# *A Common Sense Approach to Defining*

# *Data, Information, and Metadata*

Dimitris A. Dervos
*I.T. Dept., Alexander Technology Educational Institute,*
*Thessaloniki, Greece*
*(**dad@it.teithe.gr**)*

Anita Coleman
*School of Information Resources & Library Science,*
*University of Arizona,*
*Tucson, USA*
*(**asc@u.arizona.edu**)*

# Defining Data

- Data have no meaning or value, because they are without context and interpretation

- Data are discrete, objective facts or observations, which are unorganized and unprocessed, and do not convey any specific meaning

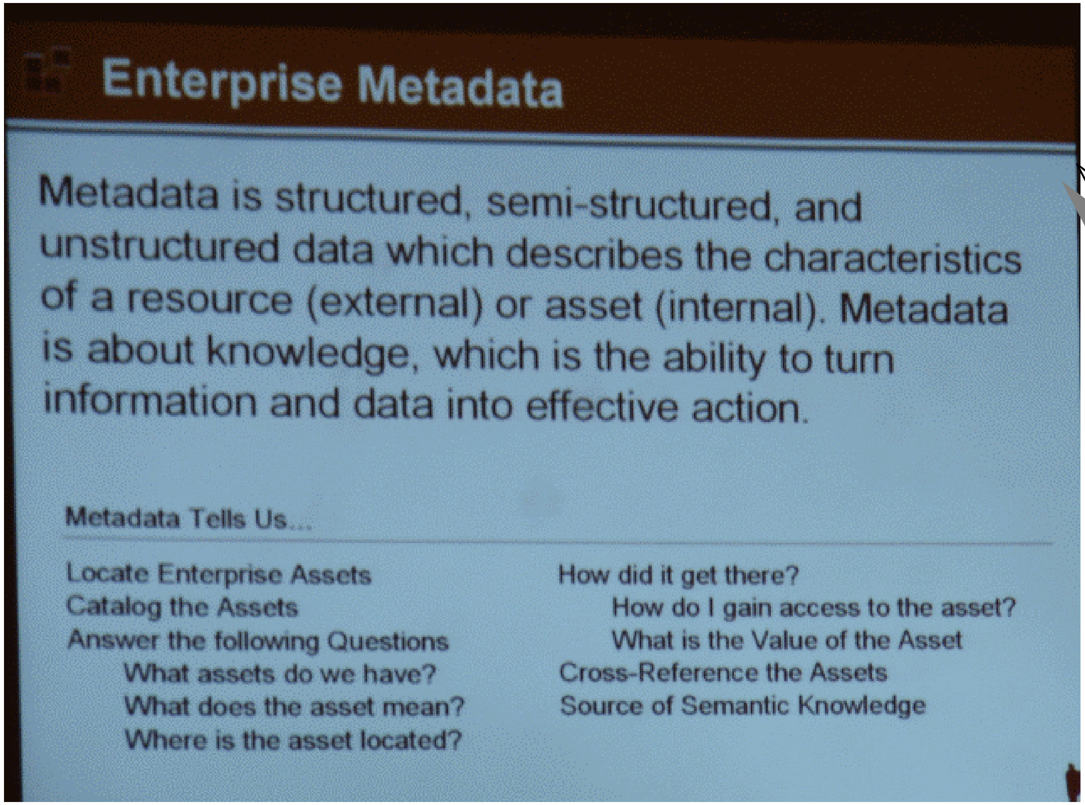
- Data items are an elementary and recorded description of things, events, activities and transactions

- Data are the raw material for building **information**

# Defining Information

- Information **is formatted data**… defined as a representation of reality
- Information **is data** which adds value to the understanding of a subject
- Information **is data** that have been shaped into a form that is meaningful and useful to human beings
- Information **is data** that have been organized so that they have meaning and value to the recipient
- Information is any physical form of representation (or surrogate) **of knowledge**
  (Faradane)

# Divergence across disciplinary perspectives



## Enterprise Metadata

Metadata is structured, semi-structured, and unstructured data which describes the characteristics of a resource (external) or asset (internal). Metadata is about knowledge, which is the ability to turn information and data into effective action.

Metadata Tells Us...

Locate Enterprise Assets
Catalog the Assets
Answer the following Questions
- What assets do we have?
- What does the asset mean?
- Where is the asset located?

How did it get there?
- How do I gain access to the asset?
- What is the Value of the Asset

Cross-Reference the Assets
Source of Semantic Knowledge

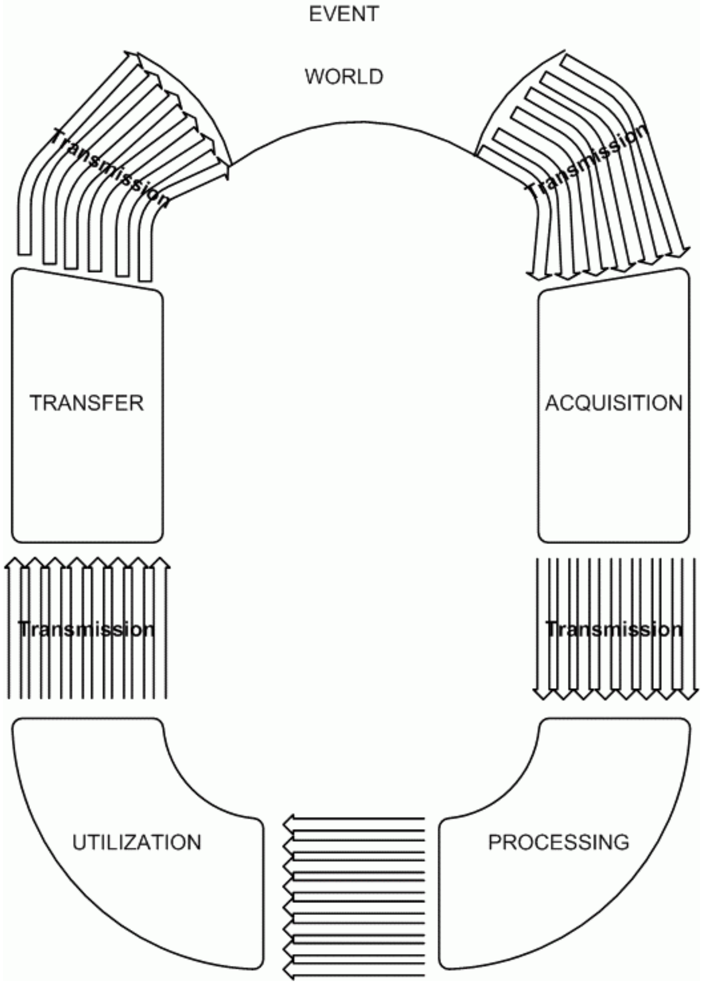
*Anonymous, **2005 International Conference on Knowledge Management**, Charlotte, N.C., U.S.A.*

… gets in the way of true interdisciplinary collaboration between computer scientists, library/information scientists, etc.

# Debons' "living species" approach

## Two Preconditions

1. Social/organizational systems are not addressed at this stage. Technology is ignored. The focus is on the individual living organism.
2. Reasoning builds upon a finite number of simple assumptions made initially



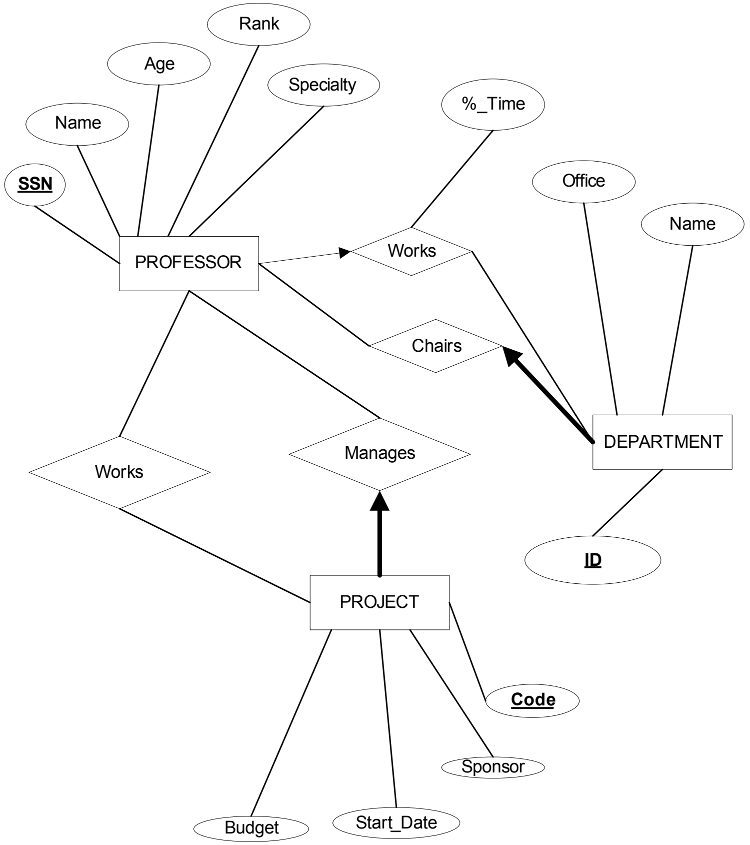
# Human Mind (HM): Favors Associations

Five years after Iro
Fifth day of month
One month earlier…

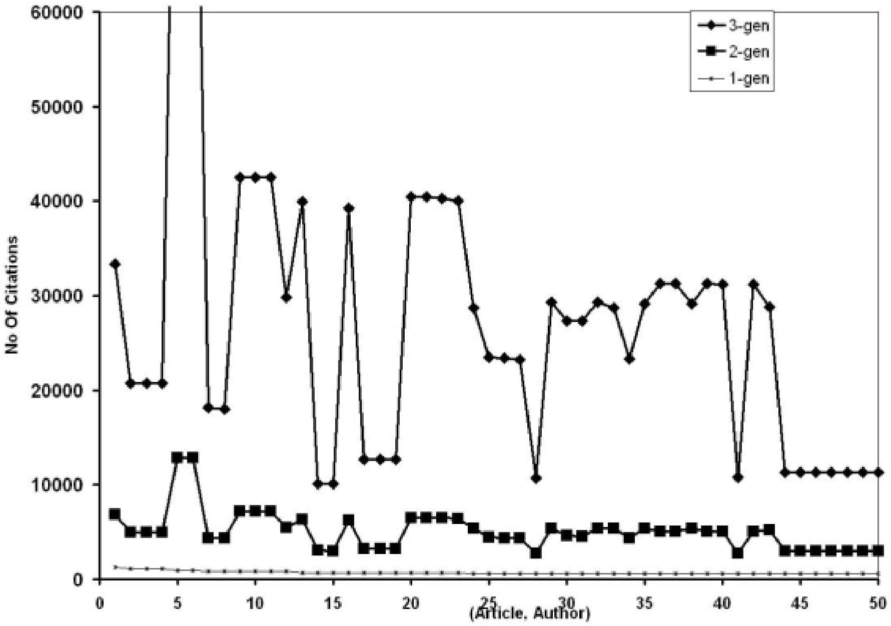
| Name | Birthday | Day of Birth |
|---|---|---|
| Antonis | 05 April 1988 | Monday |
| Claudine | 15 February 1991 | Thursday |
| David | 27 November 1982 | Saturday |
| Iro | 05 May 1983 | Thursday |
| Jaakko | 13 July 1995 | Thursday |
| Jose | 8 January 1993 | Thursday |
| Kathleen | 24 March 1980 | Monday |
| Maria | 30 December 1982 | Thursday |

Fifth day of week
Fifth day of month
Fifth month of year

# HM: Favors Shortcuts



ER Diagram



Graph

# HM: Favors Normalization

Problem 1:

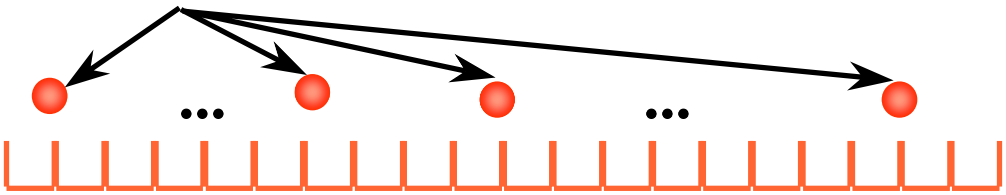
Select 20 people at random. What is the probability that two of them celebrate their birthdays on the same day in the year?

Problem 2:

365 “killer” traffic junctions. What is the probability of having two accidents take place at the same junction for every 20 accidents that occur?

Same Model (One Solution):

Place 20 balls into 365 “urns”, with replacement

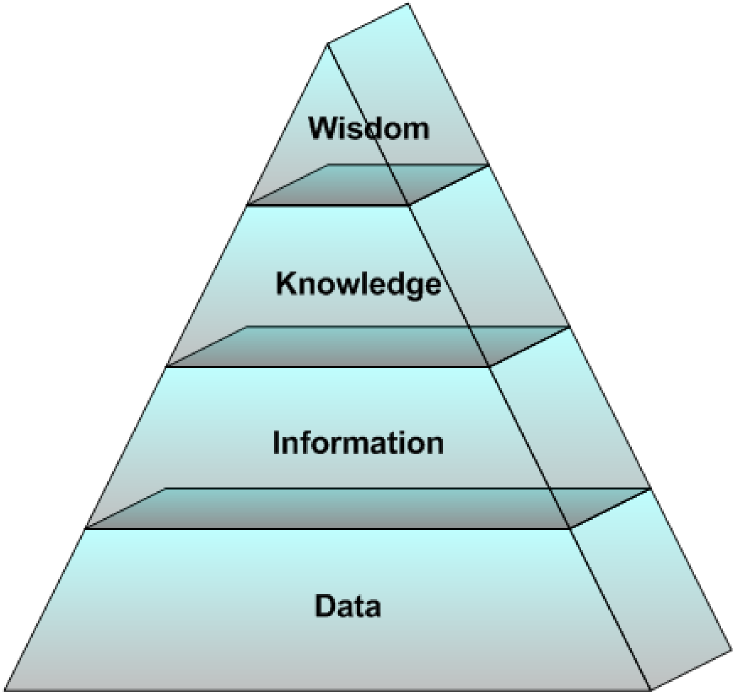


# The Common Sense Approach (CSA): Two Assumptions

1. Information is at a higher level than data
2. Knowledge is at a higher level than information

…as in the DIKW hierarchy

(*R.L. Ackoff, From Data to Wisdom,*
***Journal of Applied Systems Analysis,***
*16, 3-9, 1989*)



# CSA: Data

**Definition:**
***Data Represent Real World Facts***

**Examples:**
Rainfall measurements over time, for a given set of geographical regions,
Attribute values registered with a database application,
etc.

# CSA: Information

**Definition:**

***Information is revealed each time data are interpreted successfully in the direction of increasing benefit, profit, or pleasure, as the latter are realized by some intellectual activity***

**Examples:**

A plot that occupied 1/3 of an A4 page,
An ER diagram,
A rule, e.g. *Heavy smokers have a high probability to develop lung cancer,*
An association, e.g. *People who purchase potato chips tend to also purchase beer,*
Raw data, e.g. *Global Getaways offer their Chalkidiki package for 1/3 of the regular price, this week*

# CSA: Metadata

**Definition:**

***Metadata are tags/labels assigned to data instances and structures that make them comprehensible and/or facilitate the processing that extracts information from data corpora***

**Examples:**

*Student ID*, *Department*, *Year of Entry*, *Course ID*, *Grade*, etc.
*Format*, *Form*, *Creator*, *Title*, etc.
*Process*, *Object*, *Phenomenon*, etc.

# CSA Discussion: Data

1. Concept well defined, fully understood
2. Quantified: size remains invariant from system to system, provided the representation technology remains the same
3. Unit of measurement: the bit
4. In the developed part of the world, today: everyday human activity is shaped, to a great extent, by technology-assisted data storing/processing /management operations

# CSA Discussion: Information

1. Concept realized only indirectly, not directly
2. Not yet quantified, except from in special cases.
3. Unit of measurement: ? *(infotron?)*
4. Everyday human activity is still far from utilizing technology in a way that machines: (a) model user interests/preferences, (b) sense the current context of the user, (c) compute information relevant to the context and user profile, and (d) proactively offer 'just-in-time' information in a subtle, non-intrusive way[1]

# CSA: Two Corollaries

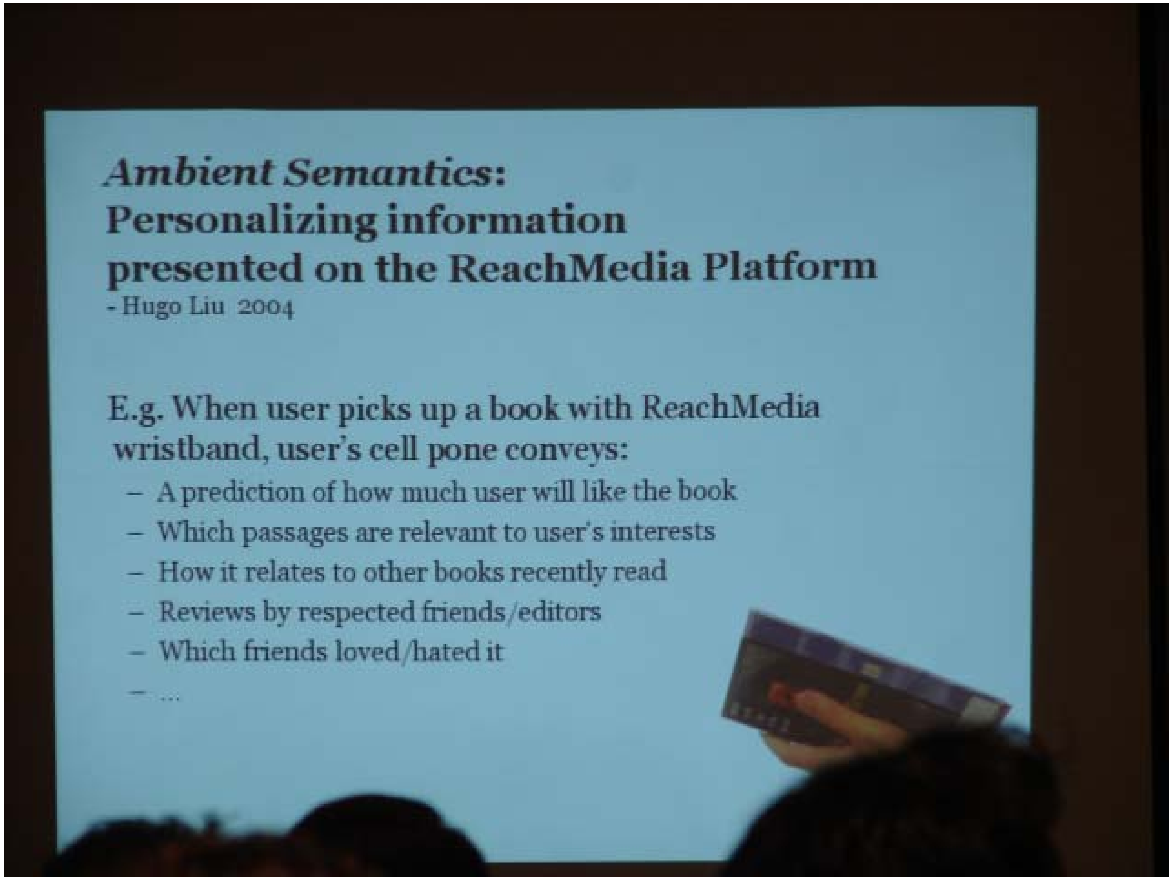
**Corollary-1**

*Information remains to be quantified, modeled, and be fully understood as a concept*

**Corollary-2**

*The forefront of our civilization is still in the data processing age. More research and technology advances are required for the information society to come of age*

# CSA: The Information Society



*Ambient Semantics:*
**Personalizing information presented on the ReachMedia Platform**
- Hugo Liu 2004

E.g. When user picks up a book with ReachMedia wristband, user's cell pone conveys:

- A prediction of how much user will like the book
- Which passages are relevant to user's interests
- How it relates to other books recently read
- Reviews by respected friends/editors
- Which friends loved/hated it
- ...

[1] *P. Mayes, Just-in-Time Information,*
*Key Note Speaker,* ***ASIS&T 2005 Annual Meeting****,*
*Charlotte, N.C., November, 2005*

# Conclusion

- Competing and divergent definitions for data, information, and metadata are noted to exist today
- The above get in the way of true interdisciplinary collaboration
- As a consequence, they hinder the development of systems that can truly move us into the information society era

A Common Sense Approach is adopted that meets the two preconditions set forth by Prof. Debons:

(a) A '*living species*' approach is established, and

(b) Reasoning builds upon a finite number of simple assumptions made initially

Data, Information, and Metadata are defined

in a way making it clear that:

(a) We are still in the data processing era

(b) The concept of information remains to be fully understood and quantified/measured

(c) Concept(s) in higher levels (e.g.: *knowledge*) remain to be defined.

# Thank You for Your Attention!

# Information as Entropy

Appendix

# Appendix: A Special Setup

**The event:** a bit of data arrives

(a) how much information is there to the given event?
(b) how much knowledge does an observer have with regard to the bit's (1/0) value PRIOR to seeing it?

**Two (extreme) cases:**

(a) The event is unbiased (i.e. unclassified): the bit carries the maximum possible amount of information, and it is equally probably for the observer to see a '1' or a '0'
(b) The event is 100% biased (classified): the observer knows the bit's value prior to seeing it: the bit carries no information (it does not need to arrive, actually)

# Appendix: Information as Entropy

Let $I[p_1, p_2]$ be the function that calculates information:

- $p_1$ is the probability the bit has to be a '1'
- Analogously, $p_2$ is its probability to be a '0'
- Constraint: $p_1+p_2=1$
- $I[1,0]=I[0,1]=0$ ← requirement-1
- $MAX(I[p_1,p_2]) = I[0.5,0.5]$ ← requirement-2

Generalization (say: for three possible states)

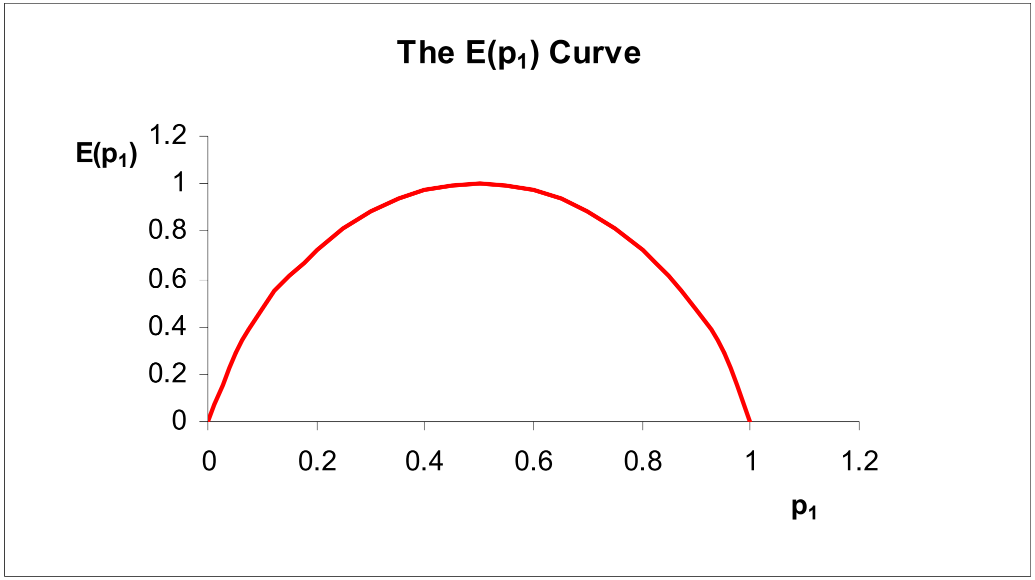
- $I[p_1, p_2, p_3]$
- $p_1+p_2+p_3=1$
- $I[1,0,0] = I[0,1,0] = I[0,0,1] = 0$ ← requirement-1
- $MAX(I[p_1,p_2,p_3]) = I[1/3, 1/3, 1/3]$ ← requirement-2
- Multistage decision property ← requirement-3:
  $I[p_1,p_2,p_3] = I[p_1,p_2+p_3] + (p_2+p_3)\ I[(p_2/(p_2+p_3), p_3/(p_2+p_3)]$

It turns out that entropy is the only one function that meets all three set requirements:

- **$\mathrm{entropy}(p_1,p_2,\ldots,p_n) = -p_1\log p_1 - p_2\log p_2 - \ldots - p_n\log p_n$**

# Appendix: Information is Measured in bits!

- Back to the one-bit event
- Calculating the logarithm in base 2
- $p_1$: the probability for the event to turn out having the value '1'



When the event is biased (classified) to turn out to be 1(0) with a probability of 20%(80%), then it carries information equivalent to that of a 0.72 (un-biased) bit.

The 20%(80%) bias is said to represent an *information gain* of 0.28 bits for the observer.

# References (for definitions – slides 2 and 3)

*E.M. Awad & HM Ghaziri,* ***Knowledge Management****, Pearson Educational International, Upper Saddle River, NJ, 2004*

*P. Bocij, et al.,* ***Business Information Systems: technology, development and management for the e-business****, 2nd Ed., FT Prentice Hall, Harlow, 2003*

*D. Buddy, A. Boonstra, and G. Kennedy,* ***Managing information systems: an organizational perspective****, 2nd Ed., FT Prentice Hall, Harlow, 2005*

*D. Chaffey & S. Wood,* ***Business information management: improving performance using information systems****, FT Prentice Hall, Harlow, 2005.*

*T.R. Croft and T.P. Jones,* ***Introduction to Knowledge Management: KM in business****, Butterworth Heinemann, Amsterdam, 2003*

*J. Faradane, The Nature of Information,* ***Journal of Information Science****, 1: 13-17, 1979*

*L.M. Jessup & J.S. Valacich,* ***Information Systems Today****, Prentice Hall, Upper Saddle River, NJ, 2003*

*K.C. Laudon & J.P. Laudon,* ***Management information systems: managing the digital firm****, 9tth Ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2006*

*F. McCrank,* ***Historical Information Science****, Medford, N.J.: Information Today, 2002*

*K.E. Pearson and C.S. Saunders,* ***Managing and using information systems: a strategic approach****, Wiley, New York, 2004*

*F. Turban, R.K. Rainer, and R.E. Potter,* ***Introduction to information technology****, 3rd Ed., New York, Wiley, 2005*