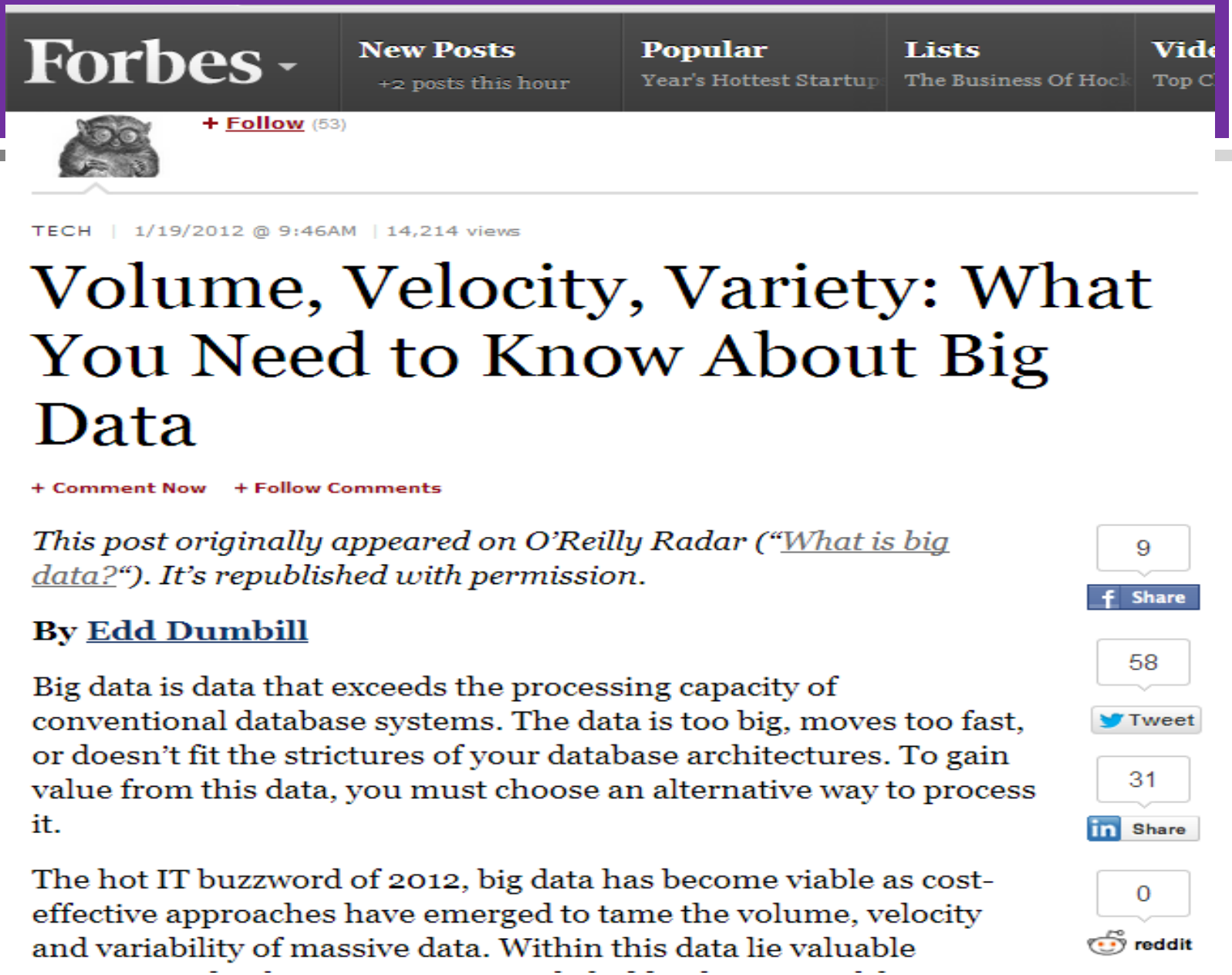
WHAT IS BIG DATA

We lead

Defining Big Data: Volume, Velocity, and Variety

- By Judith Hurwitz, Alan Nugent, Fern Halper, and Marcia Kaufman
- **Part of the Big Data For Dummies Cheat Sheet**
- Big data enables organizations to store, manage, and manipulate vast amounts of disparate data at the right speed and at the right time. To gain the right insights, big data is typically broken down by three characteristics:
- **Volume:** How much data
- **Velocity:** How fast data is processed
- **Variety:** The various types of data
- While it is convenient to simplify big data into the three Vs, it can be misleading and overly simplistic. For example, you may be managing a relatively small amount of very disparate, complex data or you may be processing a huge volume of very simple data. That simple data may be all structured or all unstructured.

URL: http://www.dummies.com/how-to/content/big-data-for-dummies-cheat-sheet.html



http://strata.oreilly.com/2012/01/what-is-big-data.html

Type	Explanation
METADATA	data about data
LINKED DATA	a method of publishing structural data so that it can be interlinked and become more useful
OPEN DATA	The data is free
BIG DATA	a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications

http://www.oclc.org/content/dam/oclc/events/2013/CollectiveInsightBaltimore/CollectiveInsight_Baltimore02_JimNeal.pdf

DATA, DATA IN THE CLOUD
LIBRARIES ARE CLEARLY WELL-ENDOWED
James G. Neal

FACT ABOUT DATA

We lead

2004

- the global monthly Internet traffic passed 1 Exabyte for the first time.

Bret Swanson (January 20, 2007). "The Coming Exaflood". Wall Street Journal. Retrieved 2007-02-17.

2007

- Bret Swanson of the Discovery Institute coined the term **exaflood** for a supposedly impending flood of Exabyte that would cause the Internet's congestive collapse. 295 Exabyte.

Bret Swanson (January 20, 2007). "The Coming Exaflood". Wall Street Journal. Retrieved 2007-02-17.

2009

- The global data volume at the end of 2009 had reached 800 Exabyte

http://www.genevaassociation.org/PDF/Risk_Management/GA2010-RM47.pdf

FACT ABOUT DATA

We lead

- an Exabyte could hold a hundred thousand times all the printed material, or 500 to 3,000 times all content of the Library of Congress. That means,

Hard disk (Digital)		Libraries of Congress (Physical)
33.33 Exabyte	CONVERT	**100,000**

http://blogs.loc.gov/digitalpreservation/2012/04/a-library-of-congress-worth-of-data-its-all-in-how-you-define-it/

Prof. Kirk Borne

Refer presentation: Big Data, Small World

Data Scientist, Astrophysicist,
Professor of Astrophysics and Computational Science,
Big Data Science Consultant, & Public Speaker

URL: https://www.youtube.com/watch?v=Zr02fMBfuRA

3 characteristics of big data

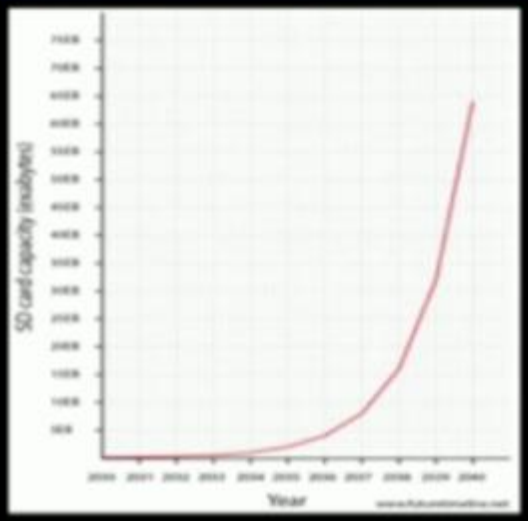
We lead

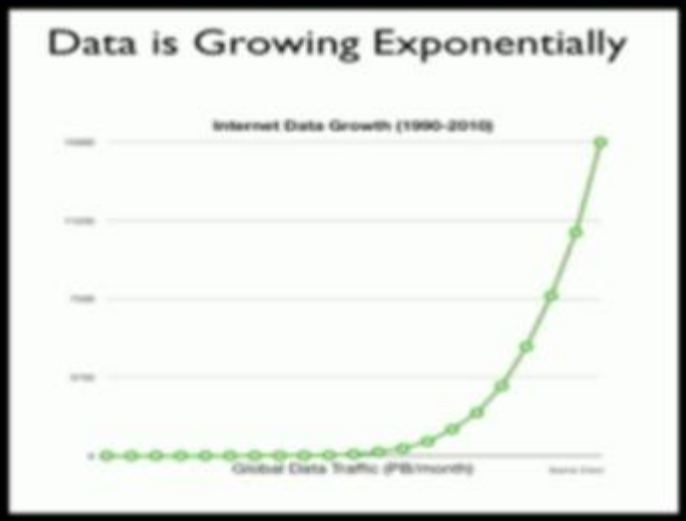
#1

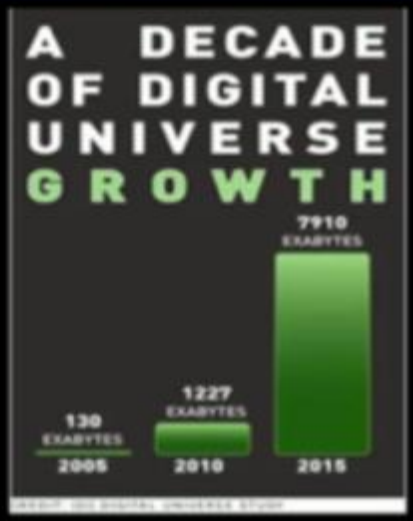
There are **huge** volumes of data in the world:

- From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data.
- In 2011 the same amount was created every two days.
- In 2013, the same amount is created every 10 minutes.

http://money.cnn.com/gallery/technology/2012/09/10/big-data.fortune/index.html





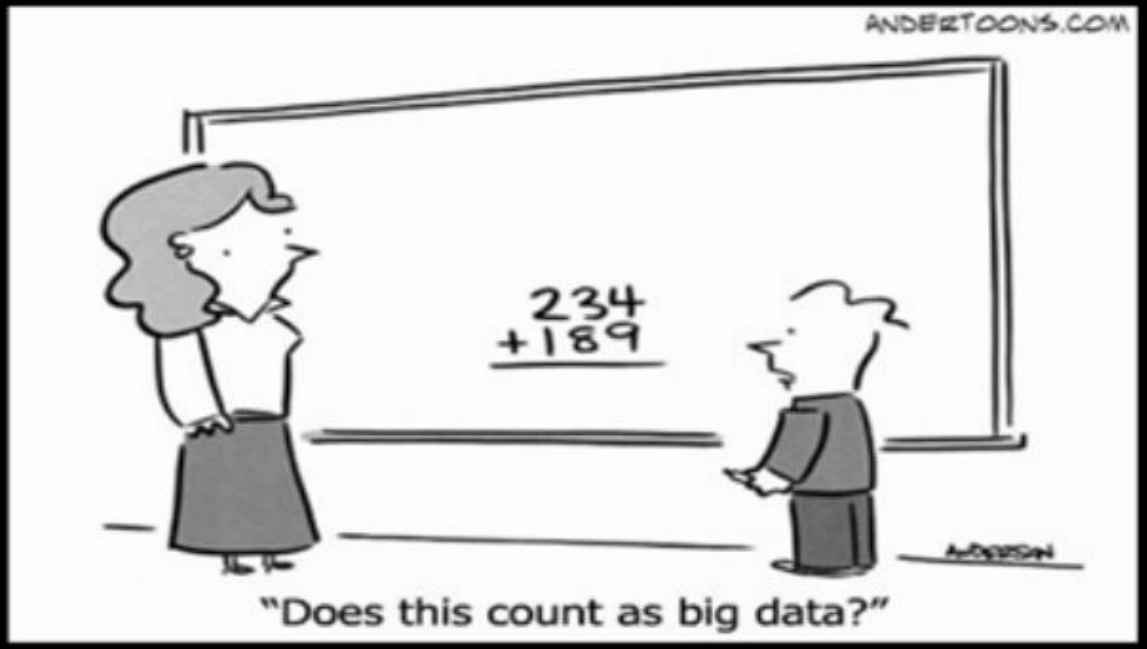


Source: http://www.youtube.com/watch?v=Zr02fMBfuRA

3 Characteristics of Big Data – #123

Huge quantities of data are acquired everywhere:

- **Big Data** is a big issue in all aspects of life: science, social networks, transportation, business, healthcare, government, national security, media, education, etc.

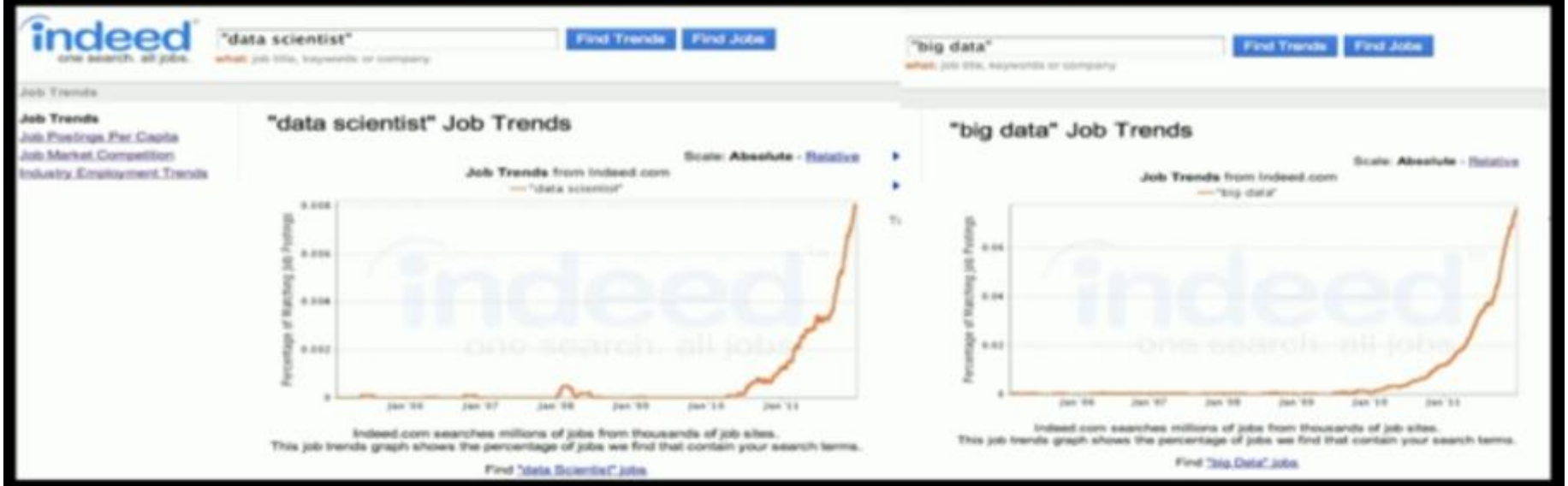


Source: http://www.youtube.com/watch?v=Zr02fMBfuRA

3 Characteristics of Big Data – #123

Job opportunities are sky-rocketing:

- Extremely high demand for Big Data analysis skills
- Demand will continue to increase
- Old: "100 applicants per job". New: "100 jobs per applicant"



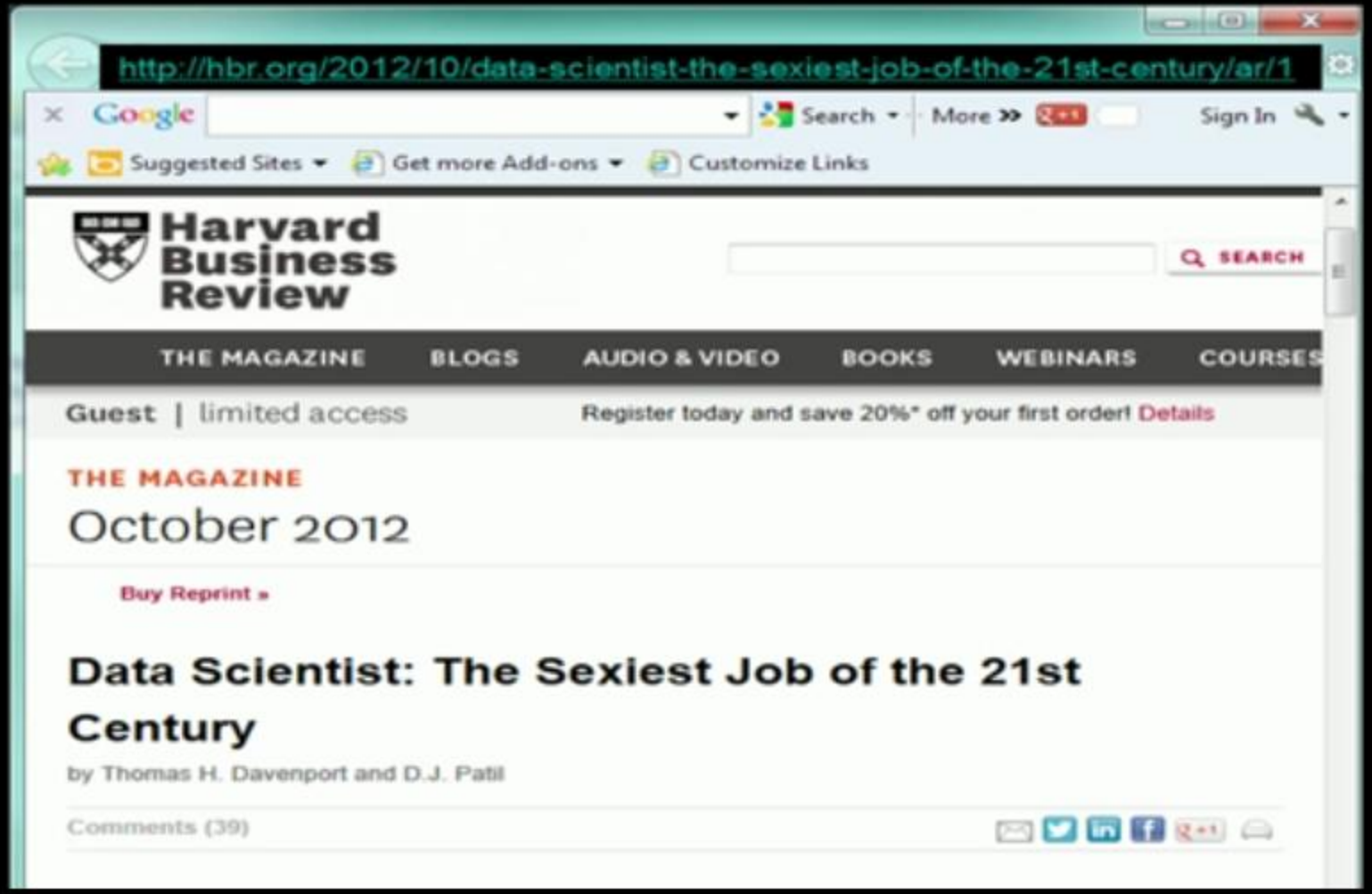
Source: http://www.youtube.com/watch?v=Zr02fMBfuRA

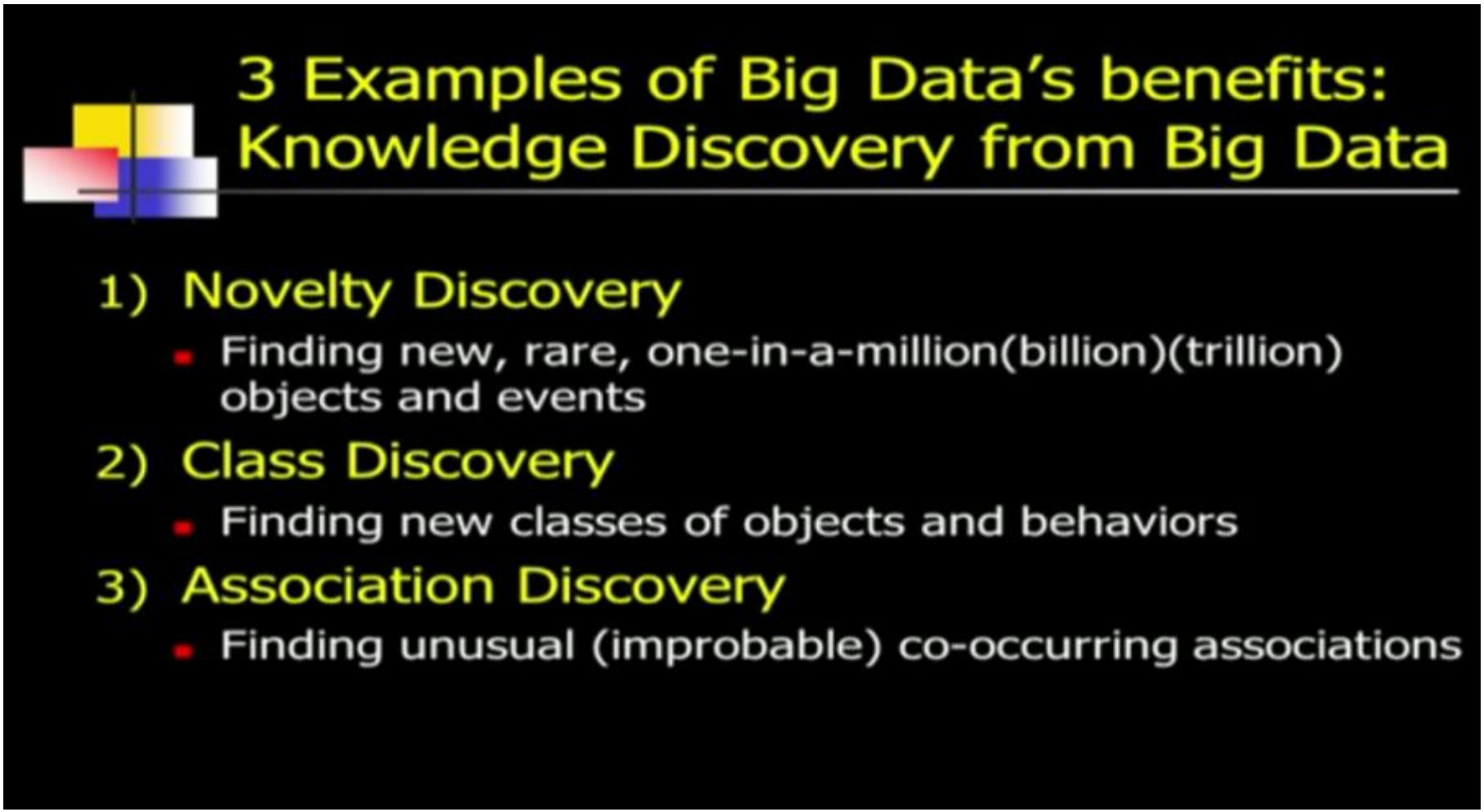
Source: http://www.youtube.com/watch?v=Zr02fMBfuRA

News #2 - Promising: Big Data leads to Big Insights and New Discoveries

Source: http://www.youtube.com/watch?v=Zr02fMBfuRA

News #3 - Good: Big Data is Sexy

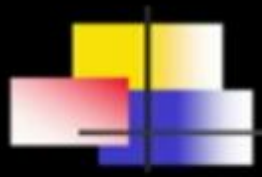



3 Examples of Big Data's benefits:
Knowledge Discovery from Big Data
1) Novelty Discovery
Finding new, rare, one-in-a-million(billion)(trillion) objects and events
2) Class Discovery
Finding new classes of objects and behaviors
3) Association Discovery
Finding unusual (improbable) co-occurring associations

Simple Example of Novelty Discovery:

Discovering the one that is different ...

Simple Example of Information Classification:



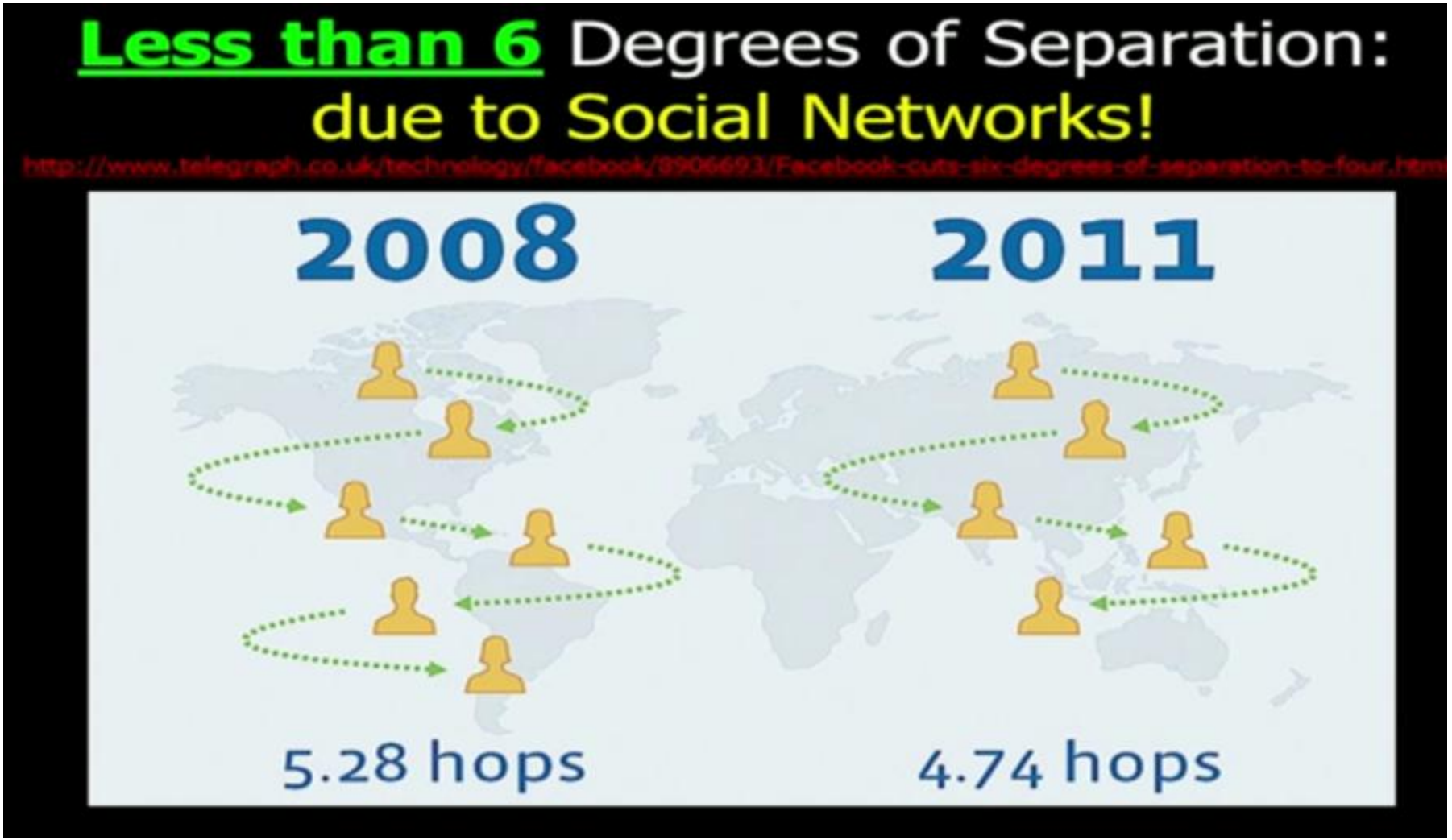
What is Association Discovery?

- Identifying connections between different things (people or events)
- Finding unusual (improbable) co-occurring combinations of things (for example: in your shopping cart)
- Finding things that have much fewer than "six degrees of separation"

6 Degrees of Separation:

Everyone is on average approximately 6 steps away from any other person on Earth (through their relationships with each other).

http://info.logicmanager.com/bid/86132/ERM-and-the-Six-Degrees-of-Separation-Theory


Less than 6 Degrees of Separation:
due to Social Networks!
http://www.telegraph.co.uk/technology/facebook/8906693/Facebook-cuts-six-degrees-of-separation-to-four.html
2008
2011
5.28 hops
4.74 hops

www.telegraph.co.uk/technology/facebook/8906693/Facebook-cuts-six-d

By Emma Barnett, Digital Media Editor
1:50PM GMT 22 Nov 2011

Follow 25.9K followers

24 Comments

Since the American social psychologist, Stanley Milgram, conducted his famous 'small world experiment' in the 1960s, it has been commonly accepted that most people have six degrees of separation between them.

However, a vast new study by **Facebook**'s data team and the University of Milan, which assessed the relationships between 721 million active users (more than 10 per cent of the global population) of the social network, has found that the average number of connections between people has dropped to four.

The **huge piece of research**, which took a month to conduct and analysed 69 billion connections across the site, found that 92 per cent of facebook users are connected by only 5 hops, or four degrees of separation.

"Using state-of-the-art algorithms…we were able to approximate the number of 'hops' [degrees of separation] between all pairs of individuals on Facebook.

"We found that six degrees actually overstates the number of links between typical pairs of users: While 99.6% of all pairs of users are connected by paths with 5 degrees (6 hops), 92% are connected by only four degrees (5 hops)," the **Facebook Data team said**.

Related Articles

Print this article

Share 1K

Facebook 881

Twitter 431

Email

LinkedIn 0

g+1 0

Facebook

News »
Media and Telecoms »
Media »
Technology »
Digital Media »

In Technology »

Galaxy S4 in pictures

In Facebook

The Six Degrees of Kevin Bacon!
... K.B. is now only 2 steps from K.B.

http://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon

Kevin Bacon,
appeared in the play
"An Almost Holy Picture"

Heather McDonald, **GMU**
Playwright, author of
"An Almost Holy Picture"

Kirk Borne :
The Borne Identity,
Professor of Data Science,
GMU

ebay

6 Degrees of Separation works with people, and it works with any type of network, including the network of things that you like or purchase linked to things that other people buy or like.

http://nosql.mypopescu.com/post/46508012660/graph-based-recommendation-systems-at-ebay

4 Examples: how Big Data is shrinking your world – Small World Connections through Associations

amazon.com

NETFLIX

Walmart

Example #1: how Big Data is shrinking your world – Small World Connections through Associations

- **Classic Textbook Example of Data Mining** (Legend?): Data mining of grocery store logs indicated that men who buy diapers also tend to buy beer at the same time.

Example #2: how Big Data is shrinking your world – Small World Connections through Associations

- Amazon.com mines its customers' purchase logs to recommend books to you: *"People who bought this book also bought this other one."*

Example #4: how Big Data is shrinking your world – Small World Connections through Associations

- **Wal-Mart** studied product sales in their Florida stores in 2004 when several hurricanes passed through Florida.
- Wal-Mart found that, before the hurricanes arrived, people purchased 7 times as many of ***{one particular product}*** compared to everything else.

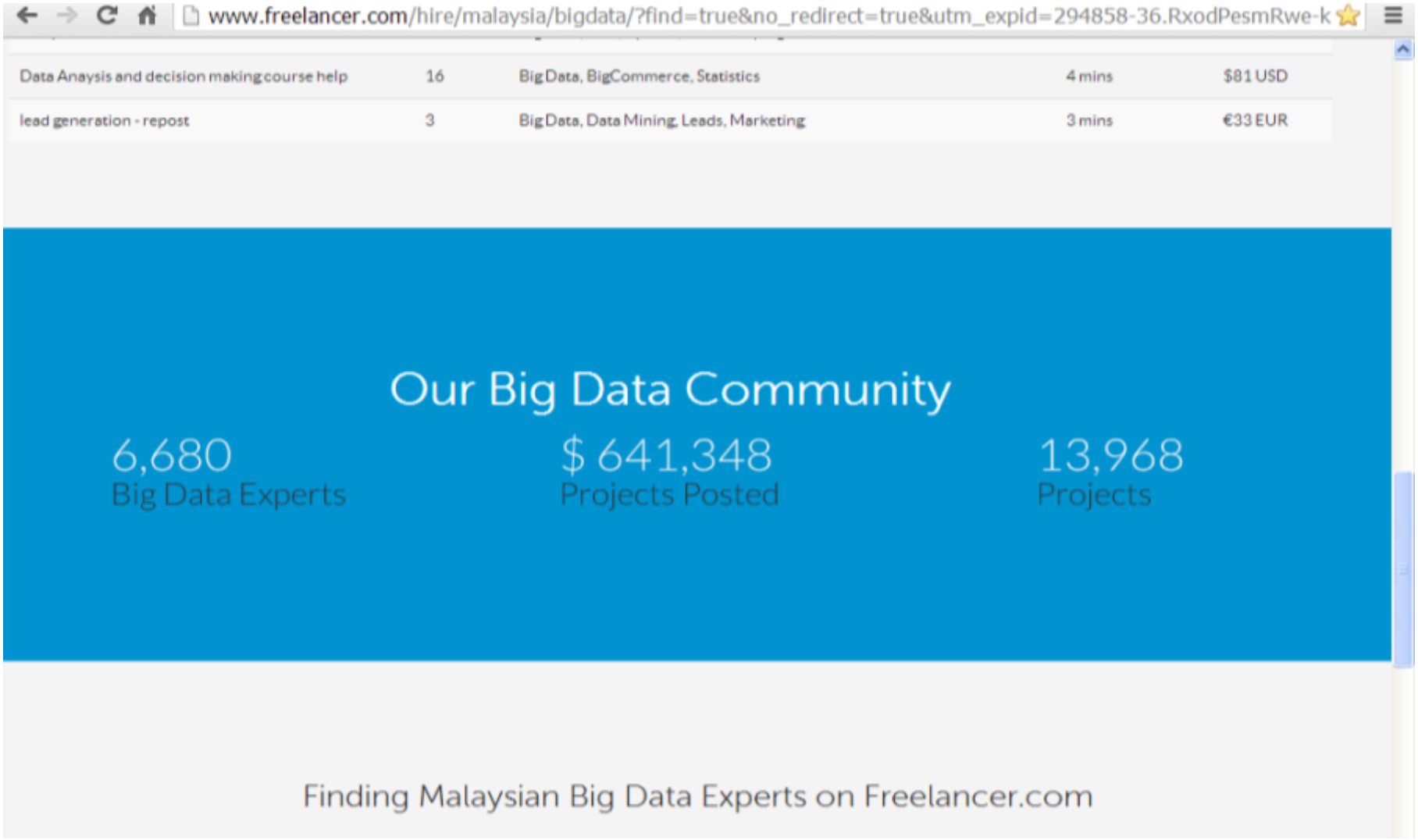
700% increase sales

Strawberry pop tarts???

http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html
http://www.hurricaneville.com/pop_tarts.html

BIG DATA = MONEY?

We lead



venturebeat.com/2013/12/09/big-data-startups-pull-in-big-money-in-2013/

Big Data
Business
Cloud
Deals
Dev
Education
Enterprise
Entrepreneur
Gadgets
Green
Health
Lifestyle
Media
Mobile
New York
Science
Security
Small Biz
Social
GAMESBEAT

Big data startups pull in big money in 2013

The money is flowing to big data startups in a big way.

December 9, 2013 11:43 AM
Jordan Novet

Investors have pumped $3.6 billion into startups focused on big data this year. Not too shabby: It's almost three quarters of all the money that's gone into

www.nst.com.my/latest/shabery-malaysia-must-p

13 October 2013 | last updated at 01:37PM Email Print

Shabery: Malaysia must produce many data scientists

Share 6 Tweet 2 Google + 8 Pinterest 0 0 comments

KEMAMAN: Malaysia has to produce many data scientists to drive its effort to become a data management centre so as to pioneer new economic opportunities on a global scale, Communication and Multimedia Minister Datuk Ahmad Shabery Cheek said today.

He said it was the desire of Prime Minister Datuk Seri Najib Razak to see Malaysia not only as a consumer in the information technology sector but also as a major player in IT development.

"Education must also be geared towards that direction. Hitherto, we have sent our children to become engineers, chemists and aeronautical

M'sia perlu hantar pelajar saintis data besar-besaran

Online Metro | October 13, 2013 | Reply

Malaysia memerlukan lebih ramai saintis data dalam usaha menjadikan negara ini sebagai pusat pengurusan data, dalam usaha memacu negara mempelopori peluang-peluang ekonomi baharu di peringkat global.
Selaras dengan hasrat Perdana Menteri Datuk Seri Najib Tun Razak yang mahu Malaysia bukan sahaja menjadi pengguna dalam sektor teknologi maklumat (IT), Menteri Komunikasi dan Multimedia Datuk Seri Ahmad Shabery Cheek berkata negara perlu melahirkan ramai saintis data bagi memacu negara sebagai peserta utama dalam perkembangan terkini pembangunan industri IT itu.

adi jurutera, ahli

besaran, dan kita

Who are the right professional that knowledgeable in running mission critical data centres?

- **IBM says 'big data' will transform schools, hospitals — and malls**

Devin Coldewey NBC News

http://www.youtube.com/watch?v=hTA5GyWamR0&src_vid=yKNSOwLcrkE&feature=iv&annotation_id=annotation_1030989583

James G. Neal

James G. Neal
Vice President for Information Services and University Librarian, Columbia University

He is currently Chair of the ALA Budget Analysis and Review Committee (BARC) and has served on the Council and Executive Board of the American Library Association. He has served on the Board and as President of the Association of Research Libraries, on the Board and as Chair of the Research Libraries Group (RLG), and as Chair of the RLG Programs Committee of the OCLC Board.

Refer presentation: *Data, Data in the Cloud, Libraries are Clearly Well-Endowed*

URL: http://www.youtube.com/watch?v=a0kg-Low74E

NEW TECHNOLOGIES AND THE 21ST CENTURY ACADEMIC LIBRARY

We lead

- Mobiles and Tablets (single, portable multi-purpose device)
- Cloud Computing (distributed processing and applications)
- Geo-Everything (geolocation and geotagging)
- Personal Web (customized management of online content)
- **Linked Data (connecting and relating structured information)**
- Semantic-Aware Applications (meaning to provide answers)
- Smart Objects (links physical world with information)
- Open Content (wide distribution and repurposing)
- MOOC (massive open online courses)
- Electronic Book (platforms, applications, redefinition)
- **DATA/BIG SCIENCE (RESEARCH INFORMATION MANAGEMENT)**
- Games As Learning Tools (participation and interaction)
- Visualization and Simulation (more meaningful and intuitive)

Refer: Data, Data in the Cloud, Libraries are Clearly Well-Endowed

BUILD THE DIGITAL LIBRARY QUALITY =

- Published/Licensed Content
- Primary Content
- Open Web Content
- Institutional Content
- **Research Data Content**
- Multimedia Content
- Integrated Services
- Software Tools

ADVANCE THE REPOSITORY MOVEMENT

We lead

- Discipline Repositories
- Institutional Repositories
- **Data Repositories**
- Departmental/School Repositories
- Individual Repositories
- Learning Repositories
- Government Repositories
- National Repositories
- Publisher Repositories
- **Research Data Repositories**

SUPPORT THE NEEDS OF RESEARCH

We lead

- Navigate, Analyze, Synthesize
- Open Research/Continuous Scholarly Communication
- Scholarly Products to Scholarly Process
- Expertise Databases/Subject Ontologies
- **Data Management Consulting**
- Integration of Disparate Sources/Grey Literature
- Special Library/Informationalist Model

SUPPORT THE NEEDS OF BIG DATA

We lead

- Federal/Funding Agency Mandate
- **Massive Data Sets**
- Unstructured Data/Curation
- Extraction
- Distribution
- Collaboration
- Visualization
- Simulation
- Preservation

DATA MEET THE "HUMAN" OBJECTIVES

We lead

- Success (turn out well, attain desired end)
- Happiness (well-being and contentment)
- Productivity (achieving results or benefits)
- Progress (forward movement or betterment)
- Relationships (personal connections or attachments)
- Experiences (observation or participation)
- Impact (significant effect)

DATA MEET USERS?

- Students (diversity abounds)
- Faculty (expectations galore)
- Researchers (tribal differences)
- Administration (the bottom line)
- Community (local politics)
- Working Professionals (practical applications)
- Alumni and Donors (largely ignored)
- World on the Web (the new majority)

SOME ISSUES FOR DISCUSSION

We lead

- Identifier Proliferation
- Commercialization
- Openness/Protection
- Maintenance
- Authority
- Unintended Consequences/Harvesting

BIG DATA AFFECTS DECISION IMPACT

Questions meet data needs

1) Among total article publication listed in the Emerald, how many article written by accountants storied about bibliometrics?

Patterns of co-authorship and research collaboration in Malaysia

Ming Yu Cheng, Kai Wah Hen and Hoi Piew Tan
Faculty of Accountancy and Management, Universiti Tunku Abdul Rahman, Kajang, Malaysia, and
Kuk Fai Fok
ASEAN Academy of Engineering and Technology (AAET), Puchong, Malaysia

Received 11 December 2012
Revised 19 June 2013
Accepted 19 June 2013

Abstract

Purpose – By exploring the patterns of co authorship, this paper aims to identify the degree and type of research collaboration in Malaysia.

Design/methodology/approach – A total of 22,244 publication records from five research universities in Malaysia were retrieved from Scopus database. Journal articles published for the period between 2008 and October 2011 were collected. Indicators such as number of authors, subject areas, number of local institutions and foreign countries, were analysed using simple statistical tools to identity the degree and type of collaboration.

Findings – The findings reveal that in Malaysia, researchers tend to work in teams but collaboration is more dominant in science based research than social sciences. Academics published extensively with their colleagues from the same university or from other academic institutions, but there is little collaboration with researchers from public research institutes or industry. In terms of international collaboration, Iran, India, UK, Japan and the USA are the top five collaborating countries. Disciplines with significant international collaboration are physics and astronomy; chemistry; agricultural and biological sciences; engineering; health profession and computer sciences.

Originality/value – This paper is among the few that study the patterns of co authorship in Malaysia and most probably the first to examine the patterns in the Malaysian research universities. The study highlights the skewed distribution of co authorship patterns where there is limited evidence of cross sectors collaboration in journal publication. The findings call for policy makers as well as universities to look into the constraints as well as drivers that would enhance the linkage of different actors in the national research system.

Keywords Co authorship, Malaysia, Academic research, Bibliometrics, Joint authorship, Scientific collaboration

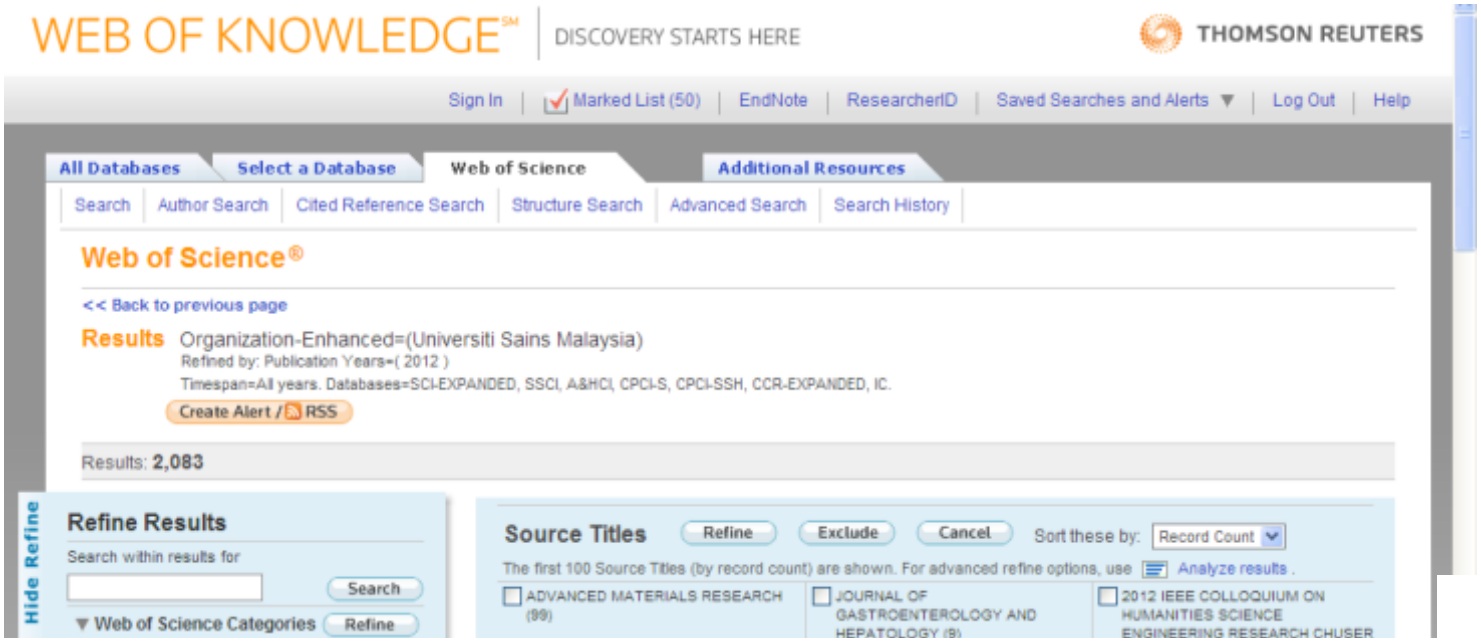
DECISION IMPACT

- SO ON FUTURE TOWARDS BIBLIOMETRICS STUDY, DO GOVERNMENT SUPPORT more BUDGET (RESEARCH GRANTS) TO INFORMATION MANAGEMENT EXPERTS or anybody else that have more interesting findings?

Questions meet data needs in USM

We lead

2) Among total article publication listed in the Web of Science, how many percent USM contribute in 2012?



ANSWER

Results: **2,083**

2,083 records matched your query of the 43,577,624 in the data limits you selected.

Key: = Structure available

2) in 2012 , 0.0047% of total publication listed in Web of Science were from USM

edge.

DECISION IMPACT

We lead

- SO ON 2013, DO UNIVERSITY TOP MANAGEMENT GIVE MORE ON RESEARCH GRANTS? OR LIBRARY?

Questions are meet data needs in USM

We lead

3) Among total BOOKS ON THE SHELVES, which subject are highly use in 2012 and LESS/DROP BORROWED in 2012?

In 2012 , <?? %> of total books on the shelves were in <what subject> are highly use and <?? %> of total books were in <what subject> are less use.

ANSWER

DECISION IMPACT

- SO, on year 2013, WHICH SUBJECT BOOK SHOULD BE PRIORITISE TO BUY?

CONLUSIONS

1. Need more research on Big Data issues.
2. Big Data have impacts to libraries
3. Librarians should study more on Big Data issues.

Thank you

Presented by
Muhammad Akmal Ahmat| PHS