Expert Panel

ALLDATA/MMEDIA/KESA

Wednesday, April 22, 5:30 pm Barcelona

# Big Data Processing
# Can We Control the Value of Lost Data?

**Moderator**

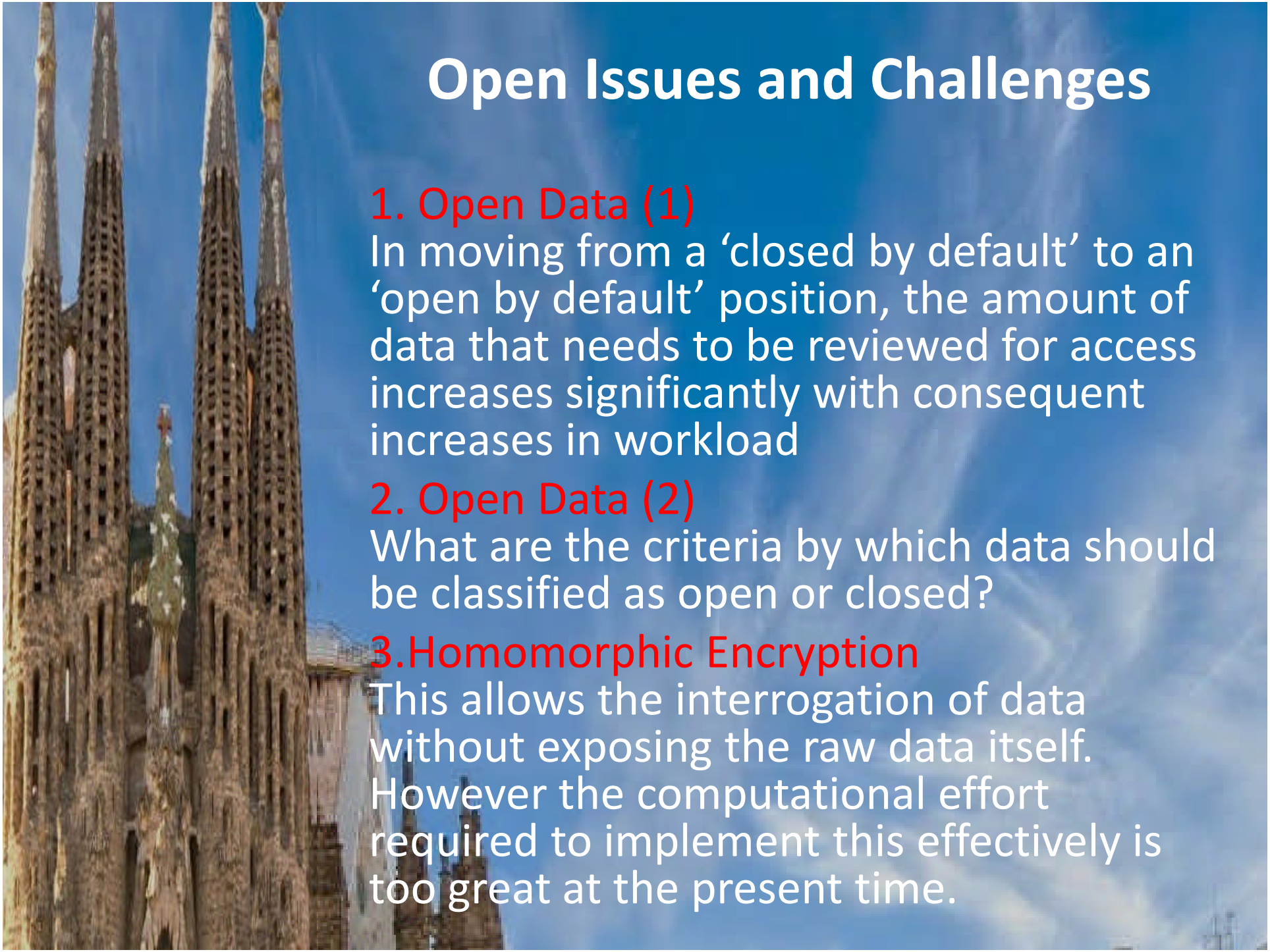
**Philip Davies, Bournemouth University, UK**

**Panelists**

**Venkat N. Gudivada, Marshall University, USA**

**Jolon Faichney, Griffith University, Australia**

**Jerzy Grzymala-Busse, University of Kansas, USA**



# Open Issues and Challenges

1. Open Data (1)

In moving from a ‘closed by default’ to an ‘open by default’ position, the amount of data that needs to be reviewed for access increases significantly with consequent increases in workload

2. Open Data (2)

What are the criteria by which data should be classified as open or closed?

3.Homomorphic Encryption

This allows the interrogation of data without exposing the raw data itself. However the computational effort required to implement this effectively is too great at the present time.

# SHOULD OPEN DATA BE: "OPEN BY DEFAULT"?


Dr Jolon Faichney

School of Information and Communication Technology

Griffith University, Australia

j.faichney@griffith.edu.au


# The motivations behind Open Data

- Governments in democratic countries serve the public:
  - Therefore public should have access to data

- Transparency
- Accountability
- Productivity

# Open Data Institute

- Founded by Tim Berners-Lee and Nigel Shadbolt in 2012
- UK-based, worldwide
- Promote the concept of:

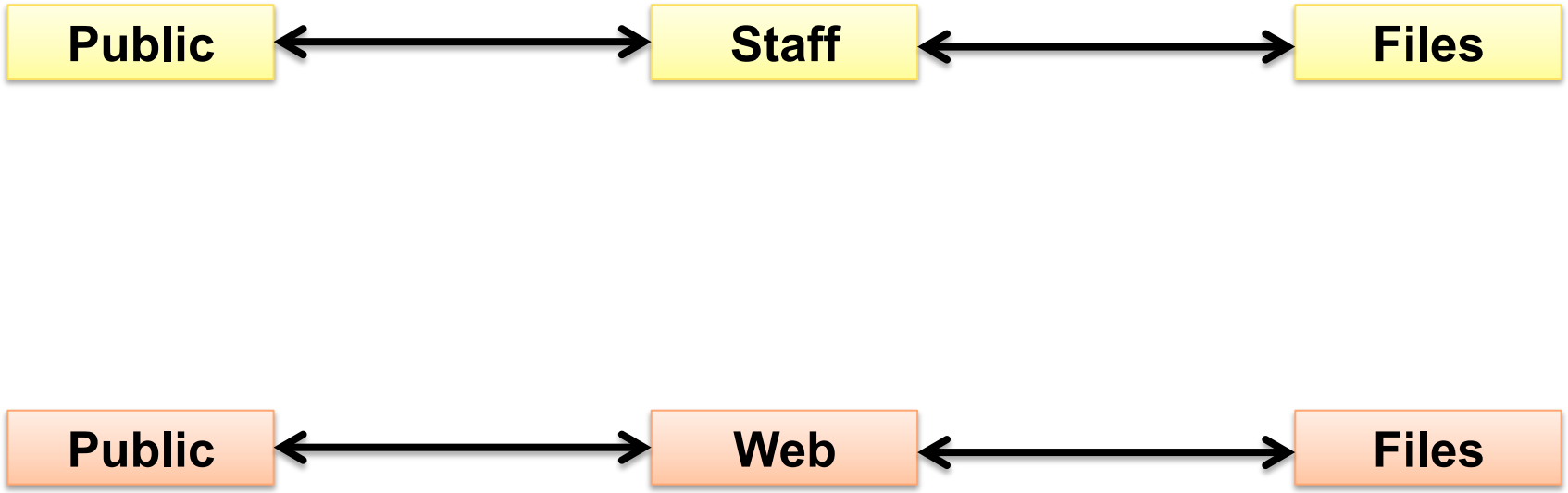
**"Open by Default"**

# Traditional Government Data

# Pre-Web Government Data


Previously, access to Government data required person-to-person interaction

# Post-Web Government Data



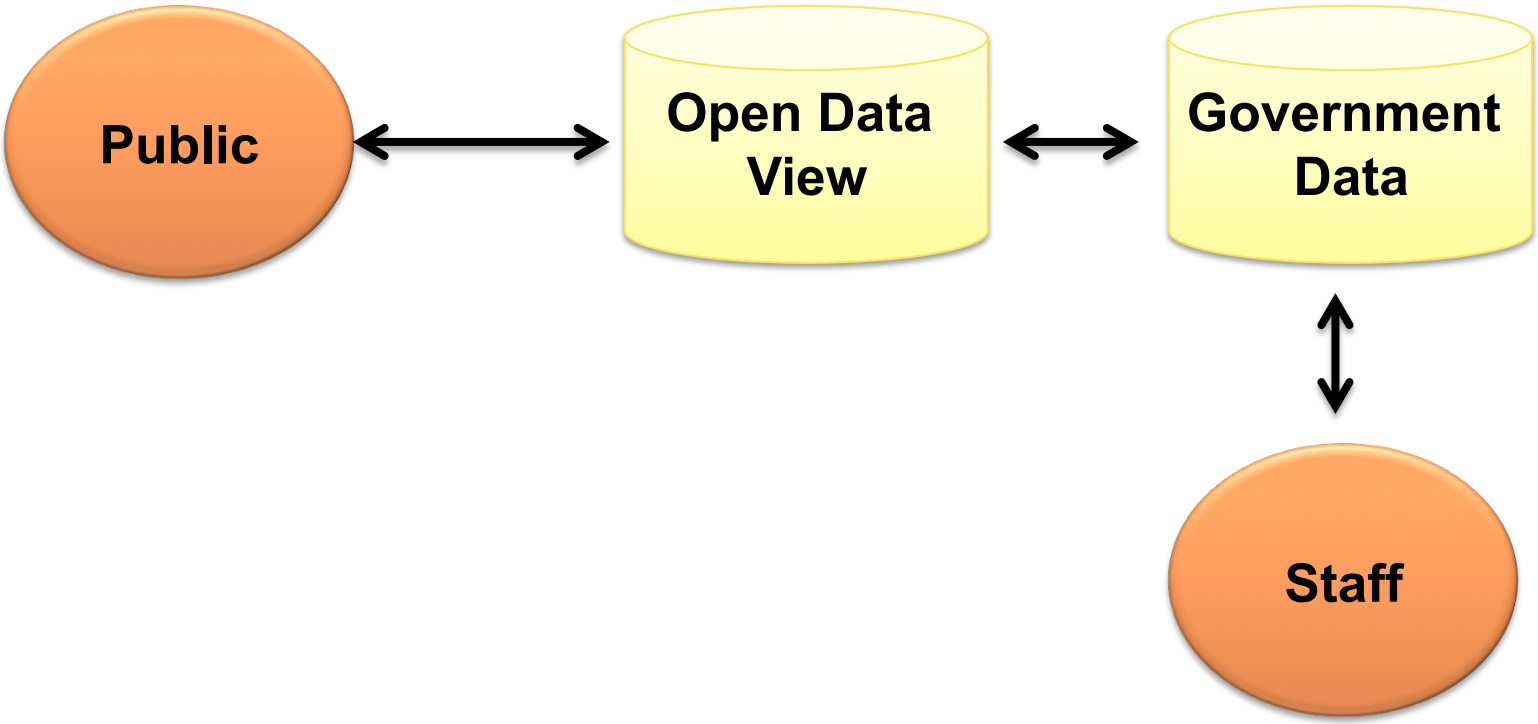
# "Open by Default" Challenges

- What are the criteria for not opening data?
- Since data is open by default, we must assess all data to ensure that no data is being released that shouldn't.
- What resources are required to make data open?
- De-identifying data:
  - Remove/obscure identifying attributes
    - Risks
  - Provide statistical summaries

# An Information Systems Approach

- Current approaches to Open Data view data in the traditional paper model:
  - Documents that must be made available
- However these documents rarely exist.
- Traditional paper documents are now views/reports on (mostly) relational databases.
- Should we instead provide access to the relational data and allow clients to form their own queries?
- Databases must be designed from the beginning with Open Data in mind.

# An Information Systems Approach

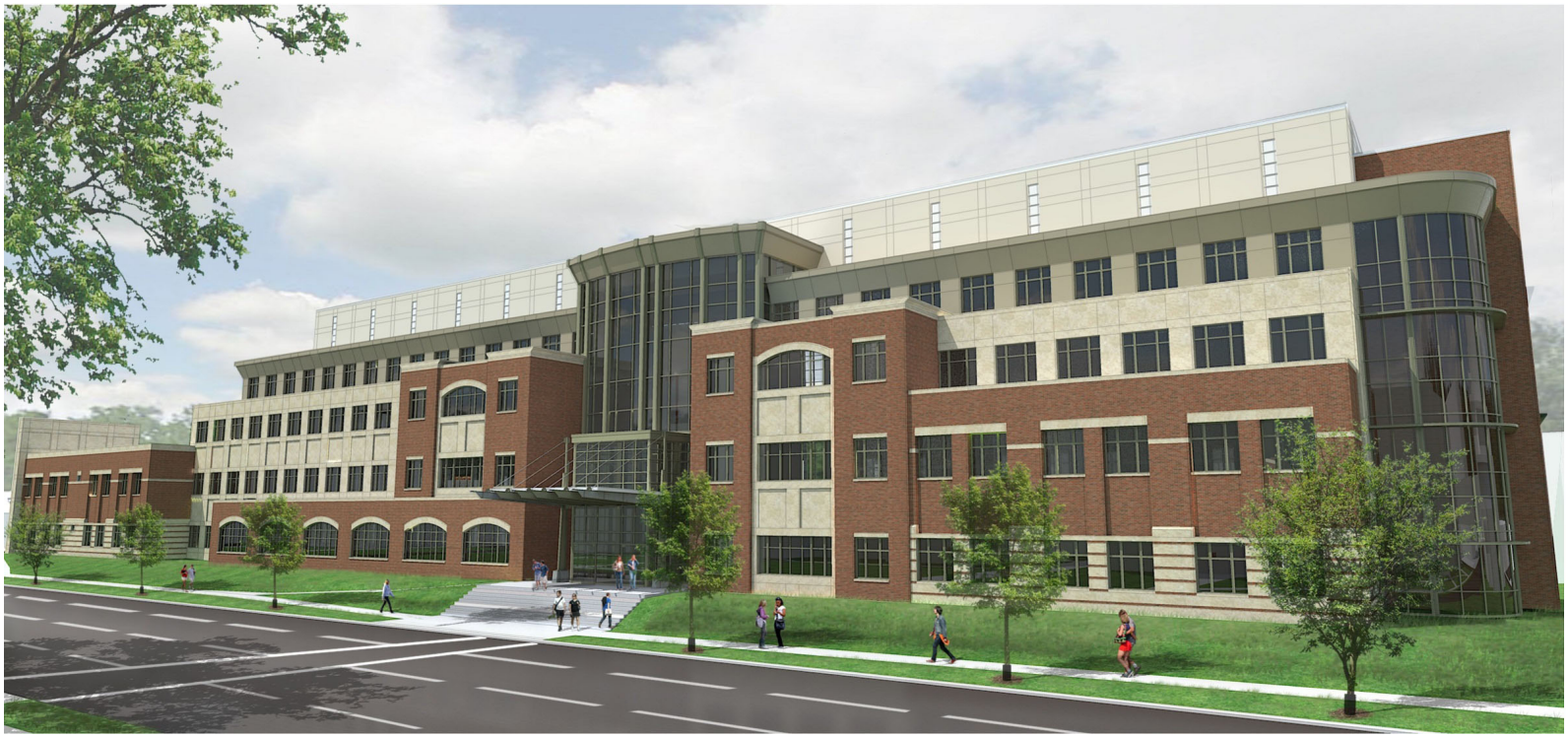


# Topics for Discussion

- Is "Open by Default" too ambitious/impractical?
- What mechanisms can we use to de-identify data?
  - Are these flawed?
- Can we allow the public access to government relational databases with Open Data interfaces?

# Big Data Processing: Can We Control the Value of Lost Data?

Venkat N Gudivada
Weisberg Division of Computer Science
Marshall University
Huntington, WV, USA



# Big Data: Promises and Problems

- Big Data has potential for groundbreaking scientific discoveries, business innovation, and increased productivity.
- Provides as many challenges as the number of new opportunities it ushers in.
- Several problems need solutions before the full potential of Big Data is realized.
- March 2015 theme issue of IEEE Computer - Big Data: Promises and Problems.
- In 2014, the White House commissioned a study to examine how Big Data will transform the way we live and work.

# Five V's of Big Data

- Volume
- Velocity
- Variety
- Veracity
- Value

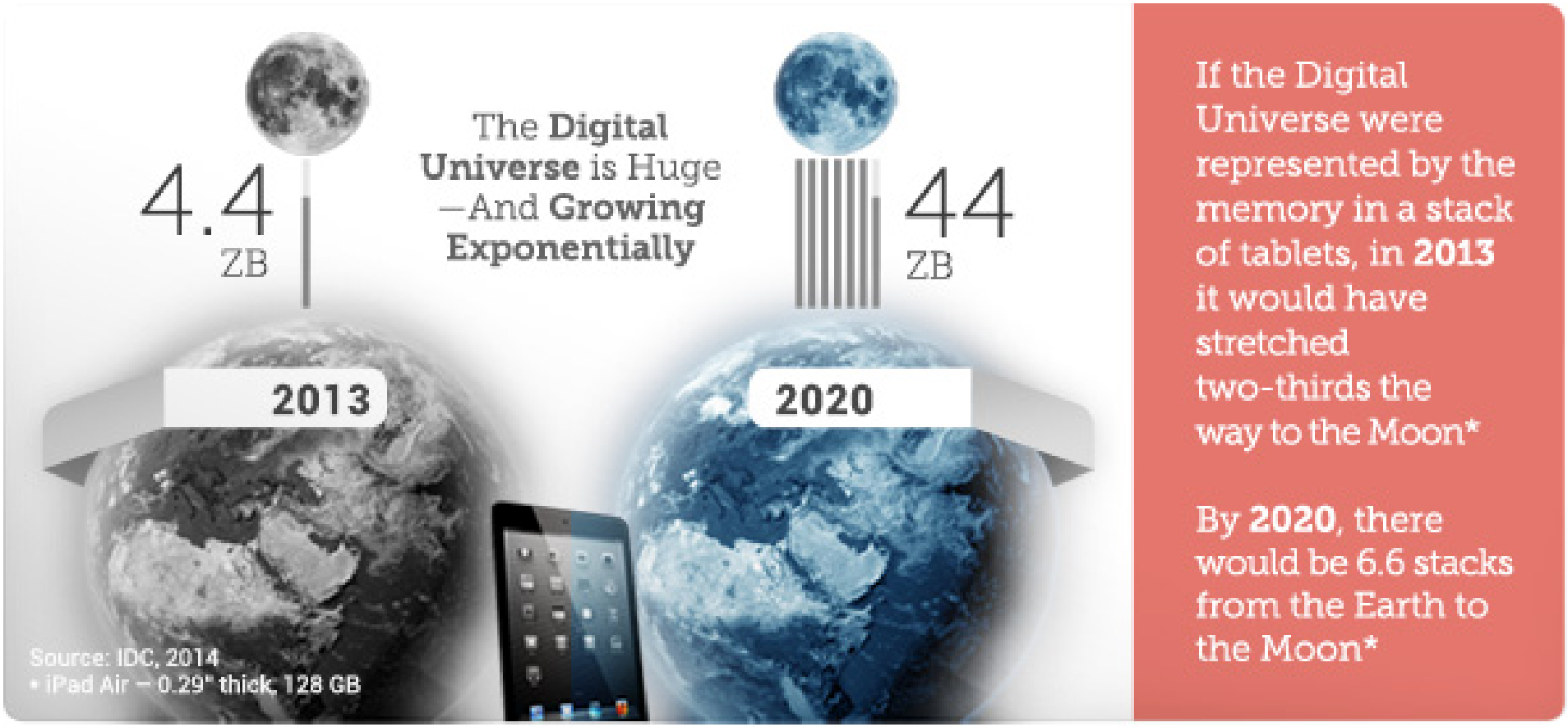
# How is Data Generated?

- Supercomputers and supercolliders.
- Smart phones and other hand-held devices.
- Internet of Things (IoT).
- Wireless sensor and camera networks.
- Earth-orbiting satellites.
- Social media applications.

# Computer is the $21^{st}$ Century Laboratory

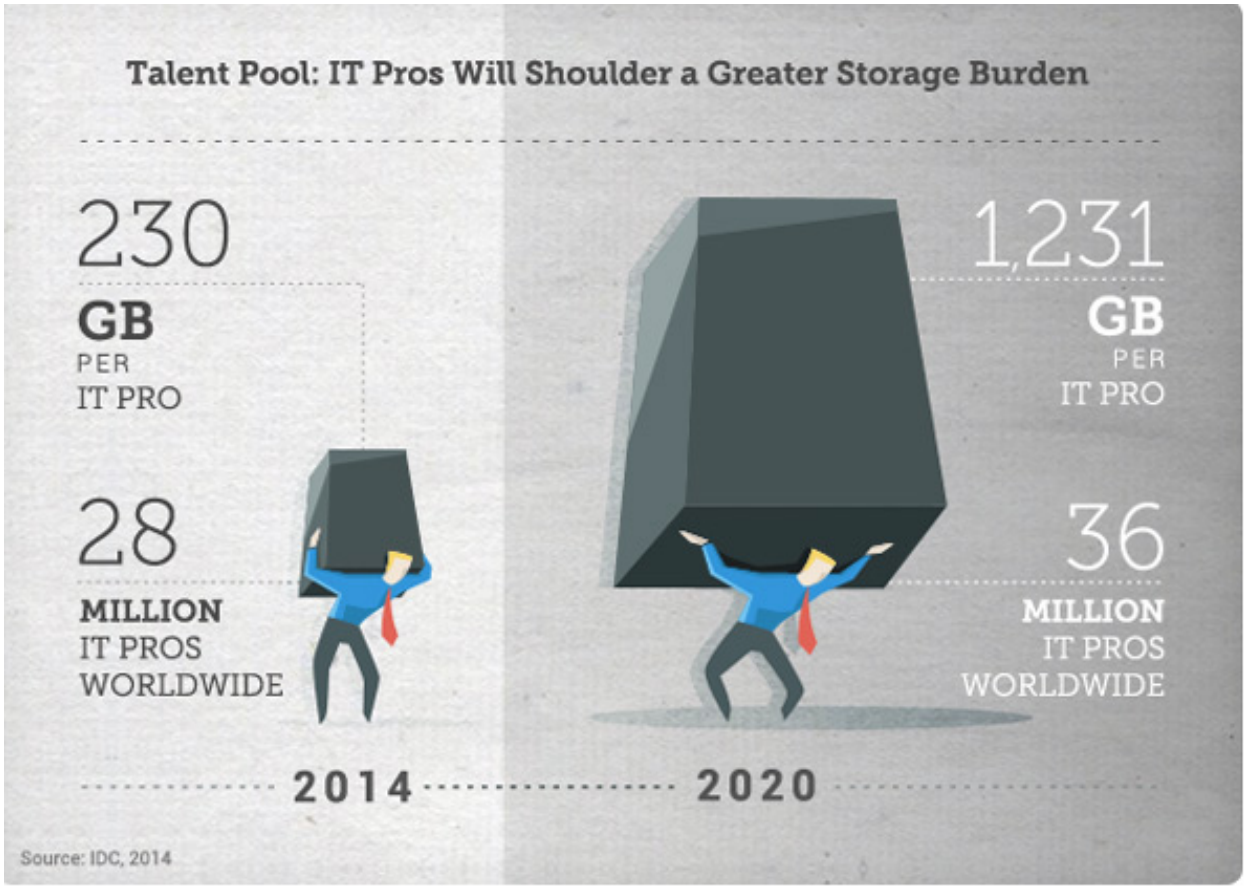
- 2013 Nobel Prize in chemistry – involved measuring and visualizing the behavior of 50,000 or more atoms in a reaction over the course of a fraction of a millisecond.
- Large Hadron Collider machine – there are 150 million sensors capturing data about nearly 600 million collisions per second.
- Square Kilometer Array (SKA) radio telescope project – will produce 2.8 gigabytes of astronomy data per second – to create the biggest map of the Universe ever made.

# Big Data Growth




http://www.emc.com/leadership/digital-universe/2014iview/

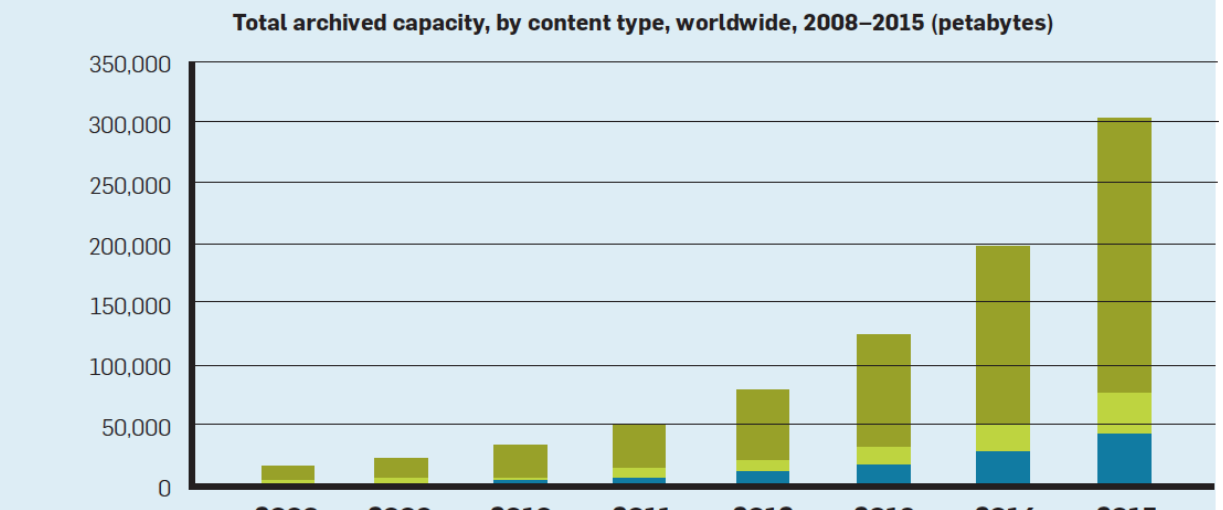
# Big Data Burden on IT Professionals




http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

# Growth of Unstructured and Structured Data

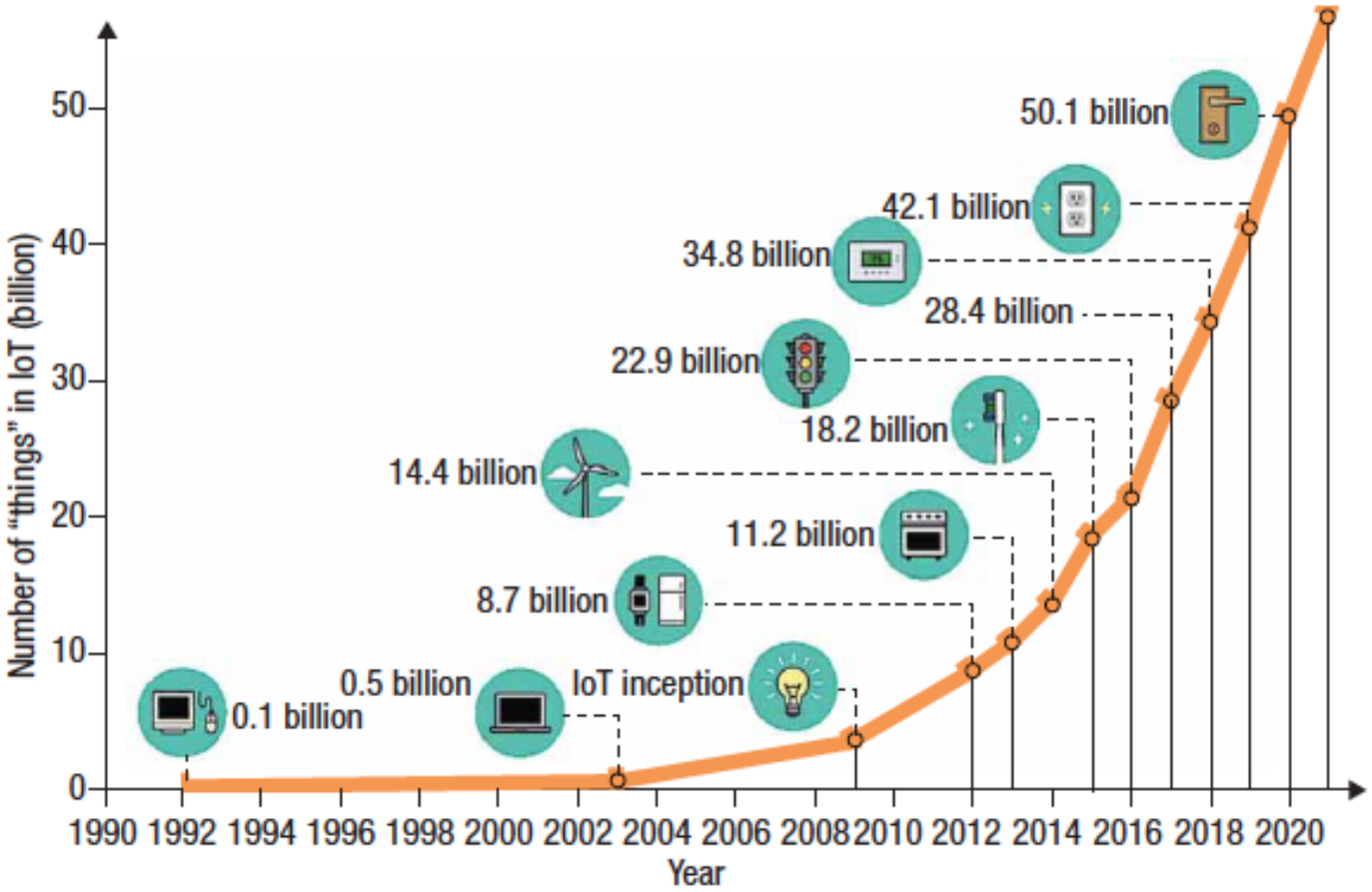
**Figure 1. Projected growth of unstructured and structured data.**




| | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|
| Unstructured | 11,430 | 16,737 | 25,127 | 39,237 | 59,600 | 92,536 | 147,885 | 226,716 |
| Database | 1,952 | 2,782 | 4,065 | 6,179 | 9,140 | 13,824 | 21,532 | 32,188 |
| Email | 1,652 | 2,552 | 4,025 | 6,575 | 10,411 | 16,796 | 27,817 | 44,091 |

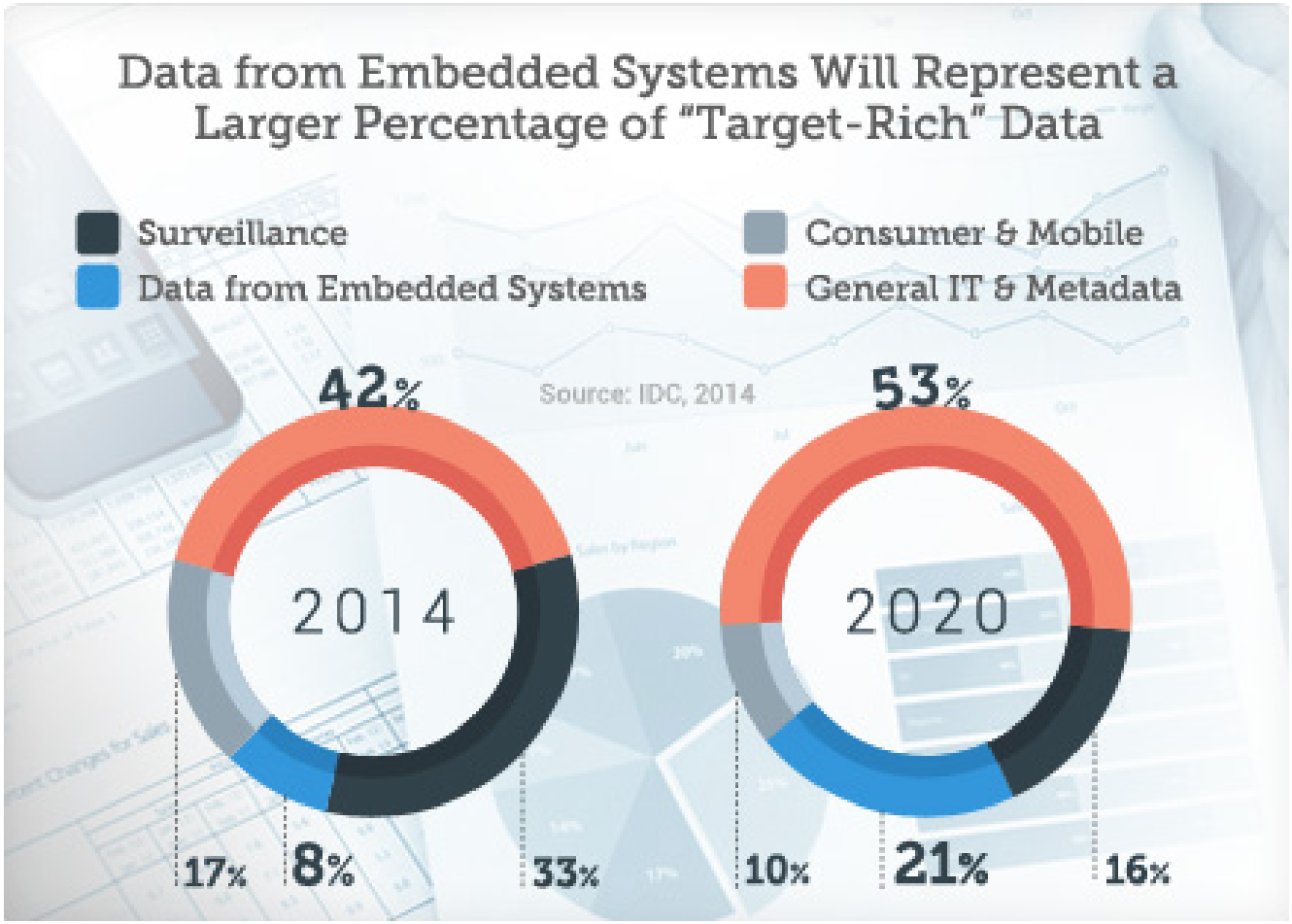
Vasant Dhar. *Data Science and Prediction*, Comm. of the ACM, 56(12), pp. 64 - 73, Dec, 2013.

# IoT Growth



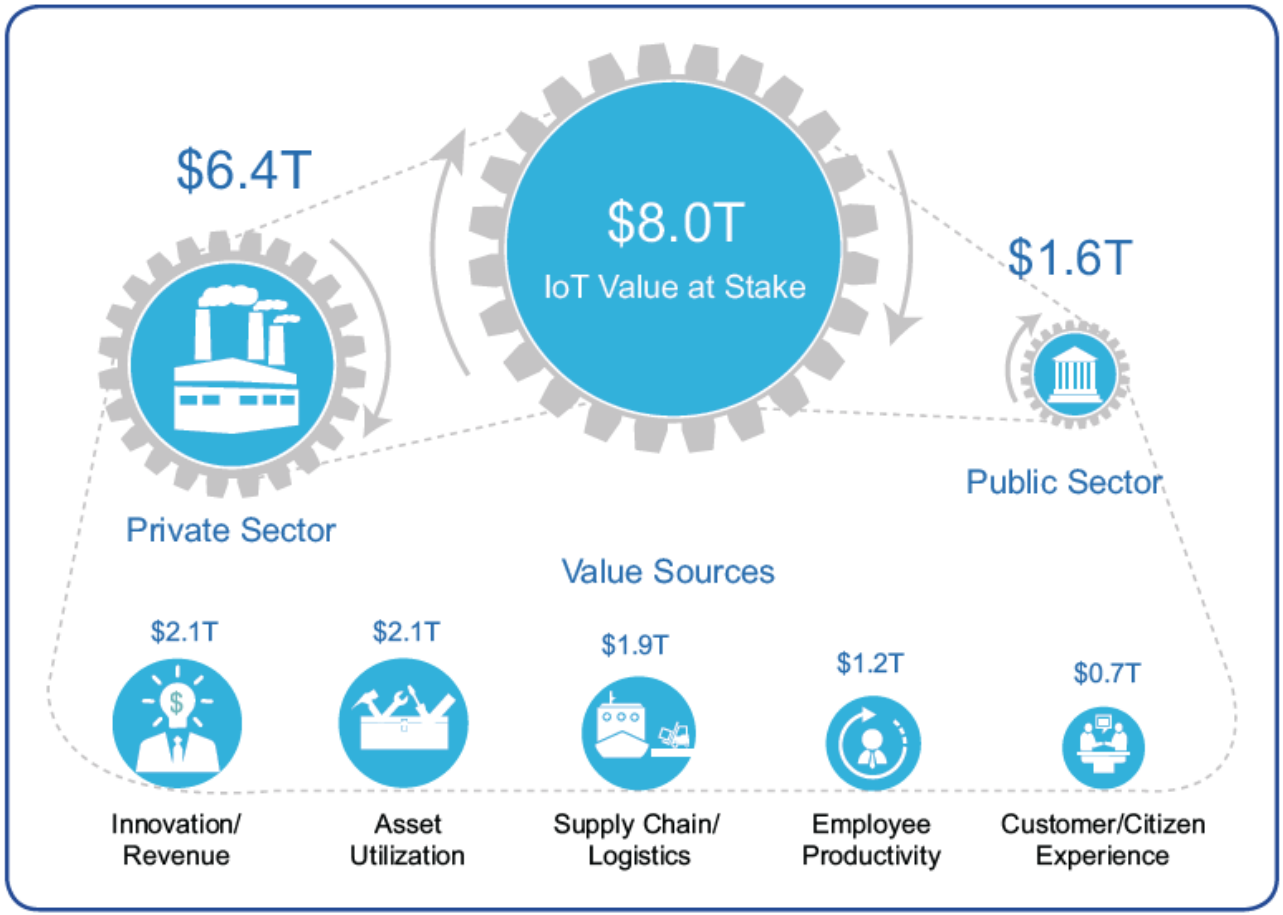

Peter Fonash and Phyllis Schneck (US Department of Homeland Security). Cybersecurity: From Months to Milliseconds. IEEE Computer, January 2015, pp. 42-50.

# High Value Data




http://www.emc.com/leadership/digital-universe/2014iview/high-value-data.htm

# IoT Value




Source: Cisco Consulting Services, 2014

http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

# IoT Value

- Precision Agriculture
- Intelligent Transportation Systems
- Smart Cities
- Home and Building Automation
- Energy (Generation and Distribution) - Smart Grid
- Environmental Monitoring
- Infrastructure Management
- Manufacturing
- Medical and Healthcare Systems

# Big Data Value Costs

- Data acquisition.
- Data quality.
- Data provenance.
- Meta data and semantic annotations.
- Access control and differential privacy.
- Perceived vs. real value

# Big Data Value Questions

- What data is available?
- What data is useful for my context?
- What are the costs of data errors?
- How do I deal with incomplete or missing data?
- How do I detect duplicate data?
- How do I cross-link data from multiple vendors?
- How do I maintain data validity and consistency across recent and older datasets?

# Interpretations of Missing Attribute Values (A Rough Set Approach)

**Jerzy W. Grzymala-Busse'"**

' University of Kansas, Lawrence, KS 66045, USA

" Department of Expert Systems and Artificial Intelligence,

University of Information Technology and Management, 35-225 Rzeszow, Poland

# Incomplete Data Sets

Three interpretations of missing attribute values:

*lost values*

*attribute-concept values*, and

*"do not care" conditions*

# Lost Values

We assume that the original attribute value is lost,

e.g., was erased,

and that we should induce rules form existing,
specified attribute values

# Attribute-concept Values

Such missing attribute values may be replaced by any actual attribute value restricted to the concept to which the case belongs.

If our concept is a specific disease, e.g., a diastolic pressure, and all patients affected by the disease have high or very high diastolic pressure, a missing attribute value of the diastolic pressure for a sick patient will be high or very-high.

# “Do not care” Conditions

It does not matter what is the attribute value.

Such value may be replaced by any value from the set of all possible attribute values.