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 → nyálelválasztás, 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

általánsságban

Információ **forrás**

A lehetséges események közül
melyik következett be?

kódolás

eseményeket SZÁMOKKAL reprezentálunk

Átviteli **csatorna**

dekódolás

SZÁMOKBÓL visszaállítjuk az eseményeketet

(át)vevő
cél(pont)

(hír)

Információátvitel - kódolás

általánsságban

Információ **forrás**

kódolás

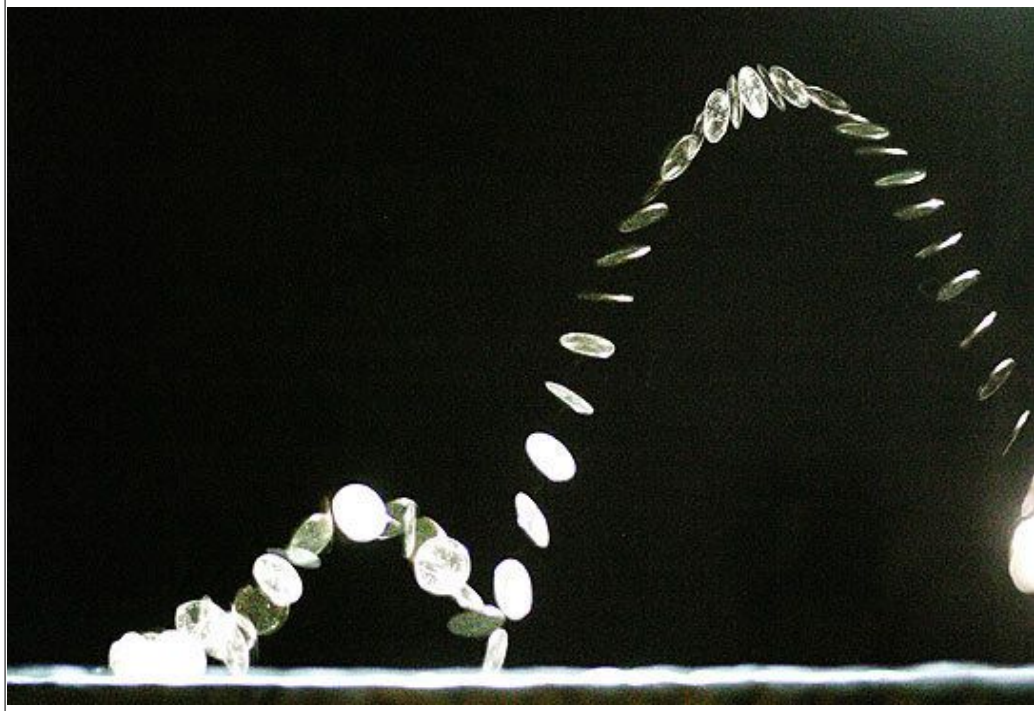
Átviteli **csatorna**

dekódolás

(át)vevő
cél(pont)

példa

Fej vagy írás?



Információátvitel - kódolás

általánsságban

Információ **forrás**

kódolás

Átviteli **csatorna**

dekódolás

(át)vevő
cél(pont)

A példában

Fej vagy írás?

kódolás
Fej/Írás → 1,0

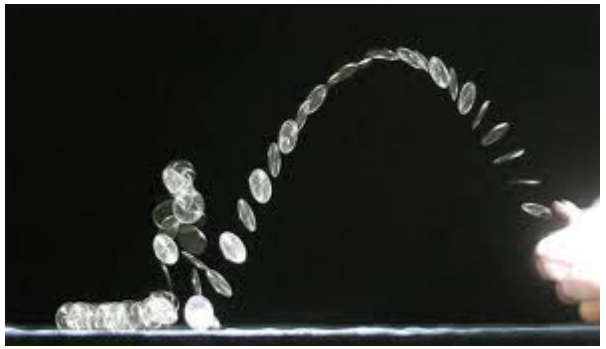
Beszéd, SMS, rádió, stb.

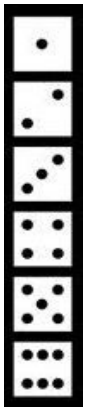
dekódolás
1,0 →Fej/Írás

Ki nyert?

Ugyanazt a **kulcsot** kell a kódoláshoz,
és a dekódoláshoz használni!

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?

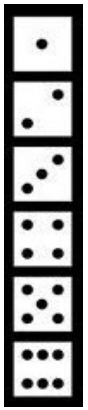
Bit: **b**inary dig**it**

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)

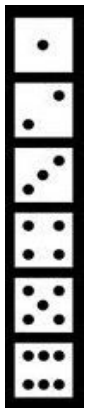
Pl: $101_2 = 1*2^2 + 0*2^1 + 1*2^0 = 4+0+1 = 5_{10}$

Információátvitel – **kódolási hatásfok**

Esemény		szám	digitális kód	Bitek száma	Maximális eseményszám
	:	1	1	**1**	**2**
	:	0	0		
	:	1	001	**3**	**8**
	:	2	010		
	:	3	011		
	:	4	100		
	:	5	101		
	:	6	110		
		7	111		
		0	000		

Itt 6 eseményt kódoltunk 3 biten, csakhogy $2^3=8$,
azaz 8 eseményt is lehetne kódolni ennyi bit segítségével

Információátvitel – **kódolási hatásfok**

Esemény	szám	digitális kód	Bitek száma	Maximális eseményszám
	: 1	001	**3**	**8**
	: 2	010		
	: 3	011		
	: 4	100		
	: 5	101		
	: 6	110		
	7	111	Itt 6 eseményt kódoltunk 3 biten, csakhogy 2^3=8,	
	0	000	azaz 8 eseményt is lehetne kódolni ennyi bit segítségével	

Egy jobb hatásfokú kódolás:

$\{X_1X_2X_3\}$ 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 hatásfok**

Az információtartalom definiálható a legjobb hatásfokú kódolással

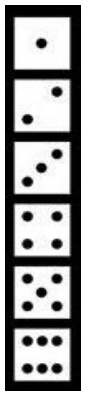
azaz:
Egy jel, esemény információtartalma megadható azzal, hogy minimálisan hány bitre van szükség az átvitléhez.
Ez egyben a kódolási hatásfok elméleti határa is.

Hogyan kapcsolódik ez az intuitív információtartalomhoz?

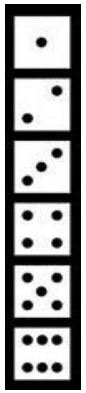
	p	q	
-Fej vagy írás?	½	½	nem tudjuk előre
-Ma reggel nincs dugó.	¼	¾	
-Holnap esni fog.	1%	99%	
-nyertem a lottón!	1/13,983,816	0.999....	valószínű nem nyert...

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

itt nem tudunk előre semmit, azaz a bizonytalanság maximális

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
			bitek számának várható értéke:		1.94

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

Átlagosan kevesebbet „tanulunk” a cinkelt kockával

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*I, azaz a bekövetkezési valószínűséggel súlyozott információtartalom.

információátvitel – az információ mértéke

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

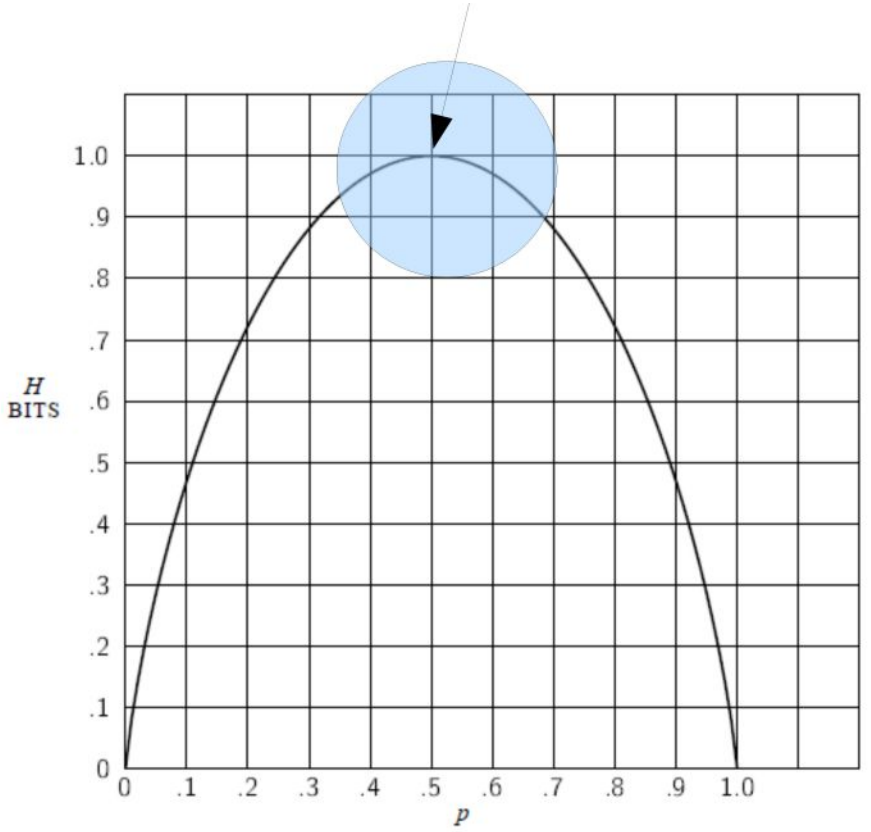
Fair érme: p = ½

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

Adatbázisok

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

Adatbázisok

Lehetne papíron is tárolni,
de ez nem hatékony

FOSTER CITY EYE CARE - OPTOMETRIC CENTER
PATIENT HISTORY QUESTIONNAIRE

Last name | First name | Mr.☐ Mrs.☐ Miss.☐ Ms.☐
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	Y/N
Other health problems				HIV/AIDS	Y/N
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

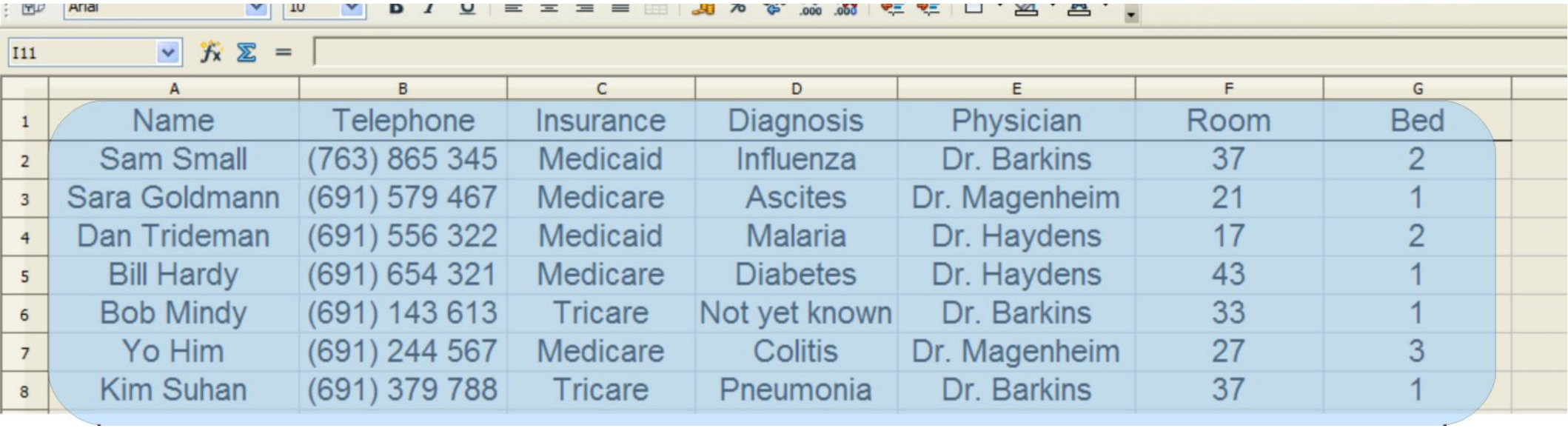
High blood pressure	Y/N Relation	Macular degeneration	Y/N Relation
Diabetes	Y/N Relation	Retinal detachment	Y/N Relation
Glaucoma	Y/N Relation	Cataracts	Y/N Relation
Other eye condition(s)	Y/N What kind?	Relation	

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	Dry eyes?	Y/N	Blurred vision?	Y/N
Other eye problems?	Y/N	What kind?					
Do you wear glasses?	Y/N	Contact lenses?	Y/N	Type			
Additional information		Are you interested in new contact lenses?					Y/N
Whom may we thank for referring you?							

Doctor's initials

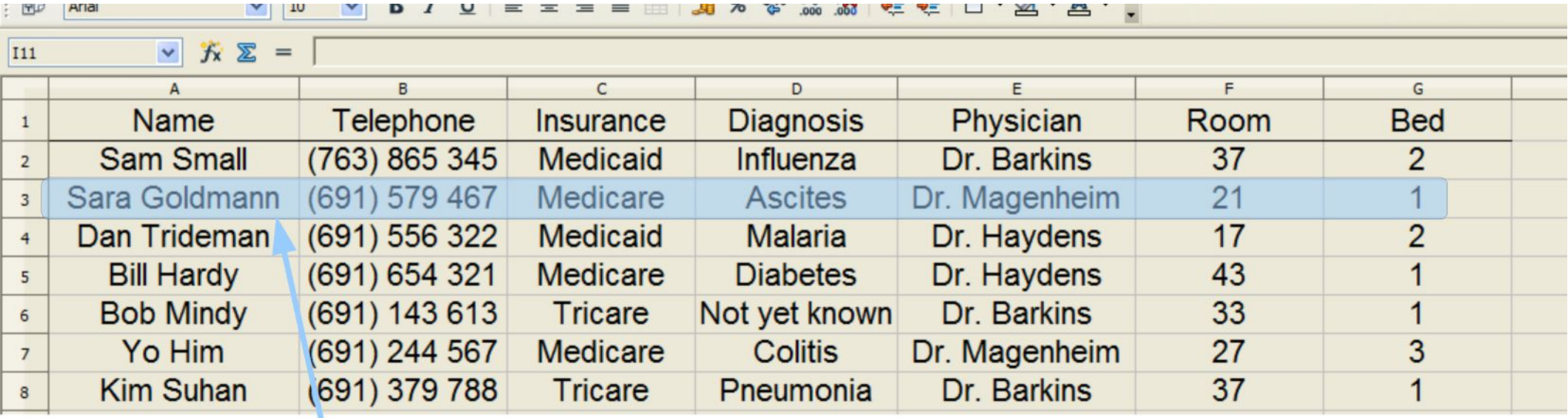
Adatbázisok - információtárolás



	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



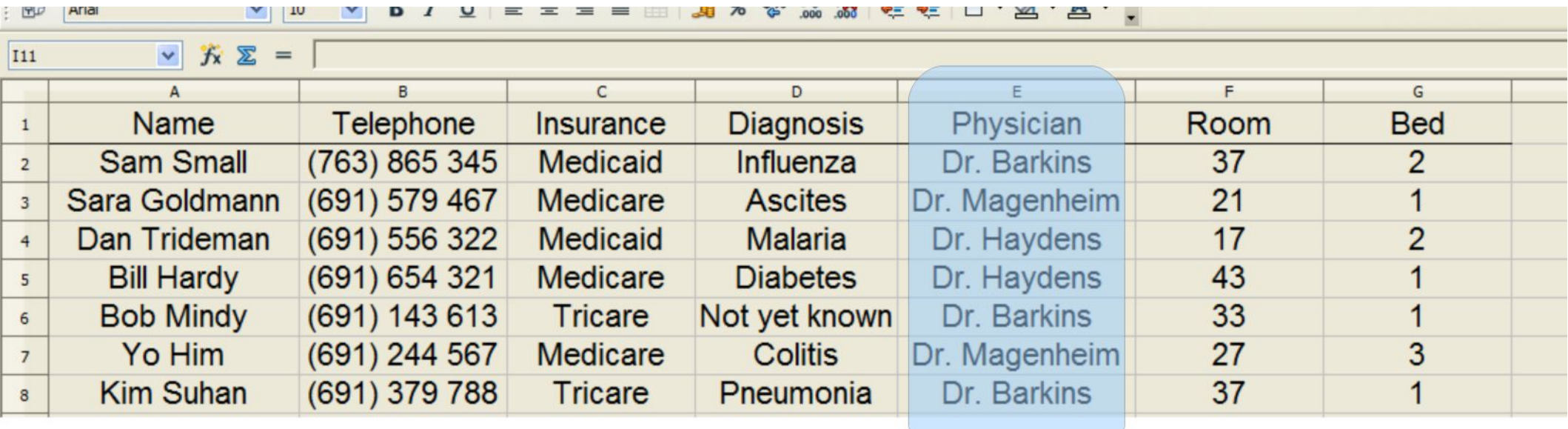
	A	B	C	D	E	F	G
1	Name	Telephone	Insurance	Diagnosis	Physician	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2
3	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1
4	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2
5	Bill Hardy	(691) 654 321	Medicare	Diabetes	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

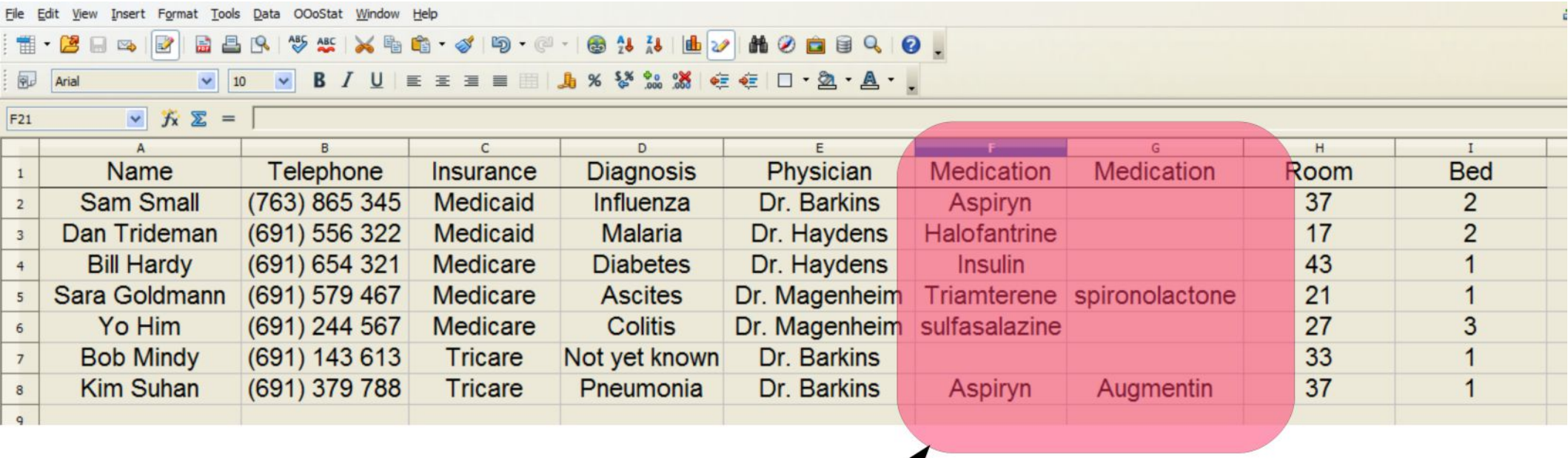
Adatbázisok - információtárolás



	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

Adatbázisok – problémák

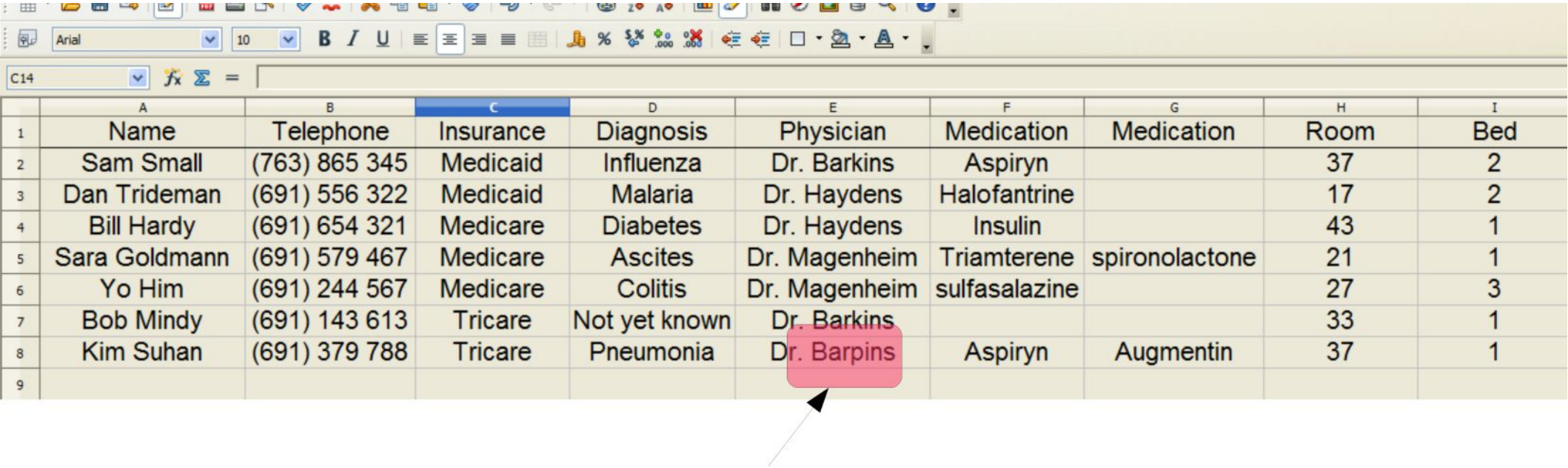


	A	B	C	D	E	F	G	H	I
1	Name	Telephone	Insurance	Diagnosis	Physician	Medication	Medication	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	Aspiryn		37	2
3	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	Halofantrine		17	2
4	Bill Hardy	(691) 654 321	Medicare	Diabetes	Dr. Haydens	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
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. 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
...

Adatbázisok – SQL

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 noninferential 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.