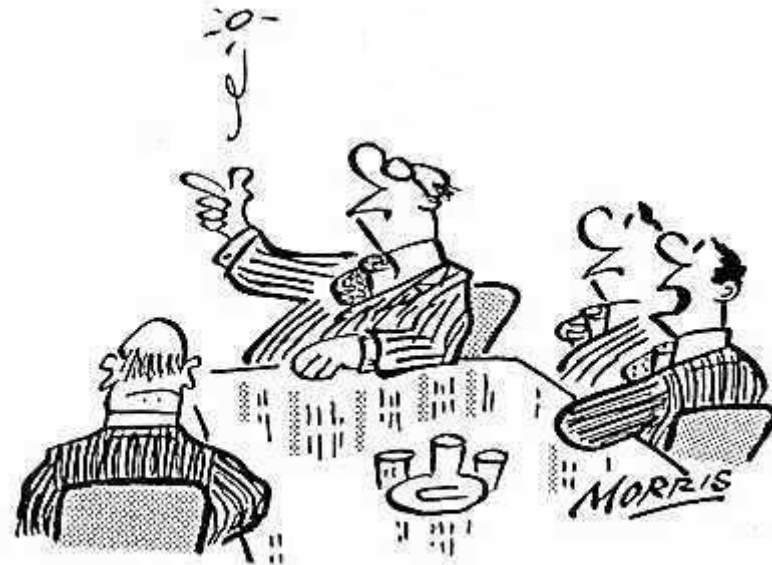


Információelmélet

Az információ fogalma (példán keresztül)

Adatok, adatfolyamok információtartalma, kódolás, továbbítás

Információ és entrópia



Bárcsak olyan nyugodt lehetnék mint J.B. amikor fontos döntésekről van szó!

Az információ fogalma (példával)

Intuitívan

"informare" (Lat.) : „**az elmét formálni**”, tanítani, utasítani valakit

Azaz: akkor tudunk tanulni, vélekedésünket megváltoztatni, ha *információhoz* jutunk

vagy:

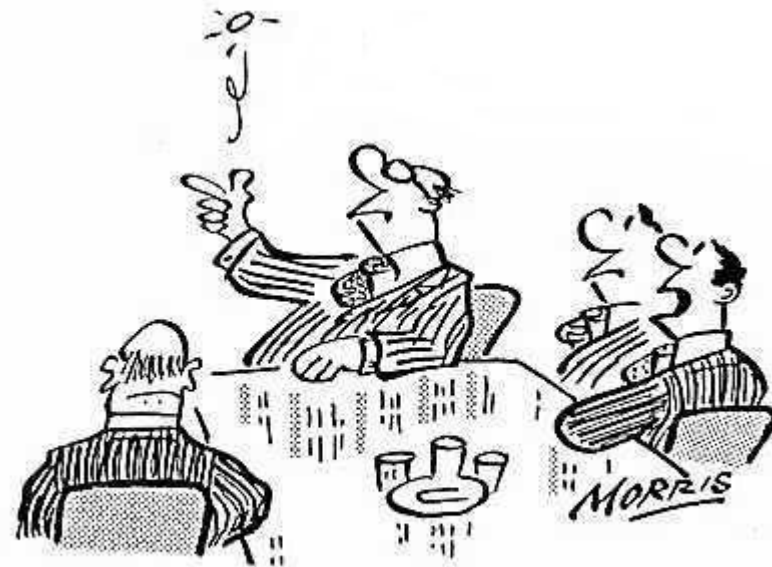
„egy eszközbe vagy élőlénybe bevitt jel, mely választ vált ki”

(Pl. Pavlovi reflex: táplálék illata → nyáleválasztás, mozdulatok)

vagy:

„ az információ olyan mintázat amely más mintázato

(Pl. DNS szekvencia → fehérje szerkezet)



Bárcsak olyan nyugodt lehetnék mint J.B. amikor fontos döntésekről van szó!

Információ átvitel – információ tartalom

Esemény és információ:
„mi történt?”

Az egyes események információtartalma eltérő

-megint dugó van reggel

-holnap esni fog.

-nyertem a lottón!

Hogyan *kódolhatjuk* az információt?
Mi kell az információátvitelhez?



Információátvitel - kódolás

általánosságban

Információ forrás



(át)vevő
cél(pont)

A lehetséges események közül
melyik következett be?

(hír)

Információátvitel - kódolás

általánsságban

Információ forrás



Átviteli csatorna



**(át)vevő
cél(pont)**

A lehetséges események közül
melyik következett be?

(hír)

Információátvitel - kódolás

általánosságban

Információ forrás

A lehetséges események közül
melyik következett be?

kódolás

eseményeket SZÁMOKKAL reprezentálunk



Átviteli csatorna

dekódolás

SZÁMOKBÓL visszaállítjuk az **eseményeket**



**(át)vevő
cél(pont)**

(hír)

Információátvitel - kódolás

általánosságban

Információ forrás

kódolás



Átviteli csatorna

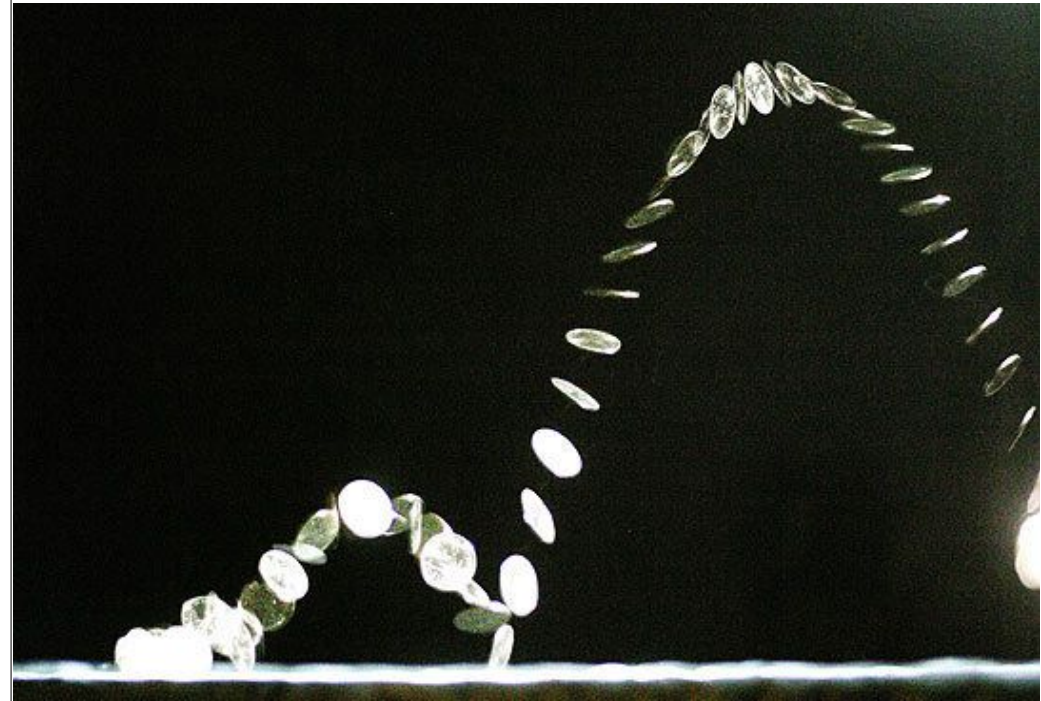
dekódolás



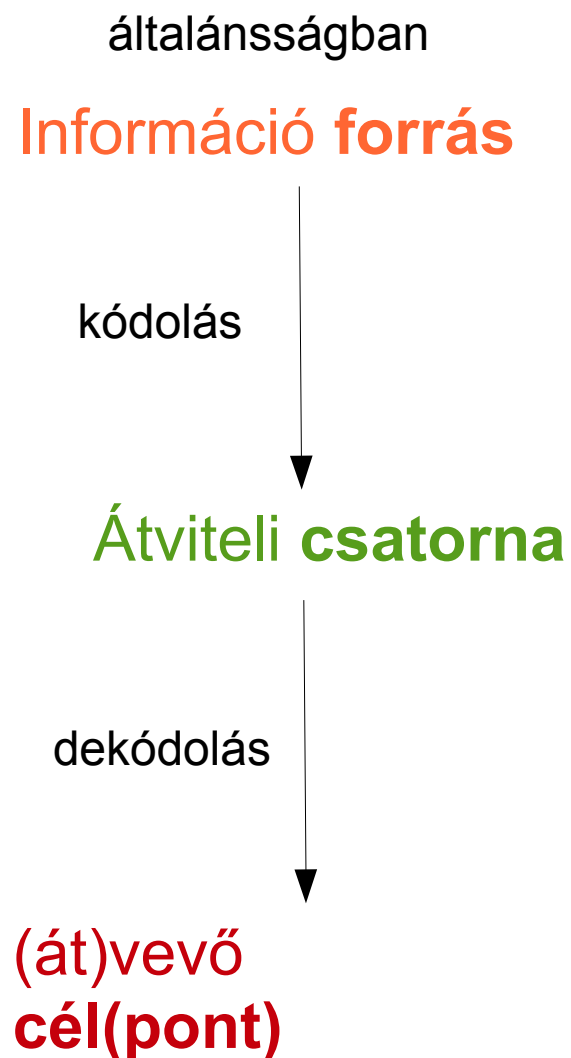
**(át)vevő
cél(pont)**

példa

Fej vagy írás?



Információátvitel - kódolás



A példában

Fej vagy írás?



Ki nyert?

Információátvitel - kódolás

általánosságban

Információ **forrás**

kódolás

Átviteli **csatorna**

dekódolás

(át)vevő
cél(pont)



A példában

Fej vagy írás?

kódolás
Fej/Írás → 1,0



Ki nyert?

Információátvitel - kódolás

általánosságban

Információ **forrás**

kódolás

Átviteli **csatorna**

dekódolás

(át)vevő
cél(pont)



A példában

Fej vagy írás?

kódolás

Fej/Írás → 1,0

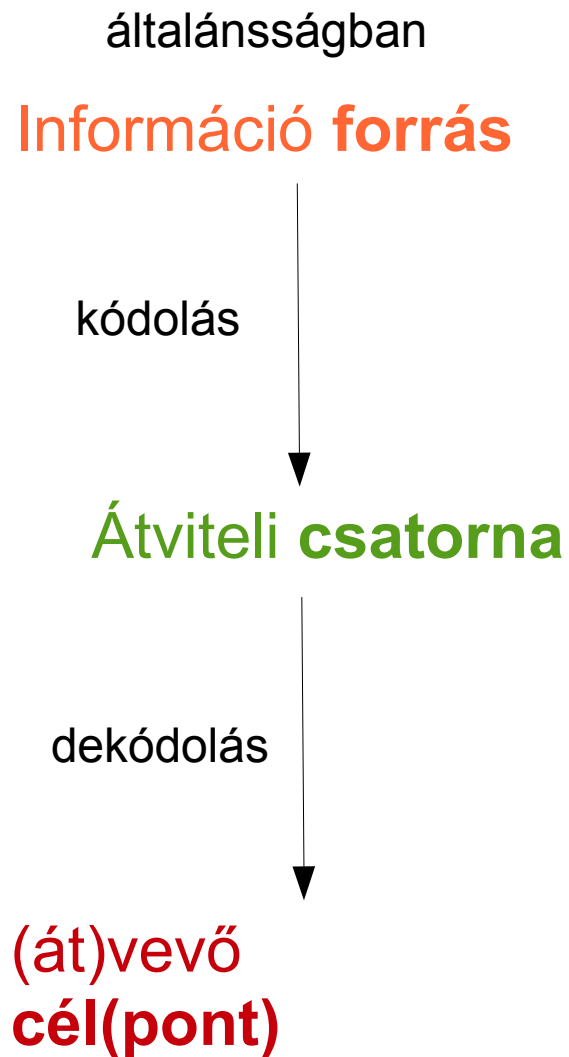


Beszéd, SMS, rádió, stb.



Ki nyert?

Információátvitel - kódolás



Információátvitel - kódolás

SZÁMOKKÁ
alakítjuk az információt.

általánosságban

Információ **forrás**

kódolás

Átviteli **csatorna**

dekódolás

(át)vevő
cél(pont)



A példában

Fej vagy írás?

kódolás

Fej/Írás → **1,0**

Beszéd, SMS, rádió, stb.

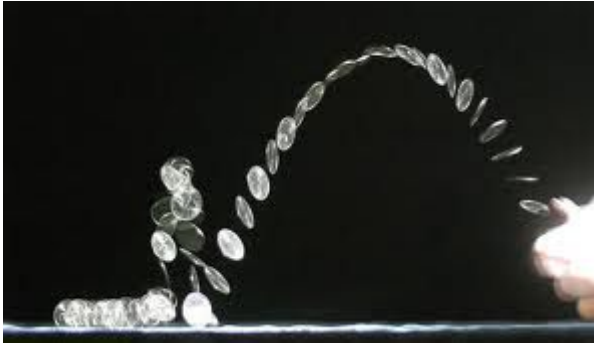
dekódolás



1,0 → Fej/Írás

Ki nyert?









Információátvitel – digitális kódolás



Esemény	szám	digitális kód
	: 1	1
	: 0	0



	: 1	001
	: 2	010
	: 3	011
	: 4	100
	: 5	101
	: 6	110

Információátvitel – digitális kódolás

Hány **bitre** van szükségünk a kódoláshoz?

Bit: **b**inary dig**it**

0 vagy 1

Számjegyek : csak *két* jegy van: 0 és 1.
(10-es rendszerben 9 jegy van: 0,1,2,...,9)
A számjegyeknek helyiértéke van, amit az alap
hatványai adnak meg:
 2^2 , 2^1 , 2^0
(10-esben $10^2=100$, $10^1=10$, $10^0=1$)

$$\text{Pl: } 101_2 = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 4 + 0 + 1 = 5_{10}$$

Információátvitel – digitális kódolás

ennyi **bit** kell?

Esemény	szám	digitális kód
---------	------	---------------



:

1

1



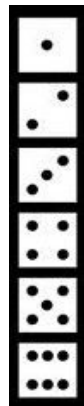
:

0

0

Bitek száma

1



:

1

001

:

2

010

:

3

011

:

4

100

:

5

101


:

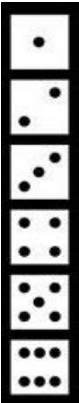
6

110

3

Információátvitel – kódolási hatások







Esemény	szám	digitális kód	Bitek száma	Maximális eseményszám
	:	1	1	2
	:	0		

	:	1	001	3	8
	:	2	010		
	:	3	011		
	:	4	100		
	:	5	101		
	:	6	110		

7
0

111 Itt 6 eseményt kódoltunk 3 biten, csak hogy $2^3=8$,
000 azaz 8 eseményt is lehetne kódolni ennyi bit segítségével

Információátvitel – kódolási hatások

Esemény	szám	digitális kód	Bitek száma	Maximális eseményszám
	: 1	001	3	8
	: 2	010		
	: 3	011		
	: 4	100		
	: 5	101		
	: 6	110		

7
0







111 Itt 6 eseményt kódoltunk 3 biten, csak hogy $2^3=8$,
000 azaz 8 eseményt is lehetne kódolni ennyi bit segítségével

Egy jobb hatásfokú kódolás:

$\{X_1 X_2 X_3\}$ csoportosítsuk az eseményeket 3-asával

Ez az előbbiek szerint
3x3 bit = **9** bit igényt jelent

Információátvitel – kódolási hatások

Esemény	szám	digitális kód	Bitek száma	Maximális eseményszám
	1	001	3	8
	2	010		
	3	011		
	4	100		
	5	101		
	6	110		

7
0

111 Itt 6 eseményt kódoltunk 3 biten, csak hogy $2^3=8$,
000 azaz 8 eseményt is lehetne kódolni ennyi bit segítségével

Egy jobb hatásfokú kódolás:

$\{X_1 X_2 X_3\}$ csoportosítsuk az eseményeket 3-asával

Ez az előbbiek szerint
3x3 bit = 9 bit igényt jelent

Összesen $6^3 = 216$ lehetőség van
de $2^8=256$, így 8 bit is elég
(sőt sok is).

1 bittel kevesebb is bőven elég!

Információátvitel – kódolási határfok

Az információtartalom definiálható a legjobb hatásfokú kódolással

azaz:

Egy jel, esemény információtartalma megadható azzal, hogy minimálisan hány bitre van szükség az átvitléhez.

Ez egyben a kódolási hatásfok elméleti határa is.

Többnyire nem egyetlen átvitelre vagyunk kíváncsiak, hanem arra, hogy egy több lehetséges eseményből álló *eseménytér* hogyan kódolható.

Információátvitel – kódolási hatások

Az információtartalom definiálható a legjobb hatásfokú kódolással

azaz:

Egy jel, esemény információtartalma megadható azzal, hogy minimálisan hány bitre van szükség az átvitléhez.

Ez egyben a kódolási hatások elméleti határa is.

Hogyan kapcsolódik ez az intuitív információtartalomhoz?

-Fej vagy írás?

-Ma reggel nincs dugó.

-Holnap esni fog.

-nyertem a lottón!

Információátvitel – kódolási hatások

Az információtartalom definiálható a legjobb hatásfokú kódolással

azaz:

Egy jel, esemény információtartalma megadható azzal, hogy minimálisan hány bitre van szükség az átviteléhez.

Ez egyben a kódolási hatások elméleti határa is.

Hogyan kapcsolódik ez az intuitív információtartalomhoz?

	p	q
-Fej vagy írás?	$\frac{1}{2}$	$\frac{1}{2}$
-Ma reggel nincs dugó.	$\frac{1}{4}$	$\frac{3}{4}$
-Holnap esni fog.	1%	99%
-nyertem a lottón!	1/13,983,816	0.999....

Információátvitel – kódolási hatások

Az információtartalom definiálható a legjobb hatásfokú kódolással

azaz:

Egy jel, esemény információtartalma megadható azzal, hogy minimálisan hány bitre van szükség az átviteléhez.

Ez egyben a kódolási hatások elméleti határa is.

Hogyan kapcsolódik ez az intuitív információtartalomhoz?

-Fej vagy írás?	p $\frac{1}{2}$	q $\frac{1}{2}$	nem tudjuk előre
-Ma reggel nincs dugó.	$\frac{1}{4}$	$\frac{3}{4}$	
-Holnap esni fog.	1%	99%	
-nyertem a lottón!	$1/13,983,816$	0.999....	valószínű nem nyert...

Információátvitel – kódolási hatások

Az információtartalom definiálható a legjobb hatásfokú kódolással

azaz:

Egy jel, esemény információtartalma megadható azzal, hogy minimálisan hány bitre van szükség az átviteléhez.

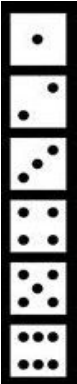
Ez egyben a kódolási hatások elméleti határa is.

Hogyan kapcsolódik ez az intuitív információtartalomhoz?

-Fej vagy írás?	p $\frac{1}{2}$	q $\frac{1}{2}$	nem tudjuk előre
-Ma reggel nincs dugó.	$\frac{1}{4}$	$\frac{3}{4}$	
-Holnap esni fog.	1%	99%	
-nyertem a lottón!	$1/13,983,816$	0.999....	valószínű nem nyert...

A nyert információ az esemény valószínűségével fordítottan arányos!

információátvitel – az információ mértéke

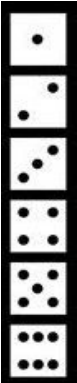
„rendes”	P_i	valószínűség	kódolási példa	bit igény	p^* (bitek száma)
	1/6	0.17	000	3	0.5
	1/6	0.17	001	3	0.5
	1/6	0.17	010	3	0.5
	1/6	0.17	011	3	0.5
	1/6	0.17	100	3	0.5
	1/6	0.17	101	3	0.5

bitek számának várható értéke:

3

cinkelt P_i

Ebben az esetben hatékonyabban kódolhatunk:

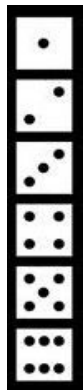
	1/2	0.5	0	1	0.5
	1/4	0.25	10	2	0.5
	1/8	0.13	110	3	0.38
	1/16	0.06	1110	4	0.25
	1/32	0.03	11110	5	0.16
	1/32	0.03	11111	5	0.16

bitek számának várható értéke:

1.94

információátvitel – az információ mértéke

„rendes”	P_i	valószínűség	kódolási példa	bit igény	p^* (bitek száma)
----------	-------	--------------	----------------	-----------	---------------------



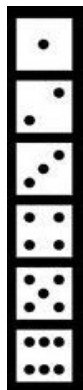
1/6	0.17	000	3	0.5
1/6	0.17	001	3	0.5
1/6	0.17	010	3	0.5
1/6	0.17	011	3	0.5
1/6	0.17	100	3	0.5
1/6	0.17	101	3	0.5

itt nem tudunk előre semmit, azaz a bizonytalanság maximális

bitek számának várható értéke: **3**

cinkelt P_i

A cinkelt kocka eredményeinek kisebb a „hírértéke”, kevesebbet tanulunk belőle



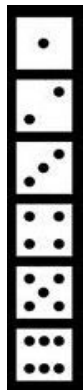
1/2	0.5	0	1	0.5
1/4	0.25	10	2	0.5
1/8	0.13	110	3	0.38
1/16	0.06	1110	4	0.25
1/32	0.03	11110	5	0.16
1/32	0.03	11111	5	0.16

itt van előismeretünk az 1-est esetleg a 2-est várjuk.

bitek számának várható értéke: **1.94**

információátvitel – az információ mértéke

„rendes”	P_i	valószínűség	kódolási példa	bit igény	p^* (bitek száma)
----------	-------	--------------	----------------	-----------	---------------------

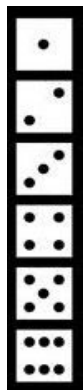


1/6	0.17	000	3	0.5
1/6	0.17	001	3	0.5
1/6	0.17	010	3	0.5
1/6	0.17	011	3	0.5
1/6	0.17	100	3	0.5
1/6	0.17	101	3	0.5

itt nem tudunk előre
semmit, azaz
a bizonytalanság
maximális

bitek számának várható értéke: **3**

cinkelt P_i



1/2	0.5	0	1	0.5
1/4	0.25	10	2	0.5
1/8	0.13	110	3	0.38
1/16	0.06	1110	4	0.25
1/32	0.03	11110	5	0.16
1/32	0.03	11111	5	0.16

itt van előismeretünk
az 1-est esetleg a 2-est
várjuk.

Átlagosan kevesebbet
„tanulunk” a cinkelt
kockával

bitek számának várható értéke: **1.94**

információátvitel – az információ mértéke

Hogyan lehet **matematikailag** definiálni az információtartalmat (H)? (Shannon 1948)

1.: H legyen p_i folytonos függvénye (ha p_i egy kicsit változik \rightarrow H sem „ugorhat” nagyot)

2.: Ritka események információtartalma nagyobb mint a gyakoriaké:

H fordítottan legyen arányos p -vel

Ha minden p_i egyforma, ($p_i = 1/n$)

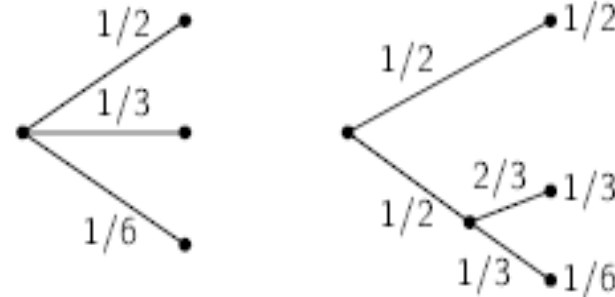
Akkor H az események számának (n) monoton növekvő függvénye legyen.

Ha több esemény lehetséges, akkor több választásra van lehetőség, nagyobb a bizonytalanság
Ekkor egy esemény bekövetkeztének nagyobb a hírértéke.

3.: Elágazó választási lehetőségek:

H értéke nem függhet attól, hogy milyen úton jutottunk az adott választáshoz (esemény bekövetkezéséhez)

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} \cdot H\left(\frac{2}{3}, \frac{1}{3}\right)$$



információátvitel – az információ mértéke

Shannon :

$$H = p \cdot \log_2 \left(\frac{1}{p} \right)$$

\log_2 : 2-es alapú logaritmus

pl.:

$$\log_2 (2) = 1$$

$$\log_2 (4) = 2$$

$$\log_2 (8) = 3$$

információátvitel – az információ mértéke

Shannon :

$$H = p \cdot \log_2 \left(\frac{1}{p} \right)$$

H az átlagos, várható információtartalmat adja meg, ez hozható kapcsolatba a kódolással.

Szokás megadni egyetlen esemény információtartalmát is (I):

$$I = \log_2 \left(\frac{1}{p} \right)$$

Ezzel $H = p \cdot I$, azaz a bekövetkezési valószínűséggel súlyozott információtartalom.

információátvitel – az információ mértéke

Shannon

$$H = p \cdot \log_2 \left(\frac{1}{p} \right) \quad [\text{bit}]$$

Ha a teljes eseményteret la akarjuk fedni, akkor összegezni kell minden eseményre

$$H = \sum_i p_i \cdot \log_2 \left(\frac{1}{p_i} \right) = \sum_i -p_i \cdot \log_2 p_i$$

2 helyettmás log alappal is lehet:

$\log_e (\ln)$: [nat]

$\log_{10} (\lg)$: [ban]

az információ mértéke - entrópia

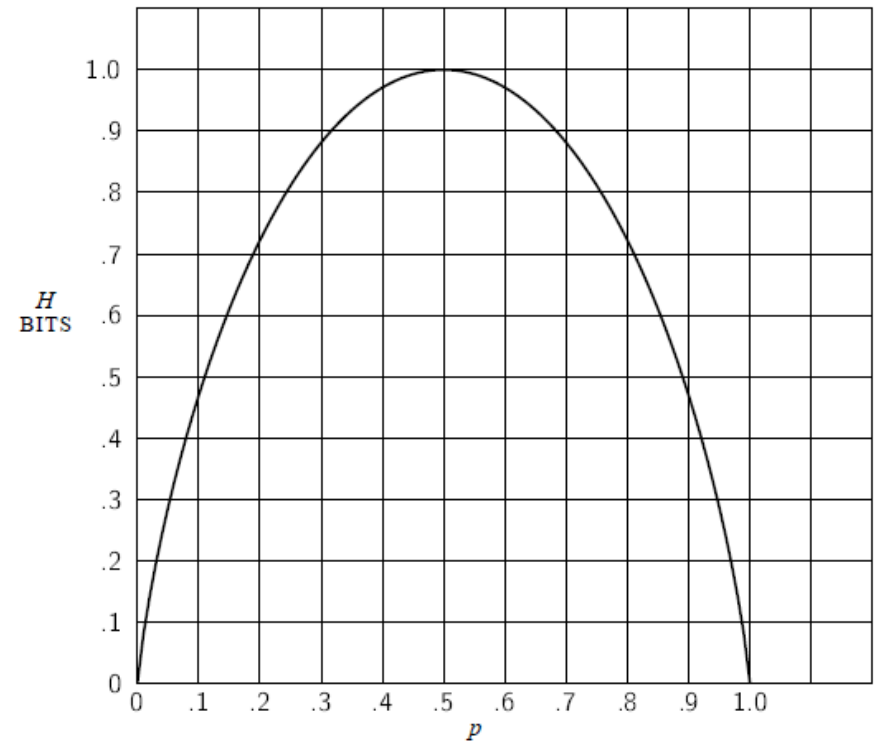
Fej vagy írás



p



$q = 1-p$



$$H = \sum_i -p_i \cdot \log_2 p_i = -p \cdot \log_2 p - q \cdot \log_2 q = -p \cdot \log_2 p - (1-p) \cdot \log_2 (1-p)$$

az információ mértéke - entrópia

Fair érme: $p = \frac{1}{2}$

Nincs előfeltevés,
Maximális bizonytalanság

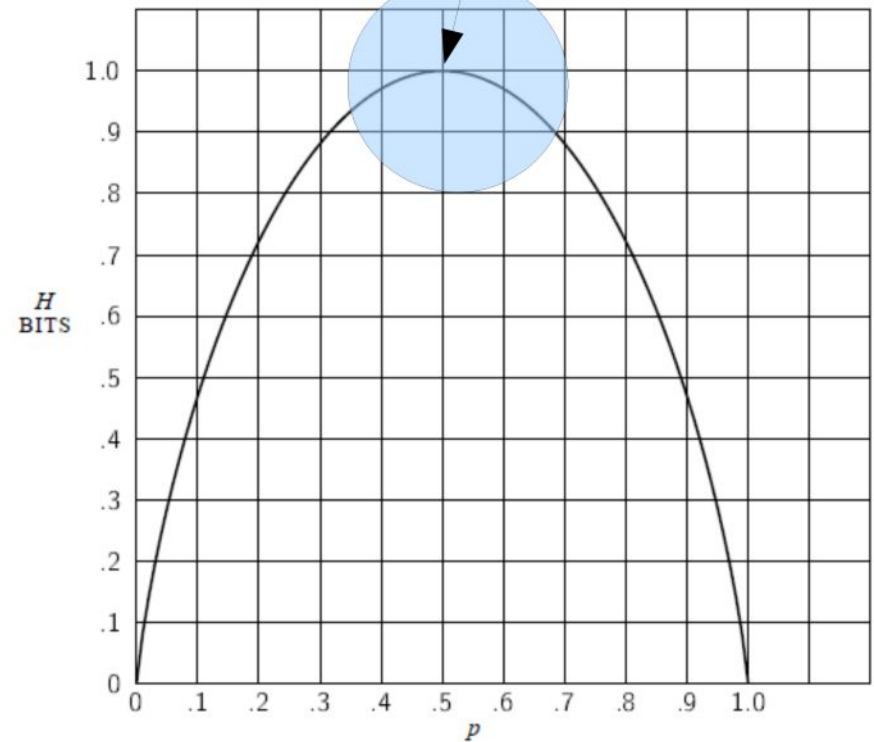
Fej vagy írás



p



$q = 1-p$



$$H = \sum_i -p_i \cdot \log_2 p_i = -p \cdot \log_2 p - q \cdot \log_2 q = -p \cdot \log_2 p - (1-p) \cdot \log_2 (1-p)$$

az információ mértéke - entrópia

Fair érme: $p = \frac{1}{2}$

Nincs előfeltevés,
Maximális bizonytalanság

Fej vagy írás

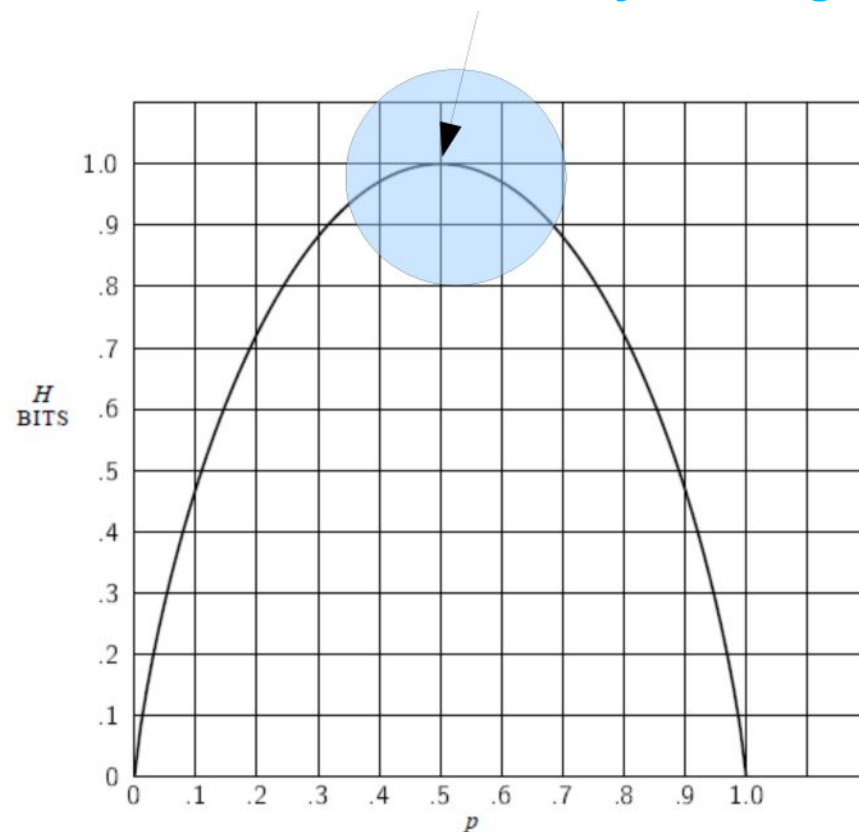


p



$q = 1-p$

H másik neve: **Shannon-entrópia**



H akkor **maximális** ha semmit sem tudunk előre azaz minden eseményhez egyforma $p_i (= 1/n)$
Minden lehetőség egyformán valószínű, legtöbbféle előfordulást látjuk



A fizikai entrópia (S) maximuma is ott van ahol a mikroállapotok száma maximális

Példa:

Egy betegségben ugyanazokat a tüneteket 3-féle kórokozó okozhatja, legyen ezek megfigyelt relatív gyakorisága $A=0.002$, $B=0.35$, $C=0.648$. Adjuk meg az egyes kórokozók előfordulásának információtartalmát, illetve a leghatékonyabb kódoláshoz szükséges bitek számát! Mennyi az információs entrópia?

$$I_A = \log_2\left(\frac{1}{p_A}\right) = 8.966 \text{ és } I_B = \log_2\left(\frac{1}{p_B}\right) = 1.514 \text{ és } I_C = \log_2\left(\frac{1}{p_C}\right) = 0.626$$

$$H = P_A * I_A + P_B * I_B + P_C * I_C = 0.018 + 0.53 + 0.40 = 0.948$$

Az információs entrópia megegyezik az elvi leghatékonyabb kódoláshoz szükséges bitek számával.

Nem biztos hogy technikailag létezik ilyen, de ennél jobbat nem lehet csinálni.

Orvosi alkalmazások

Diagnosztikus tesztekől nyert információ

Szakértői rendszerek

Genetikai adatbázisok

Proteomika

...

a_i : preteszt valósz.

b_i : poszt teszt valósz.

$$D(b||a) = \sum_{i=1}^n b_i \log_2(b_i/a_i)$$

Testing Situation	Pretest Probability of Disease	Test Operating Characteristics: Sensitivity/Specificity	Test Result	Posttest Probability of Disease	Information Gained
Breast cancer screening with mammography	0.01	0.75/0.94	Positive	0.11	0.25 bits
			Negative	0.003	0.006 bits
Mammography given palpable breast mass	0.2	0.80/0.90	Positive	0.67	0.74 bits
			Negative	0.05	0.13 bits
Screening for HIV with antibody test	0.001	0.99/0.998	Positive	0.33	2.4 bits
			Negative	0.00001	0.001 bits
Presence of tonsillar exudate in diagnosing infection with group A streptococci	0.1	0.45/0.84	Positive	0.24	0.11 bits
			Negative	0.07	0.01 bits
Colon cancer screening by fecal occult blood testing	0.005	0.40/0.90	Positive	0.02	0.02 bits
			Negative	0.003	0.0005 bits

Adatbázisok

Adatbázisokban ***információt*** tárolnak, rendszereznek és olvasnak ki.

FOSTER CITY EYE CARE - OPTOMETRIC CENTER PATIENT HISTORY QUESTIONNAIRE

Last name	First name	Mr. <input type="checkbox"/> Mrs. <input type="checkbox"/> Miss. <input type="checkbox"/> Ms. <input type="checkbox"/>
Address		
Telephone (W)	(H)	(Cell)
SSN	Date of Birth	Age
Occupation	Computer Hours Per Day	
Employer		
Emergency contact/Telephone no.		
Date of last eye exam	Dilated?	Today's Date
Hobbies or Sports		
Primary reason for today's exam		

MEDICAL INFORMATION

What is your general health:

Do you have any problems with any of these systems? (please circle all that apply)

Gastrointestinal	Y/N	Nervous	Y/N	Eyes	Y/N
Ear/Nose/Throat	Y/N	Genitourinary	Y/N	Mental	Y/N
Cardiovascular	Y/N	Musculoskeletal	Y/N	Endocrine (glands)	Y/N
Respiratory	Y/N	Integumentary (skin)	Y/N	Blood/lymph	Y/N
				Allergic/immunologic	Y/N
				Pregnant or nursing	Y/N

Please explain

Please answer all that apply:

Diabetes	Y/N	Type	Date of diagnosis
Allergies	Y/N	Allergic to what?	What happens?
Medication allergy	Y/N	What happens?	Headaches
Other health problems			HIV/AIDS
Current medication(s)			
Have you had any operations?	Y/N	Kind?	When?
Do you use cigarettes/tobacco?		Alcohol?	Other substance(s)?
Name of family doctor			Date of last visit
Date of last tetanus shot			

FAMILY HISTORY

High blood pressure	Y/N Relation	Macular degeneration	Y/N Relation
Diabetes	Y/N Relation	Retinal detachment	Y/N Relation
Glaucoma	Y/N Relation	Cataracts	Y/N Relation
Other eye condition(s)	Y/N What kind?		Relation

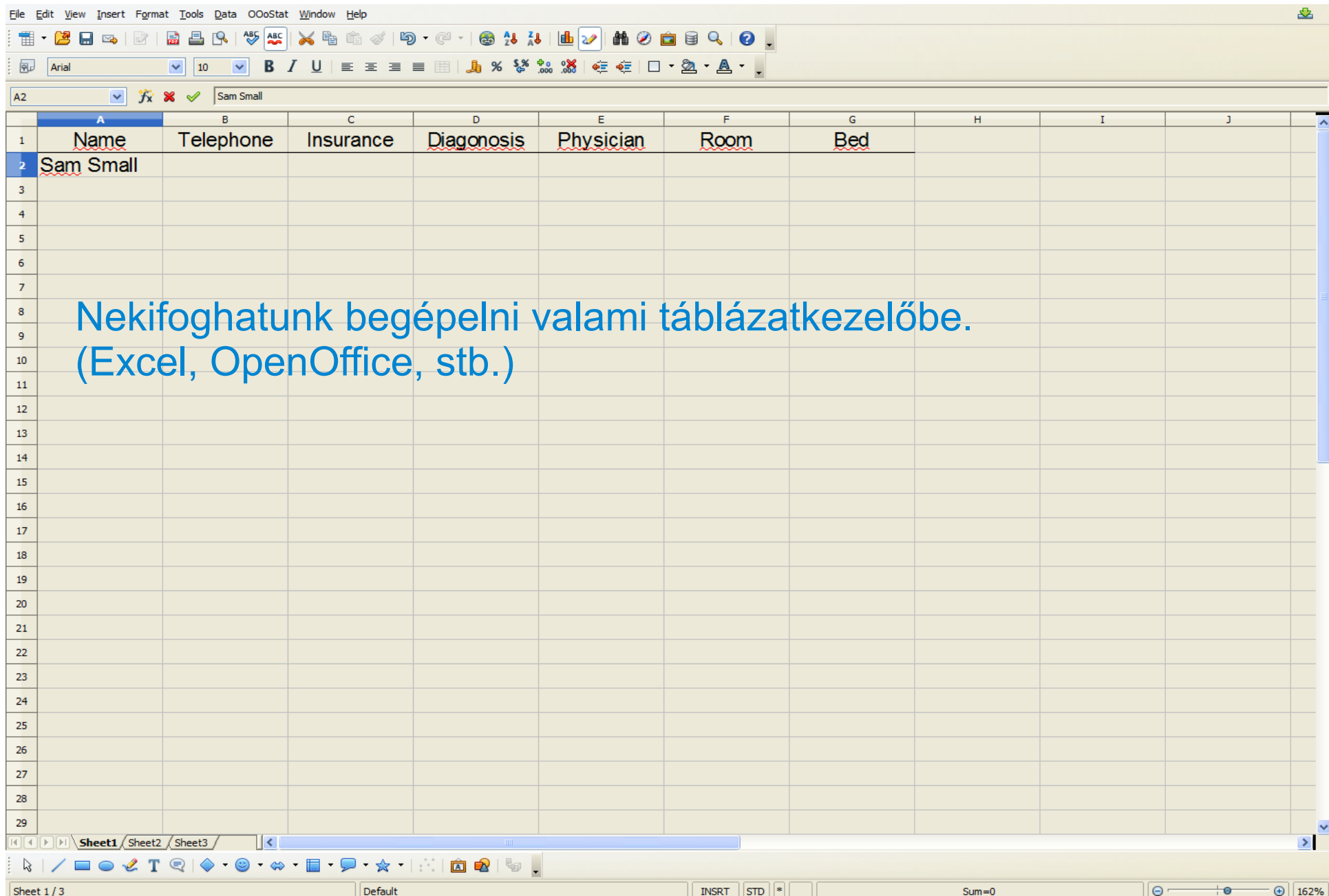
PERSONAL EYE INFORMATION

Have you had an eye operation?	Y/N	Type	Date
Have you had an eye injury?	Y/N	Kind	Date
Do you have glaucoma?	Y/N	Cataracts?	Y/N
Other eye problems?	Y/N	Dry eyes?	Y/N
Do you wear glasses?	Y/N	What kind?	Blurred vision? Y/N
Additional information		Contact lenses? Y/N	Type
Whom may we thank for referring you?		Are you interested in new contact lenses?	Y/N

Doctor's initials

Lehetne papíron is tárolni,
de ez nem hatékony

Adatbázisok - információtárolás



The screenshot shows the OpenOffice Calc application interface. The spreadsheet has a header row (row 1) with the following columns: Name, Telephone, Insurance, Diagonosis, Physician, Room, and Bed. The text "Diagonosis" is misspelled. Row 2 contains the data "Sam Small" in the Name column. The rest of the spreadsheet is empty. The status bar at the bottom shows "Sheet 1 / 3", "Default", "INSRT", "STD", "Sum=0", and a zoom level of 162%.

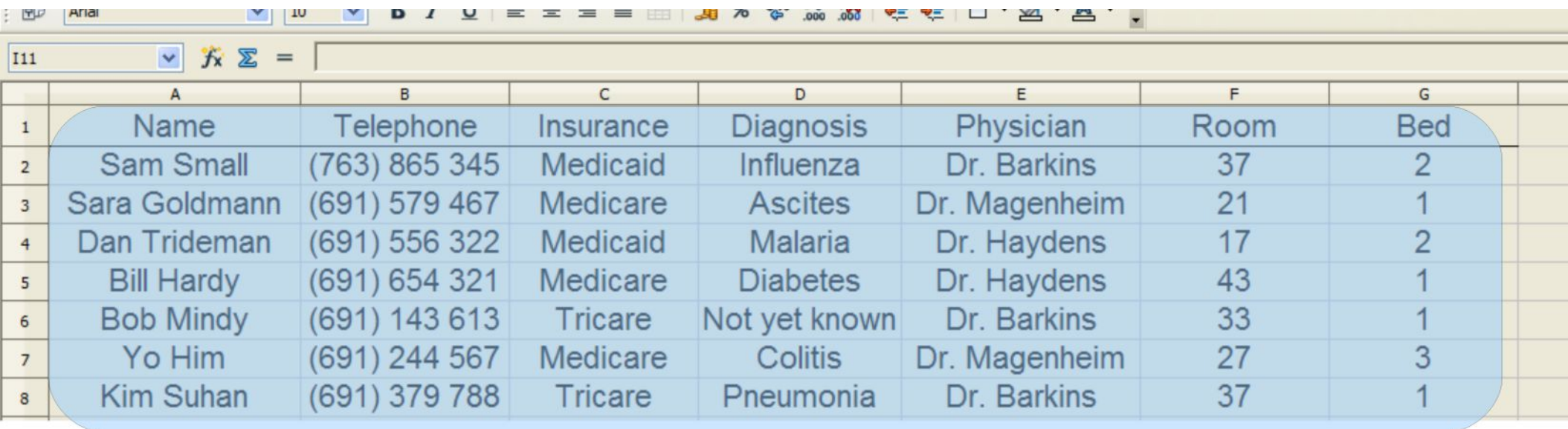
	A	B	C	D	E	F	G	H	I	J
1	<u>Name</u>	<u>Telephone</u>	<u>Insurance</u>	<u>Diagonosis</u>	<u>Physician</u>	<u>Room</u>	<u>Bed</u>			
2	<u>Sam Small</u>									
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										

Nekifoghatunk begépelni valami táblázatkezelőbe.
(Excel, OpenOffice, stb.)

Adatbázisok - információtárolás

File Edit View Insert Format Tools Data OOoStat Window Help								
G8 Arial 10 B I U								
G8 = 1								
	A	B	C	D	E	F	G	
1	Name	Telephone	Insurance	Diagnosis	Physician	Room	Bed	
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2	
3	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1	
4	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2	
5	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	43	1	
6	Bob Mindy	(691) 143 613	Tricare	Not yet known	Dr. Barkins	33	1	
7	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	27	3	
8	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barkins	37	1	
9								

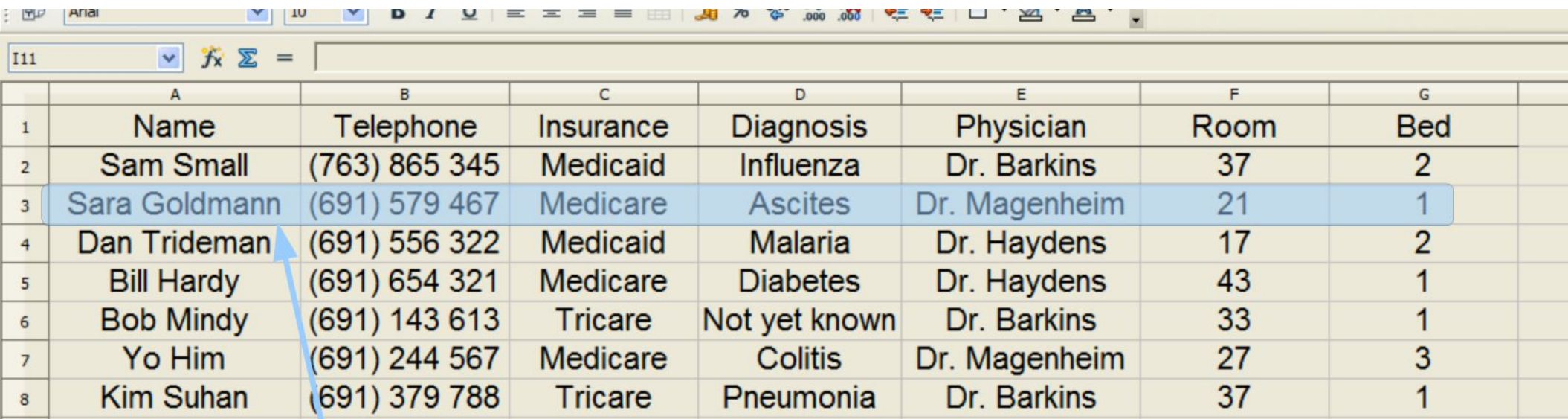
Adatbázisok - információtárolás



	A	B	C	D	E	F	G
1	Name	Telephone	Insurance	Diagnosis	Physician	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2
3	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1
4	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2
5	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	43	1
6	Bob Mindy	(691) 143 613	Tricare	Not yet known	Dr. Barkins	33	1
7	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	27	3
8	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barkins	37	1

Tábla : rendezett adathalmaz

Adatbázisok - információtárolás



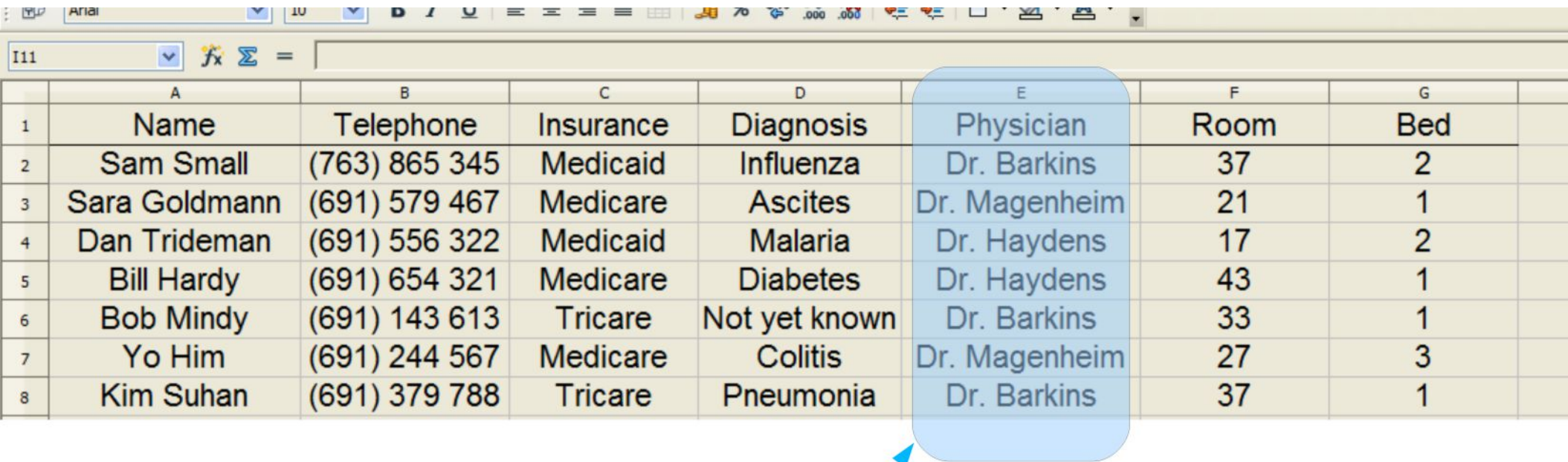
	A	B	C	D	E	F	G
1	Name	Telephone	Insurance	Diagnosis	Physician	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2
3	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1
4	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2
5	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	43	1
6	Bob Mindy	(691) 143 613	Tricare	Not yet known	Dr. Barkins	33	1
7	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	27	3
8	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barkins	37	1

Rekord : csoportosított információcsomag
(*egy sor a táblában*)

Minden sor egy adat-csoport

Minden sor szerkezete azonos

Adatbázisok - információtárolás



	A	B	C	D	E	F	G
1	Name	Telephone	Insurance	Diagnosis	Physician	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2
3	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1
4	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2
5	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	43	1
6	Bob Mindy	(691) 143 613	Tricare	Not yet known	Dr. Barkins	33	1
7	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	27	3
8	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barkins	37	1

oszlop: adat típus

rendezés

The screenshot shows the Microsoft Excel interface with the 'Data' menu open. The 'Sort...' option is highlighted. The spreadsheet contains the following data:

	A	B	C	D	E	F	G
	Name		Insurance	Diagnosis	Physician	Room	Bed
1	Sam Small		Medicaid	Influenza	Dr. Barkins	37	2
2	Sara Goldmann		Medicare	Ascites	Dr. Magenheim	21	1
3	Dan Trideman		Medicaid	Malaria	Dr. Haydens	17	2
4	Bill Hardy		Medicare	Diabetes	Dr. Haydens	43	1
5	Bob Mindy	(691) 745 610	Tricare	Not yet known	Dr. Barkins	33	1
6	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	27	3
7	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barkins	37	1

Adatbázisok - információtárolás

Rendeztethetjük valamilyen sorrend szerint az adatcsoportokat...

C2:C14								
	A	B	C	D	E	F	G	
1	Name	Telephone	Insurance	Diagnosis	Physician	Room	Bed	
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2	
3	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2	
4	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	43	1	
5	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1	
6	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	27	3	
7	Bob Mindy	(691) 143 613	Tricare	Not yet known	Dr. Barkins	33	1	
8	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barkins	37	1	
9								
10								
11								

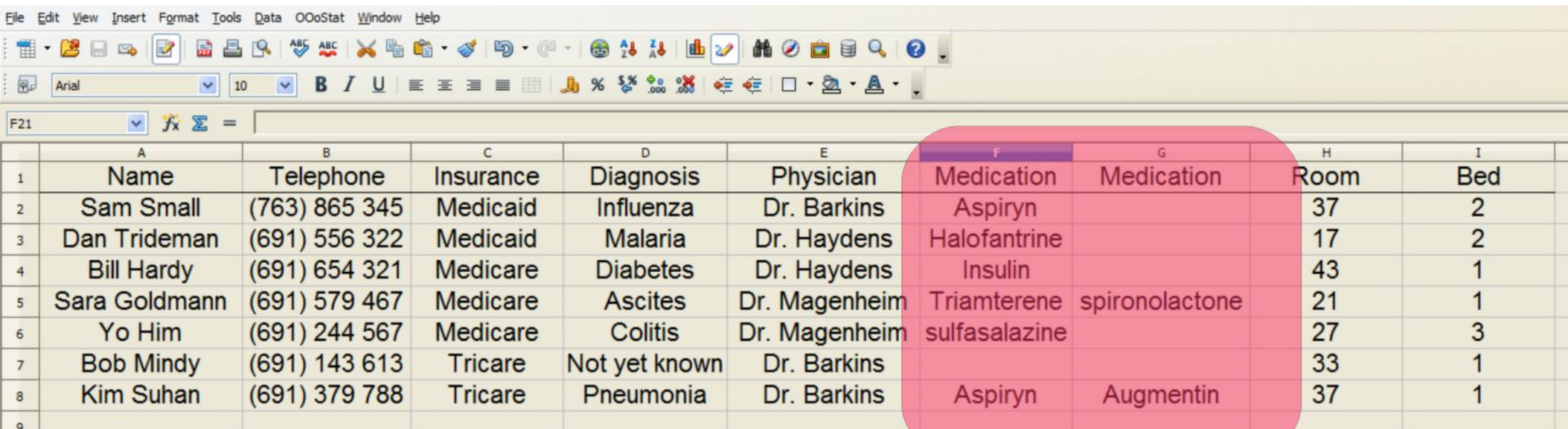
Adatbázisok – információ visszanyerés

Pl. kereséssel...

The screenshot shows a spreadsheet application interface. The 'Edit' menu is open, displaying various options including 'Undo: Sort', 'Can't Restore', 'Repeat', 'Cut', 'Copy', 'Paste', 'Paste Special...', 'Select All', 'Changes', 'Compare Document...', 'Find & Replace...' (highlighted), 'Headers & Footers...', 'Fill', 'Delete Contents...', and 'Delete Cells...'. The spreadsheet data is as follows:

	B	C	D	E	F	G	H
	Telephone	Insurance	Diagnosis	Physician	Room	Bed	
1	763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2	
2	691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2	
3	691) 654 321	Medicare	Diabetes	Dr. Haydens	43	1	
4	691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1	
5	691) 244 567	Medicare	Colitis	Dr. Magenheim	27	3	
6	691) 143 613	Tricare	Not yet known	Dr. Barkins	33	1	
7	691) 379 788	Tricare	Pneumonia	Dr. Barkins	37	1	
8							
9							

Adatbázisok – problémák



	A	B	C	D	E	F	G	H	I
1	Name	Telephone	Insurance	Diagnosis	Physician	Medication	Medication	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	Aspiryn		37	2
3	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	Halofantrine		17	2
4	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	Insulin		43	1
5	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	Triamterene	spironolactone	21	1
6	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	sulfasalazine		27	3
7	Bob Mindy	(691) 143 613	Tricare	Not yet known	Dr. Barkins			33	1
8	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barkins	Aspiryn	Augmentin	37	1
9									

Az egyes rekordok eltérő méretűek lehetnek

Helypazarló

Új adat-típusok hozzáadása nagyon körülményes

Inkonzisztens : ha egy mező üres az baj, vagy direkt van?

Adatbázisok – problémák

	A	B	C	D	E	F	G	H	I
	Name	Telephone	Insurance	Diagnosis	Physician	Medication	Medication	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	Aspiryn		37	2
3	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	Halofantrine		17	2
4	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	Insulin		43	1
5	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	Triamterene	spironolactone	21	1
6	Yo Him	(691) 244 567	Medicare	Colitis	Dr. Magenheim	sulfasalazine		27	3
7	Bob Mindy	(691) 143 613	Tricare	Not yet known	Dr. Barkins			33	1
8	Kim Suhan	(691) 379 788	Tricare	Pneumonia	Dr. Barpins	Aspiryn	Augmentin	37	1
9									

Ugyanazt többször is be kell gépelni:

Elírások

Feleslegesen többször is tároljuk

Később lehetetlen módosítani – túl sok példányban van meg

...

A Relational Model of Data for Large Shared Data Banks

E. F. CODD

IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on n -ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity

CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. INTRODUCTION

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of *data independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of *data inconsistency* which are expected to become troublesome even in nondeductive systems.

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the “connection trap”).

Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

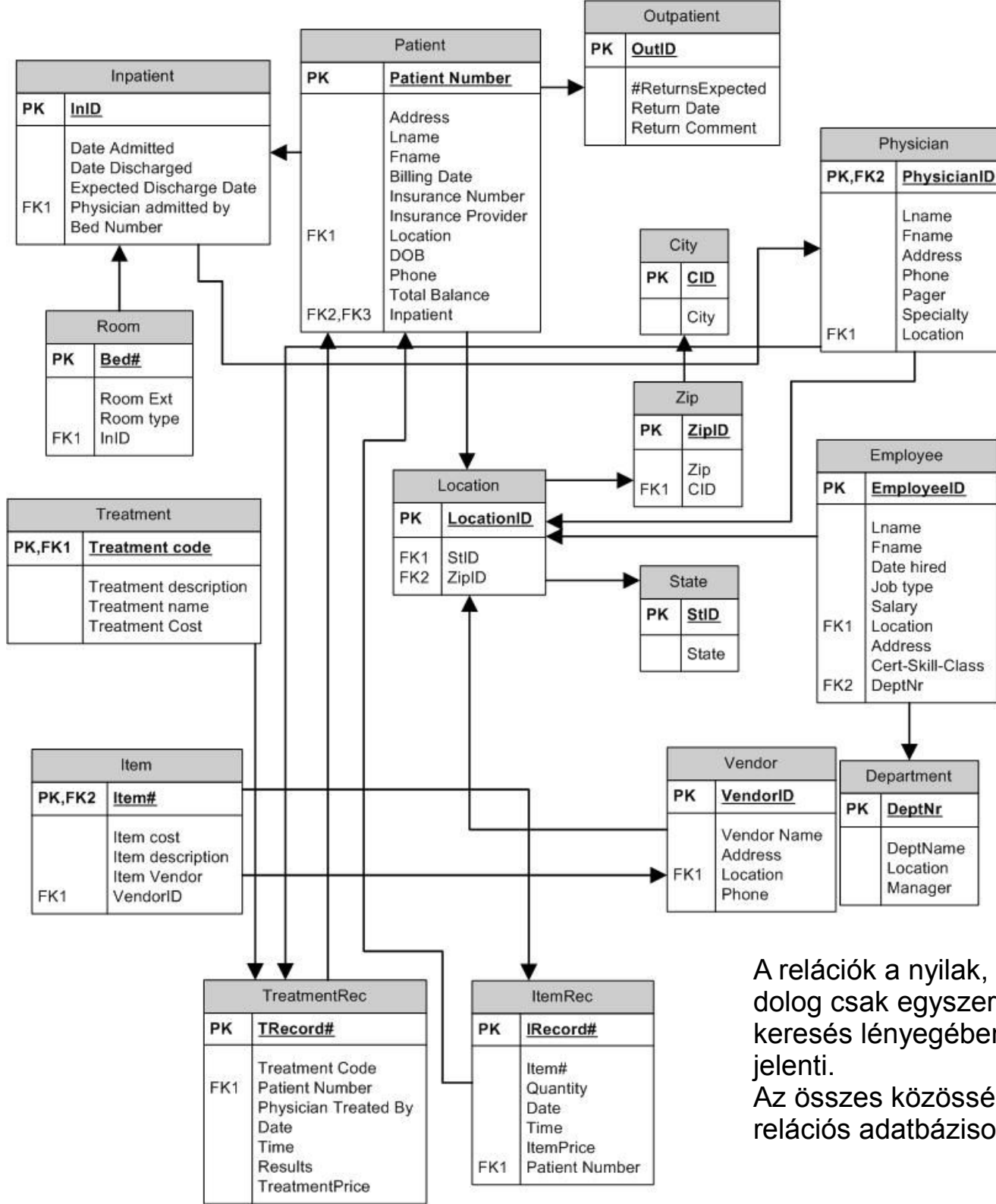
1.2. DATA DEPENDENCIES IN PRESENT SYSTEMS

The provision of data description tables in recently developed information systems represents a major advance toward the goal of data independence [5, 6, 7]. Such tables facilitate changing certain characteristics of the data representation stored in a data bank. However, the variety of data representation characteristics which can be changed *without logically impairing some application programs* is still quite limited. Further, the model of data with which users interact is still cluttered with representational properties, particularly in regard to the representation of collections of data (as opposed to individual items). Three of the principal kinds of data dependencies which still need to be removed are: ordering dependence, indexing dependence, and access path dependence. In some systems these dependencies are not clearly separable from one another.

1.2.1. Ordering Dependence. Elements of data in a data bank may be stored in a variety of ways, some involving no concern for ordering, some permitting each element to participate in one ordering only, others permitting each element to participate in several orderings. Let us consider those existing systems which either require or permit data elements to be stored in at least one total ordering which is closely associated with the hardware-determined ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from such a file is identical to (or is a subordering of) the

A relációs adatbázisban minden csak egyszer tárolódik, az egyes adat-típusok közötti kapcsolatokat tároljuk.

Ehhez egy matematikai leírás és programozási nyelv is tartozik, tetszőleges feladatra összeállítható egy adatbázis-szerkezet. Az adatbázisban a kapcsolatok az információ-áramlás irányát is jelzik.



A relációk a nyilak, az adatbázisban egy dolog csak egyszer tárolódik, és egy keresés lényegében a nyilak követését jelenti.
Az összes közösségi oldal mögött pl relációs adatbázisok dolgoznak...

Kiegészítő anyag: az információs entrópia és a fizikai entrópia kapcsolata

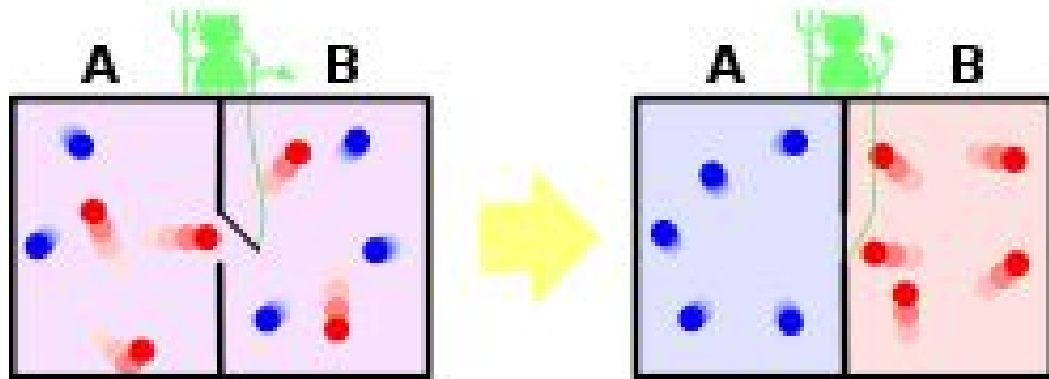
Ezt érdemes akkor megnézni amikor már fizikából az entrópia fogalma, és a termodinamika volt előadáson.

A lényeg az, hogy az információ-vesztés (pl törlés) fizikai entrópia-növekedést okoz!

Információ és fizikai entrópia

„in an isolated system, entropy never decreases.” Second Law of Thermodynamics

The Maxwell demon



Temperature of A **decreases**, B **increases** → Violation of the Second Law ?

az információ mértéke - entrópia

Fej vagy írás

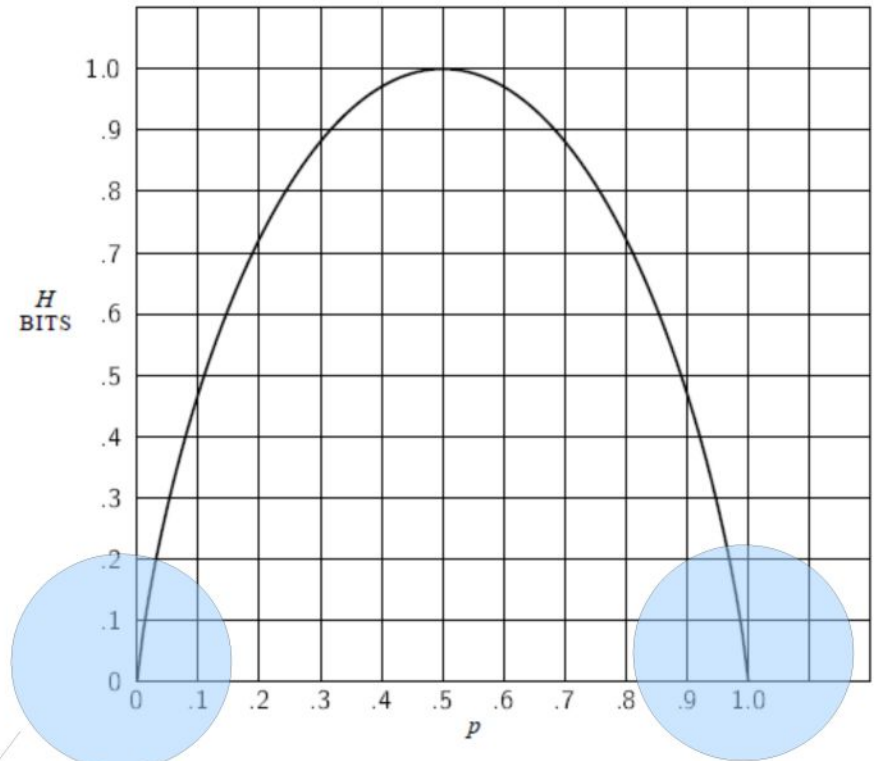


p



$q = 1-p$

H másik neve: **Shannon-entrópia**



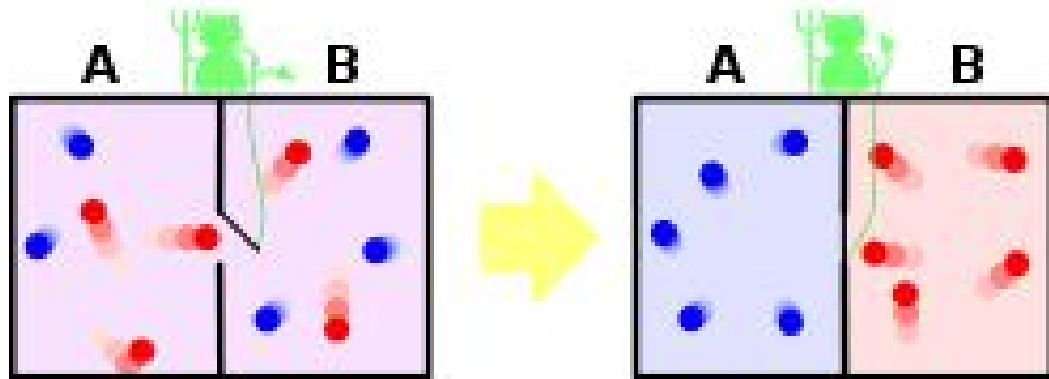
H CSAK akkor nulla, ha teljesen biztos mi fog történni: $p=0$ vagy $p=1$



A fizikai entrópia (S) CSAK akkor nulla, ha csupán 1 mikroállapot létezik

„in an isolated system, entropy never decreases.” Second Law of Thermodynamics

The Maxwell demon

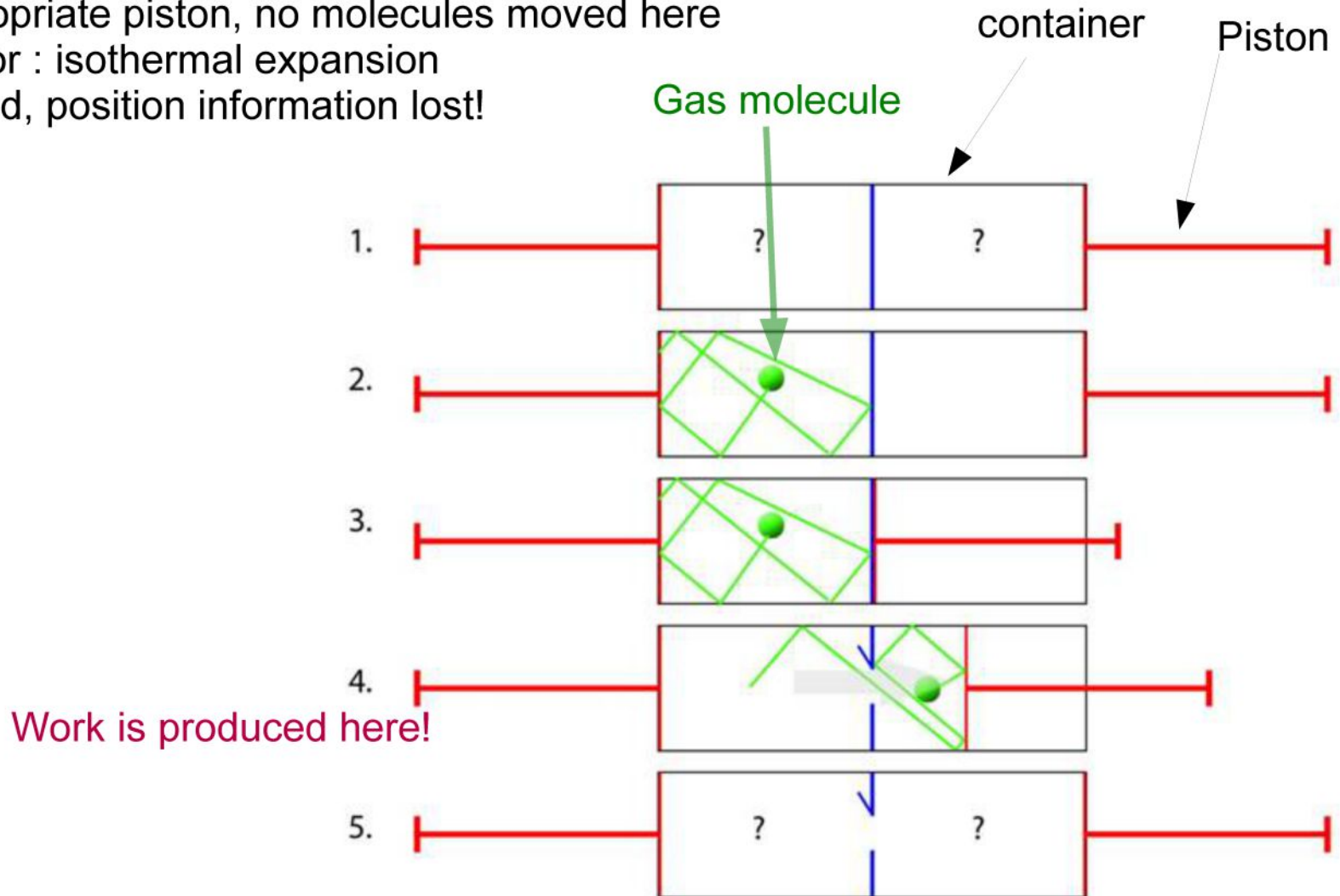


Temperature of A **decreases**, B **increases** → Violation of Law II. ?

Solution: NO, since the demon interacts with the system, it must be considered. The demon acquires **information**, and this changes its state!

Információ és fizikai entrópia

1. : molecule's position unknown
2. : measure position, information = 1 bit
3. : move appropriate piston, no molecules moved here
4. : release door : isothermal expansion
5. : door opened, position information lost!



Információ és fizikai entrópia

1. : molecule's position unknown
2. : measure position, information = 1 bit
3. : move appropriate piston, no molecules moved here
4. : **release door : isothermal expansion**
5. : door opened, position information lost!

Isothermal expansion:

$$W_{A \rightarrow B} = NkT \ln \left(\frac{V_A}{V_B} \right)$$

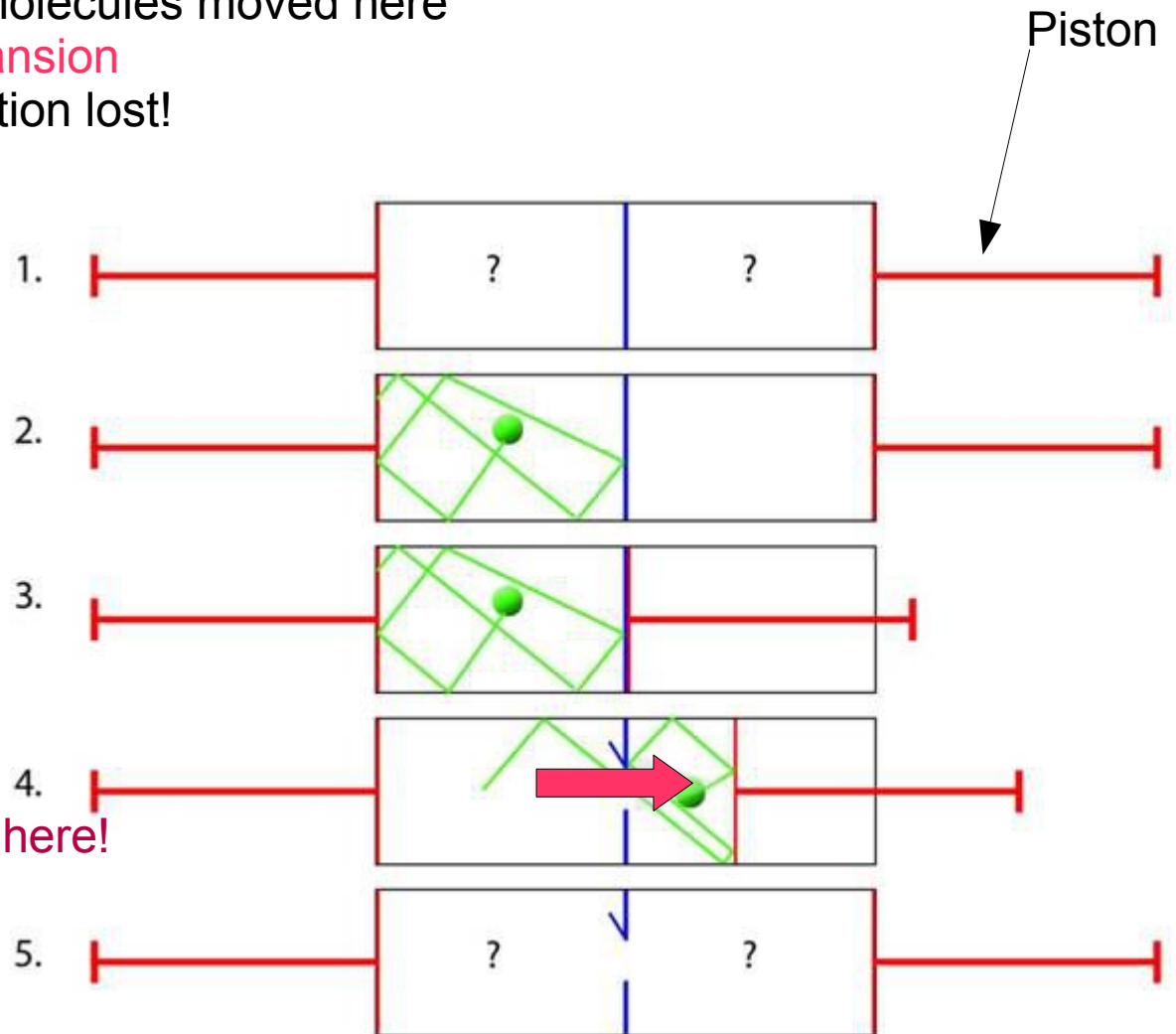
In this case:

$N=1$

$V_A/V_B = 2$

Hence

$W = kT \ln(2)$ Work is produced here!



Információ és fizikai entrópia

1. : molecule's position unknown
2. : measure position, information = 1 bit
3. : move appropriate piston, no molecules moved here
4. : release door : isothermal expansion
5. : **door opened, position information lost!**

Isothermal expansion:

$$W_{A \rightarrow B} = NkT \ln \left(\frac{V_A}{V_B} \right)$$

In this case:

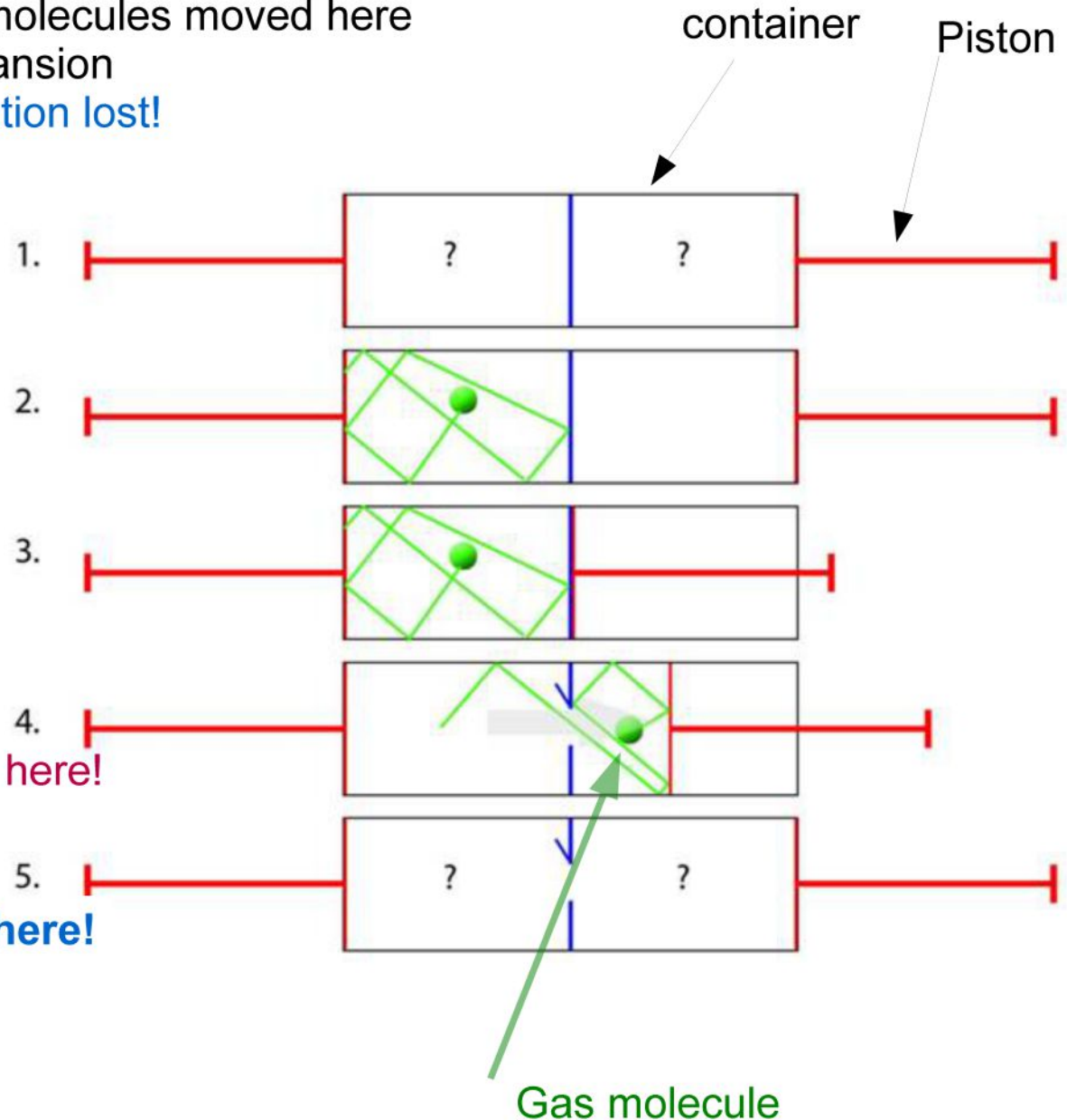
$N=1$

$V_A/V_B = 2$

Hence

$W = kT \ln(2)$ Work is produced here!

Information is lost here!



Információ és fizikai entrópia

Leo Szilárd:

From Law II. taking into account that $W = T\Delta S$

$$W_{\text{produced by piston}} = W_{\text{loss of information}}$$

$$T\Delta S_{\text{inf}} = kT \ln 2$$

$$\Delta S_{\text{1bit}} = k \ln 2$$

Erasing 1 bit of information increases physical entropy by $k \cdot \ln 2$

(Landauer 1971, logically irreversible processes, eg. AND-gate)

