Information & Quantities & Data

Sorana D. Bolboacă

OBJECTIVES

- Information Theory
- Quantity of Information
- International systems used to measure quantities and speeds and their applications
- Coding Information
- Data – Information - Knowledge
- Data vs. Constant
- Types of Medical Data

INFORMATION THEORY

- What?
  - Information = knowledge that can be used
  - Communication = exchange of information
  - Goals of information:
    - Efficient (remove redundancy & irrelevance) &
    - Reliable (something that is reliable can be trusted or believed because they work or behave well in the way you expect) &
    - Secure

INFORMATION THEORY

- Developed by Claude E. Shannon
  - Data compression (JPEG, MP3)
  - Reliable communication through noisy channels (memories, Cds, DVDs, Internet, etc.)
  - Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal 1948; 27:379–423 & 623–656.

- The field is at the intersection of mathematics, statistics, computer science, physics, neurobiology, and electrical engineering.

- Sub-fields:
  - source coding, channel coding, algorithmic complexity theory, algorithmic information theory, and measures of information.

INFORMATION THEORY

- Information theory answers two fundamental questions:
  - What is the ultimate data compression?
    - Answer: The Entropy H.
  - What is the ultimate transmission rate?
    - Answer: Channel Capacity C.

- Entropy:
  - A measure of information (Shannon)
  - Expressed by the average number of bits needed for storage or communication
  - Quantifies the uncertainty involved when encountering a random variable:
    - a fair coin flip (2 equally likely outcomes) will have less entropy than a roll of a die (6 equally likely outcomes)

INFORMATION THEORY

- Memoryless sources: generate successive independent and identically distributed outcome
- The source (S) has outcomes that occur with probabilities (p)
- The entropy of a source (S,p) in bits (binary digits) is:

$$H(S) = - \sum_i p_i \log_2 p_i$$

- The larger the entropy, the less predictable is the source output and the more information is produced by seeing it!

INFORMATION THEORY

- Information theory answers two fundamental questions:
  - What is the ultimate data compression?
    - Answer: The Entropy (H).
  - What is the ultimate transmission rate?
    - Answer: Channel Capacity (C).

- Channel Capacity (C):

$$C = \max(H(X)-H(X/Y))$$

QUANTITY OF INFORMATION: SHANNON

- Let S be a system with the following states $\{S_1,S_2,...,S_n\}$
- Let $p_1,...,p_n$ be the probability of apparition of the states
- The quantity of information produced by apparition of S_k state is given by the formula:

$$I_k = - \log_2 p_k$$

- A system with two states (0 and 1):
  - The system has two states $\{S_1, S_2\}$ with probabilities of apparition $p_1 = p_2 = ½$
  - The quantity of information produced through apparition of S_1 OR S_2 is:

$$I_{1/2} = - \log_2 ½ = 1 \text{ bit}$$

QUANTITY OF INFORMATION

- All types of information in computers are representing using binary code:
  - Numbers
  - Letters
  - Processor instructions
  - Graphics
  - Video
  - Sound

BITS AND BYTES

- Binary digit = one numeral in a binary number
  - Each 1 and 0 in the following number bellow is a binary digit: 11000101
- BIT = binary digit
- BYTE = 8 bits grouped together
- 2 symbols are used to represent binary numbers: 0 and 1

BITS AND BYTES

- A bit (b) is the smallest unit of data comprised of just {0,1}
- 1 nibble (-) = 4 bits (cutesy term with limited usage; mostly bitfields)
- 1 byte (B) = 8 bits (you could also say 2 nibbles)

QUANTITY OF INFORMATION

- Used to express storage capacity:

1. International Electrotechnical Commission (binary system)
   - 1 kibibyte (KiB) = 1,024 B = 1,024^1 B = 1,024 B
   - 1 mebibyte (MiB) = 1,024 KB = 1,024^2 B = 1,048,576 B
   - 1 gibibyte (GiB) = 1,024 MB = 1,024^3 B = 1,073,741,824 B
   - 1 kibibit (Kib) = 1,024 b = 1,024^1 b = 1,024 b
   - 1 mebibit (Mib) = 1,024 Kb = 1,024^2 b = 1,048,576 b
   - 1 gibibit (Gib) = 1,024 Mb = 1,024^3 b = 1,073,741,824 b…

QUANTITY OF INFORMATION

- Used to express commercial storage capacity:

2. International System of Units (decimal system)
   - 1 kilobyte (KB) = 1,000 B = 1,000^1 B 1,000 B
   - 1 megabyte (MB) = 1,000 KB = 1,000^2 B = 1,000,000 B
   - 1 gigabyte (GB) = 1,000 MB = 1,000^3 B = 1,000,000,000 B
   - 1 kilobit (Kb) = 1,000 b = 1,000^1 b 1,000 b
   - 1 megabit (Mb) = 1,000 Kb = 1,000^2 b = 1,000,000 b
   - 1 gigabit (Gb) = 1,000 Mb = 1,000^3 b = 1,000,000,000 b
- *kbps* = kilobits per second → data rates

QUANTITY OF INFORMATION

- binary system: hardware, memory stick, etc.
- commercial system: CD, DVD, etc.
- Speed of download/upload: … MBps = mega bytes per second
- Speed of data processing:
  - MIPS = millions of instructions per second
  - FLOPS = FLoating-point Operations Per Second
    - Microprocessors had 4 FLOPS/cycles ⟶ 2.5GHz = 10 billion FLOPS = 10 GFLOPS

CODING INFORMATION

- Coding:
  - Numbers
  - Text
  - Images
- Binary Representation
  - Binary = two possible states (0 OR 1)
  - Any information stored into computer (e.g. text, numbers, images, etc.) can take just value 0 or 1

BINARY REPRESENTATION

No.	No. UI	Message* [(message example)]	Formula*
1	2	2 [(0); (1)]	2^1
2	4	4 [(00); (01), (10), (11)]	2^2
3	8	8 [(000); (001); (010); (011); (100); (101); (110); (111)]	2^3
4	16	16 [(0000); (....); ...]	2^4
...			2^n
8	256	256 [(00000000); ...]	2^8
UI = units of information			

CODING NUMBERS: BINARY

- Binary: Symbol: 0 OR 1
- Correspondence zecimal – binary:
  - 0 = **0**
  - 1 = **1**
  - 2 = **10**
  - 3 = **11**
  - 4 = **100**
  - 5 = **101**
  - 6 = **110**
  - 7 = **111**
  - 8 = **1000**
  - 9 = **1001**
  - 10 = **1010**

- Add:
  - 0 + 0 = 0
  - 0 + 1 = 1
  - 1 + 0 = 1
  - 1 + 1 = 10 (with exceeding)

- Subtract:
  - 0 - 0 = 0
  - 0 - 1 = 1 (with loaning)
  - 1 - 0 = 1
  - 1 - 1 = 10

- Multiply:
  - $0 \times 0 = 0$
  - $0 \times 1 = 0$
  - $1 \times 0 = 0$
  - $1 \times 1 = 1$

CODING NUMBERS: OCTAL

- The numerical values are represented using eight symbols: from 0 to 7

$$120 = 1\times8^2 + 1\times8^1 + 2\times8^0$$

- For representation of octal values are necessary 3 bits: from 000 to 111
- Transformation of a binary number into an octal number is made grouping the bytes in groups of 3 from right to left:

$$110110110111001_{(2)} = 66671_{(8)}$$

- Transformation of an octal number into a binary number: $65_{(8)} = 110101_{(2)}$

- 0 = 000
- 1 = 001
- 2 = 010
- 3 = 011
- 4 = 100
- 5 = 101
- 6 = 110
- 7 = 111

CODING NUMBERS: HEXADECIMAL

- Has the base 16 and use 16 hexadecimal code noted as:
  - The code from $0_{(16)}$ to $9_{(16)}$ have the decimal equivalent values from $0_{(10)}$ to $9_{(10)}$
  - The code from $A_{(16)}$ to $F_{(16)}$ have the decimal values from $10_{(10)}$ to $15_{(10)}$.
- For their representation 4 bytes are needed
  - Starting with 0000 and ending with 1111
- Transformation of a binary number to a hexadecimal number can be performed by grouping as 4 bytes from right to left:

$$110110110111001_{(2)} = 6DD9_{(16)}$$

CODING TEXT

- **ASCII (American Standard Code for Information Interchange)**
  - Use 7 bits for representation of 128 characters
  - Is the most used schema for coding the characters

Binary	Oct	Dec	Hex	Glyph
010 0000	040	32	20	␣
010 0001	041	33	21	!
010 0010	042	34	22	"
010 0011	043	35	23	#
010 0100	044	36	24	$
010 0101	045	37	25	%
010 0110	046	38	26	&
010 0111	047	39	27	'
010 1000	050	40	28	(
010 1001	051	41	29	)
010 1010	052	42	2A	*
010 1011	053	43	2B	+
010 1100	054	44	2C	,
010 1101	055	45	2D	-
010 1110	056	46	2E	.
010 1111	057	47	2F	/
011 0000	060	48	30	0
011 0001	061	49	31	1
011 0010	062	50	32	2
011 0011	063	51	33	3
011 0100	064	52	34	4
011 0101	065	53	35	5
011 0110	066	54	36	6
011 0111	067	55	37	7
011 1000	070	56	38	8
011 1001	071	57	39	9
011 1010	072	58	3A	:
011 1011	073	59	3B	;
011 1100	074	60	3C	<
011 1101	075	61	3D	=
011 1110	076	62	3E	>
011 1111	077	63	3F	?

Binary	Oct	Dec	Hex	Glyph
100 0000	100	64	40	@
100 0001	101	65	41	A
100 0010	102	66	42	B
100 0011	103	67	43	C
100 0100	104	68	44	D
100 0101	105	69	45	E
100 0110	106	70	46	F
100 0111	107	71	47	G
100 1000	110	72	48	H
100 1001	111	73	49	I
100 1010	112	74	4A	J
100 1011	113	75	4B	K
100 1100	114	76	4C	L
100 1101	115	77	4D	M
100 1110	116	78	4E	N
100 1111	117	79	4F	O
101 0000	120	80	50	P
101 0001	121	81	51	Q
101 0010	122	82	52	R
101 0011	123	83	53	S
101 0100	124	84	54	T
101 0101	125	85	55	U
101 0110	126	86	56	V
101 0111	127	87	57	W
101 1000	130	88	58	X
101 1001	131	89	59	Y
101 1010	132	90	5A	Z
101 1011	133	91	5B	[
101 1100	134	92	5C	\
101 1101	135	93	5D	]
101 1110	136	94	5E	^
101 1111	137	95	5F	_

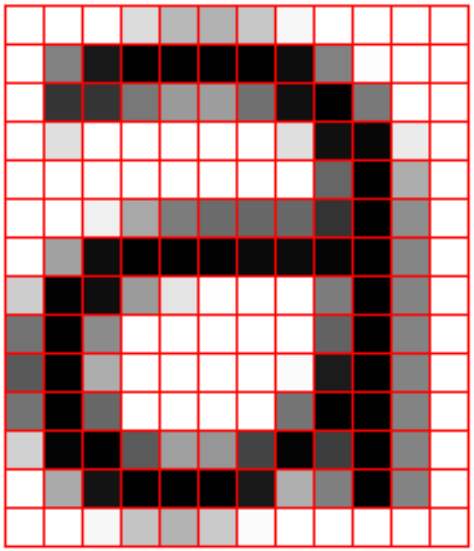
Binary	Oct	Dec	Hex	Glyph
110 0000	140	96	60	`
110 0001	141	97	61	a
110 0010	142	98	62	b
110 0011	143	99	63	c
110 0100	144	100	64	d
110 0101	145	101	65	e
110 0110	146	102	66	f
110 0111	147	103	67	g
110 1000	150	104	68	h
110 1001	151	105	69	i
110 1010	152	106	6A	j
110 1011	153	107	6B	k
110 1100	154	108	6C	l
110 1101	155	109	6D	m
110 1110	156	110	6E	n
110 1111	157	111	6F	o
111 0000	160	112	70	p
111 0001	161	113	71	q
111 0010	162	114	72	r
111 0011	163	115	73	s
111 0100	164	116	74	t
111 0101	165	117	75	u
111 0110	166	118	76	v
111 0111	167	119	77	w
111 1000	170	120	78	x
111 1001	171	121	79	y
111 1010	172	122	7A	z
111 1011	173	123	7B	{
111 1100	174	124	7C	|
111 1101	175	125	7D	}
111 1110	176	126	7E	~

CODING IMAGES

- Digital image (raster images, or bitmap images): is a representation of a two-dimensional image using ones and zeros (binary).
- Pixel = is the smallest item of information in an image
  - Are normally arranged in a 2-dimensional grid
  - Often represented using dots or squares
  - The intensity of each pixel is variable; in colour systems, each pixel has typically three or four components such as red, green, and blue, or cyan, magenta, yellow, and black.
  - The word pixel is based on a contraction of pix ("pictures") and el (for "element"). Similar formations with el for "element" include the words: voxel (a volume element, three dimensional space) and texel (fundamental unit of texture space - computer graphics).

CODING IMAGES

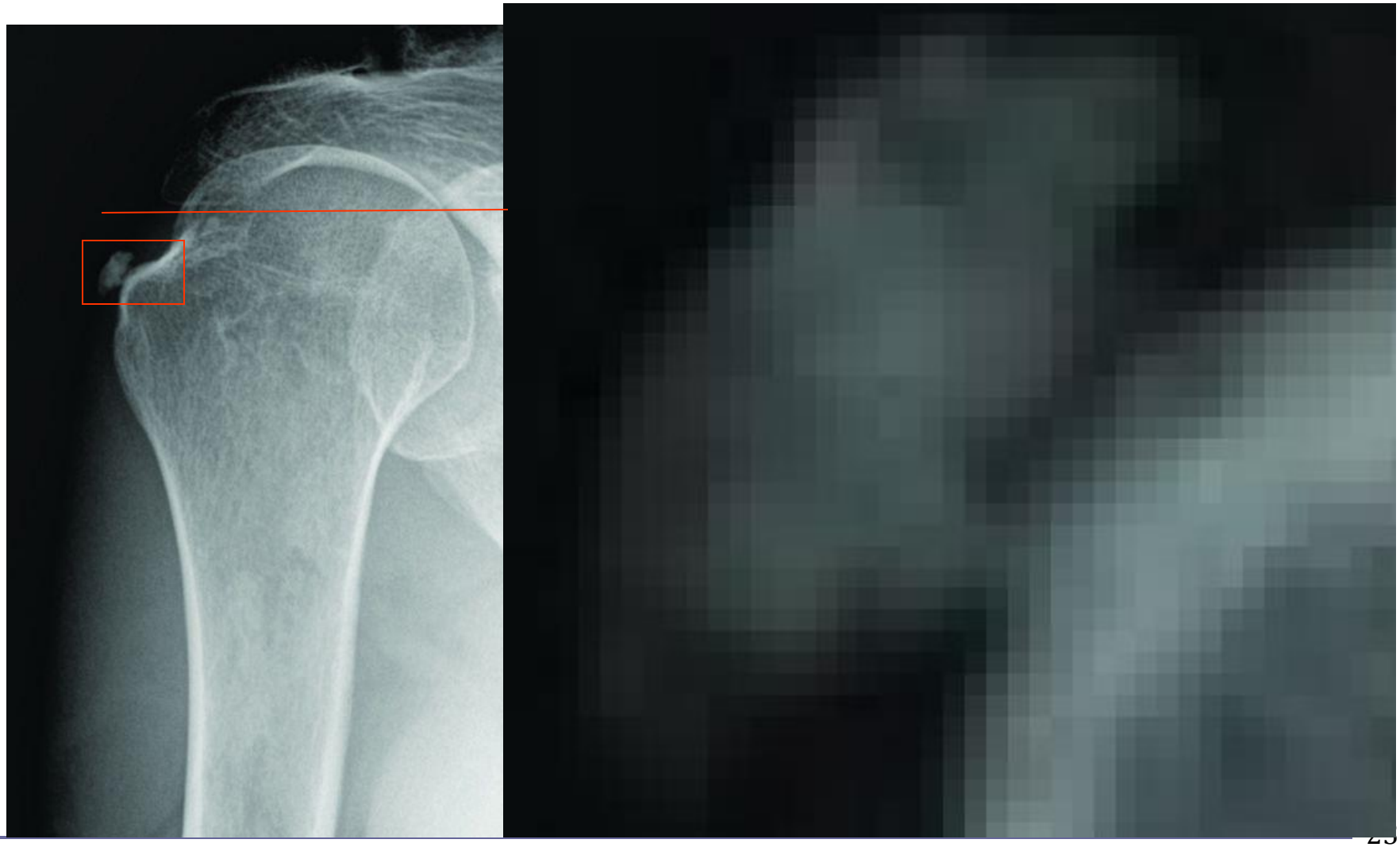
a



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

- The number of distinct colors that can be represented by a pixel depend on the number of bits per pixel (bpp)
- The maximum number of colors for a pixel are :
  - 8 bpp, 28 = 256 hue
  - 16 bpp, 216 = 65536 hues– High Color
  - 24 bpp, 224 = 16777216 hues– True Color
  - 48 bpp: continuous space of colors

Images Coding



CODING IMAGES

- The number of pixels from a image is called resolution:
  - Display resolution: 1024×768, diagonal:
    - 19", pixel dimension: 0.377 mm
  - Display resolution: 800×600, diagonal:
    - 17", pixel dimension : 0.4318 mm
  - Display resolution: 640×480, diagonal :
    - 15", pixel dimension: 0.4763 mm

MEDICAL CODING (MEDICAL CLASSIFICATION)

- The process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers
- Medical classification systems are used for a variety of applications in medicine and medical informatics:
  - Statistical analysis of diseases and therapeutic actions
  - Reimbursement; e.g., based on DRGs (Diagnosis-related group)
  - Knowledge-based and decision support systems
  - Direct surveillance of epidemic or pandemic outbreaks

MEDICAL CODING

- Diagnostic codes
- Procedural codes
- Pharmaceutical codes
- Topographical codes

Reference Classifications

- International Statistical Classification of Diseases and Related Health Problems (ICD, includes ICD9 and ICD9-CM, currently used in US)
- International Classification of Functioning, Disability and Health (ICF)
- International Classification of Health Interventions (ICHI) - under development

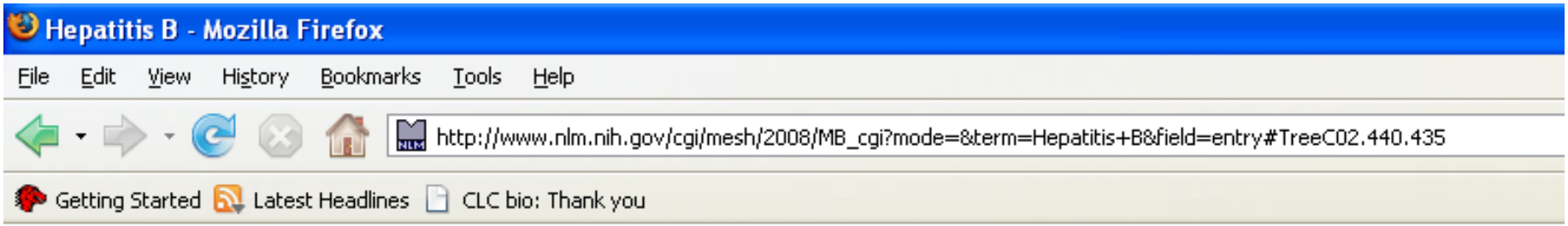
MEDICAL CODING

- **Related Classifications**
  - International Classification of Primary Care (ICPC-2)
  - International Classification of External Causes of Injury (ICECI)
  - Anatomical Therapeutic Chemical Classification System (ATC/DDD)
  - Technical aids for persons with disabilities: Classification and terminology (ISO9999)

Derived Classifications

1. International Classification of Diseases for Oncology, Third Edition (ICD-O-3)
2. ICD-10 for Mental and Behavioural Disorders
3. Application of the International Classification of Diseases to Dentistry and Stomatology, 3rd Edition (ICD-DA)
4. Application of the International Classification of Diseases to Neurology (ICD-10-NA)
5. International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY)

MeSH (Medical Subject Headings)



Virus Diseases [C02]

Hepatitis, Viral, Human [C02.440]

Hepatitis A [C02.440.420]

Hepatitis B [C02.440.435]

Hepatitis B, Chronic [C02.440.435.100]

Hepatitis C [C02.440.440] +

Hepatitis D [C02.440.450] +

Hepatitis E [C02.440.470]

WHY CODING MEDICAL INFORMATION?

- Improves the effectiveness of communication in health care systems
- Facilitates the integration of different systems
- Cuts the cost defined in terms of time, resources, etc..
- Supports health care quality management
- Supports medical research

DATA - INFORMATION - KNOWLEDGE

DEFINITIONS

- Data (datum) = a single piece of information, as a fact, statistic, or code; an item of data.
  - When data are processed, organized, structured or presented in a given context so as to make them useful, they are called **Information**.
- Information = consists of facts and data organized to describe a particular situation or condition
- Knowledge = consists of facts, truths, and beliefs, perspectives and concepts, judgments and expectations, methodologies and know-how.
  - Knowledge is accumulated and integrated and held over time to handle specific situations and challenges.

DATA

- Symbol set that is quantified and/or qualified.
- It simply exists and has no significance beyond its existence (in and of itself).
- It can exist in any form, usable or not.
- It does not have meaning of itself.
  - Example:
    - a spreadsheet generally starts out by holding data
    - data are the coded invariance

INFORMATION

- Data that are processed to be useful
- Provides answers to "who", "what", "where", and "when"
- Data that has been given meaning by way of relational connection. This "meaning" can be useful, but does not have to be.
- Is related to meaning or human intention
  - Example:
    - a relational database makes information from the data stored within it
    - the contents of databases, the web etc.

KNOWLEDGE

- application of data and information
- answers "how" questions
- is the appropriate collection of information, such that it's intent is to be useful.
  - Knowledge is a deterministic process.
  - **Knowledge** is embodied in humans as the capacity to understand, explain and negotiate concepts, actions and intentions.

"THE APPLICATION OF WHAT WE KNOW WILL HAVE A BIGGER IMPACT ON HEALTH AND DISEASE THAN ANY SINGLE DRUG OR TECHNOLOGY LIKELY TO BE INTRODUCED IN THE NEXT DECADE."

SIR MUIR GRAY, UK NATIONAL LIBRARY FOR HEALTH

KNOWLEDGE IS THE ENEMY OF DISEASE

HEALTHCARE KNOWLEDGE

- from **research** (sometimes called evidence)
- from the analysis of **routinely collected and audit data** (sometimes called statistics)
- knowledge from the **experience of clinicians and patients**

BASIC DEFINITIONS

- A characteristic can be classified in one of two ways:
  - **Quantitative:** it can be assigned a numeric value (such as 0, 1, 7.5, 17, or -¼)
  - **Qualitative***:* it cannot be assigned a compelling numeric value (such as name, birthplace, or gender)
- A characteristic is called a ***variable*** because its specific value or nature is not known before the item is examined.
- We determine a ***value*** for a characteristic or variable by making a ***measurement*** of it using an instrument (e.g. weight) or by making an ***observation*** of it by examining it (e.g. the color of the skin). The value for a characteristic of a specific item is called a ***datum***.

DATA VS. CONSTANT

- Constant
  - Something that does not or cannot change or vary
  - Unchanging in nature, value, or extent; invariable
  - A number, value, or object that has a fixed magnitude, physically or abstractly, as a part of a specific operation or discussion
    - Physics: a number expressing a property, quantity, or relation that remains unchanged under specified conditions.
    - Mathematics: a quantity assumed to be unchanged throughout a given discussion.

TYPES OF MEDICAL DATA

Qualitative (attribute)

- Sex
- Diagnosis
- Presence/Absence of a symptom
- **…**

Quantitative

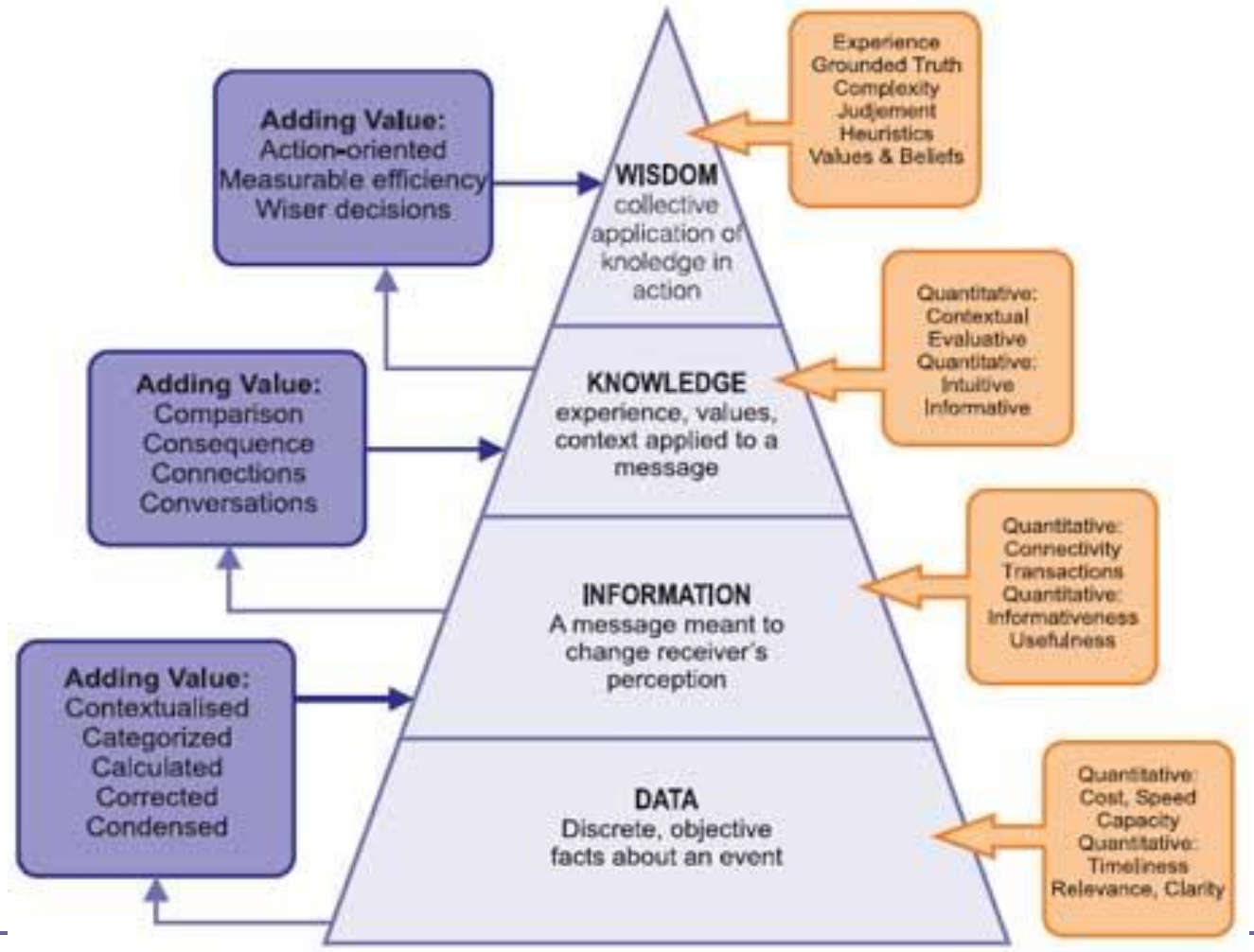
- SBP, DBP
- Level of Blood Sugar
- **…**

Signals

- EEG (Electroencephalography)
- EKG (Electrocardiography)

- **Images**
  - Echography
  - Tomography
  - Radiography
  - …

DATA – INFORMATION – KNOWLEDGE HIERARCHY



SUMMARY

- Information Theory lead to Quantity of Information
- Coding Information is important
- Data - Information - Knowledge
- Data vs. Constant
- Types of Medical Data

TASK

- Look at the following 3 abstracts:
  - http://www.ncbi.nlm.nih.gov/pubmed/24069382
  - http://www.ncbi.nlm.nih.gov/pubmed/24049294
  - http://www.ncbi.nlm.nih.gov/pubmed/23956899

- and identify:
  - Variable
  - Data
  - Information
  - Knowledge