

# Information, Databases

Szabolcs Osváth

Semmelweis University  
szabolcs.osvath@eok.sote.hu

## Definitions



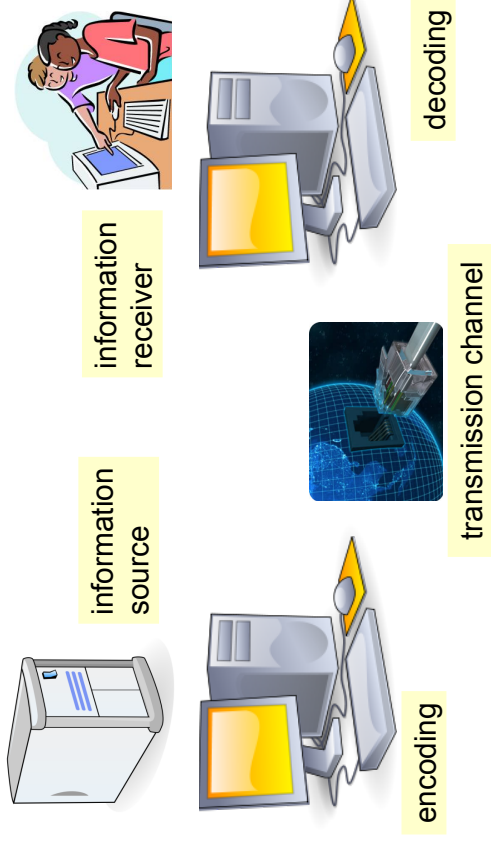
*informare* (Latin): "to give form to the mind", "to instruct"

**information:** knowledge communicated or received concerning a particular fact or circumstance

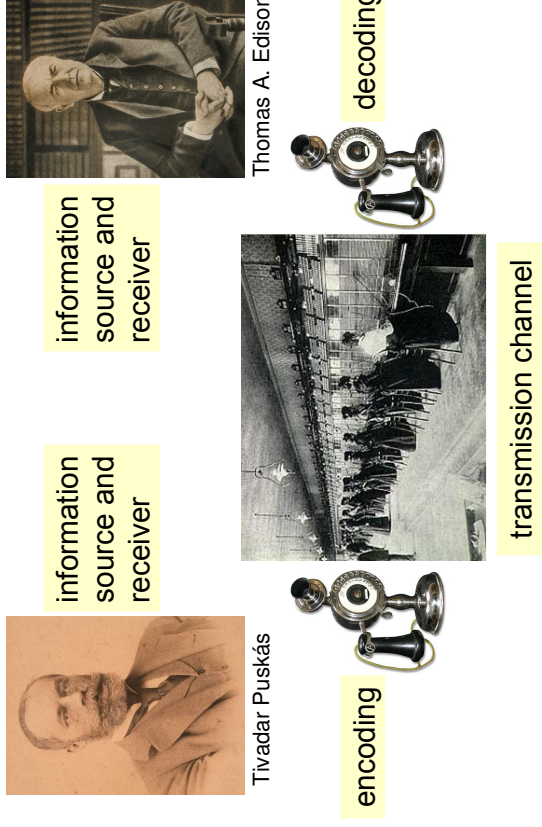
**database:** a comprehensive collection of related information organized for convenient access, generally in a computer

**information literacy:** the ability to know when information is needed and to be able to locate, evaluate, and use effectively the needed information

## Transmission of information



## Transmission of information



## Encoding information about events

**analog encoding:** use a continuous range of values to represent information

**digital encoding:** discrete values represent information; most often representation is done as binary numbers

decimal: 2011

$$2 \cdot 10^3 + 0 \cdot 10^2 + 1 \cdot 10^1 + 1 \cdot 10^0 = 2011$$

binary: 11111011011

$$1 \cdot 2^{10} + 1 \cdot 2^9 + 1 \cdot 2^8 + 1 \cdot 2^7 + 1 \cdot 2^6 + 0 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 2011$$



## Examples for information encoding

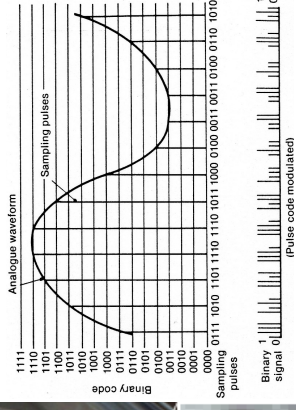
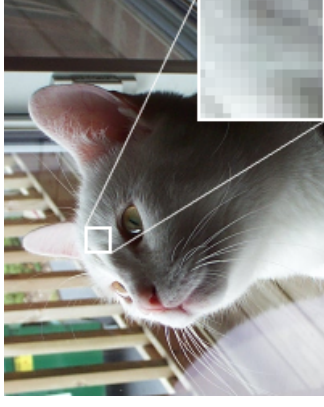
Analog encoding:

- receptor potential coding stimulus strength
- hormone concentration coding physiological messages
- photograph slide
- LP music records
- movie on videotape

Digital encoding:

- action potential coding stimulus strength
- genetic code in DNA
- photo stored on computer
- music record on CD
- movie on DVD

## Digitization



## Encoding information about events



:	0	0	
:	1	1	
:	1		
:	2		
:	3		
:	4		
:	5		
:	6		
event	number	binary code	
	1	001	
	2	010	
	3	011	
	4	100	
	5	101	
	6	110	



## Quantifying information

Self-information is a measure of the information content associated with a given outcome (event) of a random variable.

Assume that  $A$  and  $B$  are two independent events.

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

The amount of information of the proclamation that  $A$  and  $B$  both happened equals the sum of the amounts of information at proclamations of event  $A$  and event  $B$ :

$$I(A \text{ and } B) = I(A) + I(B)$$

## Quantifying information (examples)

learning the result of a coin toss

$$P(\text{heads}) = P(\text{tails}) = 1/2; \quad I(\text{heads}) = I(\text{tails}) = -\log_2(1/2) = 1 \text{ bit}$$

reading one binary digit from a disc

$$P(0) = P(1) = 1/2; \quad I(0) = I(1) = -\log_2(1/2) = 1 \text{ bit}$$

learning the result of a roll with a dice

$$\begin{aligned} P(1) &= P(2) = P(3) = P(4) = P(5) = P(6) = 1/6 \\ I(1) &= I(2) = I(3) = I(4) = I(5) = I(6) = -\log_2(1/6) = 2.584... \text{ bit} \end{aligned}$$

learning the result of the lottery

$$\begin{aligned} P(\text{win}) &= 1/13,983,816; \quad I(\text{win}) = -\log_2(1/13,983,816) = 23.737... \text{ bit} \\ P(\text{no win}) &= 1 - 1/13,983,816; \quad I(\text{no win}) = -\log_2(1 - 1/13,983,816) = -0.000000103... \text{ bit} \end{aligned}$$

## Quantifying information

Self-information is a measure of the information content associated with a given outcome (event) of a random variable.

The self-information content of the occurrence of an event  $E$  which has the probability  $P(E)$ :

$$I(E) = -\log_2(P(E)) \text{ bit}$$

$$I(E) = -\ln(P(E)) \text{ nat}$$

$$I(E) = -\log_{10}(P(E)) \text{ hartley}$$

$$1 \text{ byte} = 8 \text{ bit}$$

## Information entropy

In information theory, entropy is the expected value of the self-information of a variable.

It is a measure of the uncertainty associated with the variable.

$$H(E) = -\sum P \cdot \log_2(P) \text{ bit}$$

$$H(E) = -\sum P \cdot \ln(P) \text{ nat}$$

$$H(E) = -\sum P \cdot \log_{10}(P) \text{ hartley}$$

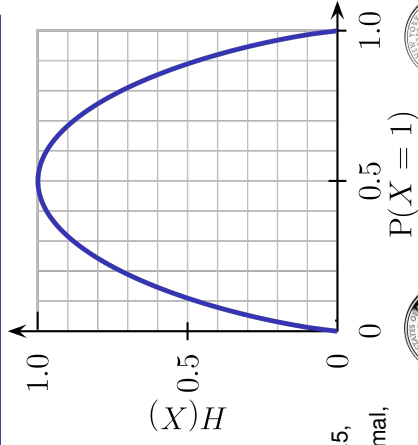
## Information entropy in a crooked coin



: 0



: 1



For the fair coin  $P(\text{heads}) = P(\text{tails}) = 0.5$ , uncertainty before the coin toss is maximal, and the entropy is maximal.



## Databases

**database:** a comprehensive collection of related information organized for convenient access, generally in a computer



Databases have grown so big, that organizing, finding and retrieving information is a big problem.

Google Inc.

mission statement: "organize the world's information and make it universally accessible and useful" + "don't be evil"

runs over one million servers in data centers around the world

processes over one billion search requests and about twenty-four petabytes ( $10^{15}$  bytes) of user-generated data every day

## Redundancy

Redundancy is the number of bits used to transmit a message minus the number of bits of actual information in the message.

It is easy to guess what is missing, because spoken and written information are redundant.

event	number	binary code	
	1	001	We could code eight events using a three digit binary code.
	2	010	
	3	011	
	4	100	
	5	101	
	6	110	
			redundancy:
			$R = -\log_2(1/8) = (-\log_2(1/6))$
			$R = -0.415... \text{ bit}$

## Finding biomedical data (textbooks)

### freebooks4doctors

100 free medical textbooks  
<http://freebooks4doctors.com/>

9. Ultrasound Imaging
13. Biochemistry (Stryer)
17. Radiology
19. Atlas of Human Anatomy
37. Human Anatomy Online
38. Muscle Atlas
39. Biochemistry Online
51. Basics of MRI
54. Chest X-ray
56. Electronic Statistics Book
58. King's Biochemistry
75. Anatomy at a Glance
79. Anatomical Images
83. Atlas of Microscopic Anatomy
100. Atlas of Human Anatomy in Cross Section

### NCBI bookshelf

online open access to 1056 biomedical books  
<http://www.ncbi.nlm.nih.gov/books/browse/>

National Center for Biotechnology Information

### e-library of the Semmelweis University

access to 137 online books  
<http://www.lib.sote.hu/nav/ebooks>  
(some may only be accessible from the university computers)



## Finding biomedical data (journals)

### NCBI PubMed

<http://www.ncbi.nlm.nih.gov/pubmed/>  
the number one open access biomedical database and search engine

### electronic journal catalog of the Semmelweis University

[http://www.lib.sote.hu/nav/journals\\_catalog](http://www.lib.sote.hu/nav/journals_catalog)  
online access to 3979 journals (some may only work from university computers)

### Science Citation Index

<http://science.thomsonreuters.com/cgi-bin/jrnlist/jloptions.cgi?PC=K>  
very powerful proprietary scientific database and search engine

### OVID technologies

<http://ovidsp.tx.ovid.com/>  
search and access scholarly books and journals (proprietary)

## Finding biomedical data (miscellaneous publications)

### SciVerse Scopus

<http://www.scopus.com/home.url>  
search scholarly publication in journals, patents, and web pages

### Google patents

<http://www.google.com/patents>  
free patent search engine and database (contains over 8 million patents)

### Google scholar

<http://scholar.google.com/>  
freely accessible search engine to find scientific journal and web publications

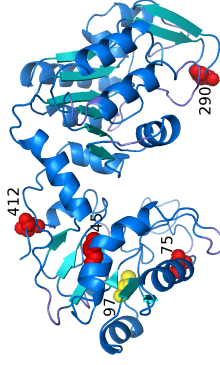
## Finding biomedical data (macromolecular structures)

### Protein Data Bank

<http://www.pdb.org>  
As of Tuesday Nov 22, 2011 contains 77394 structures obtained by x-ray diffraction or NMR, or electron microscopy.  
Download protein, DNA and RNA structures and view them in 3D using pdb viewers such as **Rasmol**, **Pymol**, **VMD**.



Immunoglobulin (Ig) domain



Phosphoglycerate kinase

## Finding biomedical data (bioinformatics)

### NCBI: BLAST

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>  
Basic Local Alignment Search Tool from NCBI  
compare primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences

### NCBI: OMIM

<http://www.google.com/patents>  
Online Mendelian Inheritance in Man from NCBI  
catalogues all the known diseases with a genetic component, and links them to the relevant genes in the human genome

### NCBI: Gene

<http://www.ncbi.nlm.nih.gov/gene>  
nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources for a wide range of species