Informationstheorie

Begriff der Information (mit Beispielen)

Informationsgehalt von Daten und Datenströmen

Entropie und Information

Ich wünsche bei schweren Entscheidungen so ruhig bleiben zu können wie JB!

Begriff der Information (mit Beispielen)

Intuitive Begriffe:

"informare" (Lat.) : „die Gedanken formen/beeinflussen”, oder jemandem belehren.

Also, wir lernen nur dann wenn wir Informationen bekommen.

Oder:
„ein Zufuhrstyp in ein Lebewesen oder Gerät”
(z.B. Sicht von Futter → Bewegung in die gegebene Richtung)

Oder:
„Information ist ein Muster, was andere Mustern beeinflusst.”
(RNS/DNS Sequenz → Eiweisstoffsfunktion)

Ich wünsche bei schweren Entscheidungen so ruhig bleiben zu können wie JB!

Informationsübertragung – Informationsgehalt

Ereigniss und Information:

Ereignisse mit unterschiedlichem Informationsgehalt:

-Heute ist -wie fast immer - wieder Stau.

-Gestern hat geregnet.

-Ich habe den Hauptgewinn bekommen!

Wie können wir Informationen *kodieren*?

Sprache, Schrift, Zeichnung,...

Informationsübertragung – Kodierung

generell

Informations**quelle**

Welches aus den möglichen Ereignisse ist aufgetreten?

Kodieren: Wir stellen **Ereignisse** mit **ZAHLEN** dar.

Übertragungskanal

Dekodieren: Wir rekonstrieren die **Ereignisse** aus den **ZAHLEN**

Informationsempfänger
Ziel

(Nachricht)

Informationsübertragung – Kodierung

generell

Informations**quelle**

Kodierung

Übertragungskanal

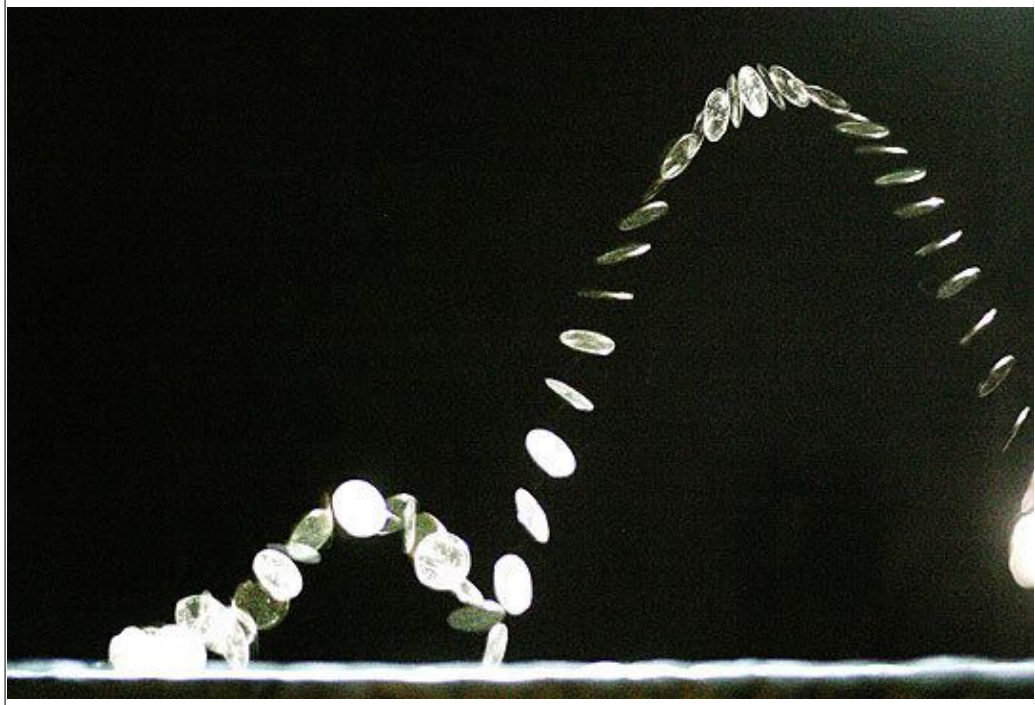
Dekodierung

Informationsempfänger

Ziel

Ein Beispiel

Münzenwerfen



Kopf oder Zahl?

Informationsübertragung – Kodierung

generell

Informations**quelle**

Kodierung

Übertragungskanal

Dekodierung

Informationsempfänger

Ziel

Beispiel

Kopf oder Zahl

Kodierung

Kopf oder Zahl
-> Zahlen: 1,0

Sprache,Zeichnen,SMS,email,...

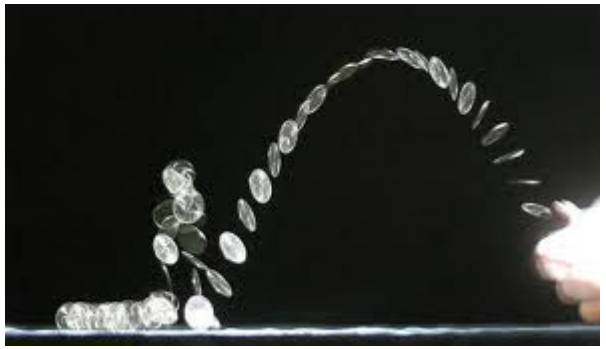
Zahlen: 1,0 ->
Kopf oder Zahl

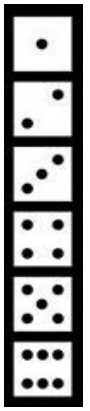
Dekodierung

Wer hat gut geraten

Bei der Kodierung/Dekodierung muss das selbe „Wörterbuch” benutzt werden.
Zahlen können mit Rechenmaschienen aufgearbeitet, gespeichert, etc. werden.

Informationsübertragung – digitale Kodierung

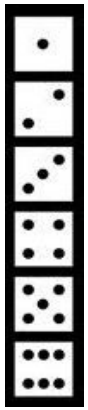


Ereignis		Zahl	digitaler Kode
	:	1	1
	:	0	0
	:	1	001
	:	2	010
	:	3	011
	:	4	100
	:	5	101
	:	6	110

2-Basis Zahlensystem, Beispiel: $101_2 = 1*2^2+0*2^1+1*2^0 = 5_{10}$

bit = „binary digit”

Informationsübertragung – Kodierungseffizienz

Ereignis		Zahl	digitaler Kode	Anzahl d. Bits	Maximaler Anzahl der Ereignisse
	:	1	001		
	:	2	010		
	:	3	011	**3**	**8**
	:	4	100		
	:	5	101		
	:	6	110		
		7	111		
		0	000		

Wir haben nur 6 Ereignisse,
Könnten aber 2^3=8 Kodieren

Eine bessere Kodierung:

{$X_1X_2X_3$} Stellen wir jeweils 3 Ereignisse zusammen:

Ohne extra Aufwand wäre
das einfach 3x3 bits = **9** bits

→ 1 Bit weniger!

Wir haben aber insgesamt
6^3 = 216 Möglichkeiten, wenn
wir 3-mal würfeln.
Weil 2^8 = 256, sind 8 Bits mehr
als genug! (7 wäre zu wenig)

Informationsübertragung – Informationsgehalt

Wir haben aber insgesamt $6^3 = 216$ Möglichkeiten, wenn wir 3-mal würfeln.
Weil $2^8 = 256$, sind 8 Bits mehr als genug! (7 wäre zu wenig)

Informationsgehalt := wie viele Bits *minimal* notwendig sind.

Dieser Zahl kann auch ein Bruchzahl sein, wenn wir Bit/Ereignis nehmen. Nochdazu, der gibt auch ein maximaler Effizienz an: die bestmögliche Kodierungsprozess benutzt genau so viele Bits. Alle andere Methoden brauchen mehr Bits für die Übertragung. (also sind weniger effektiv)

Informationsübertragung – Informationsgehalt

Informationsgehalt := wie viele Bits *minimal* notwendig sind.

Wie verbindet sich das mit der intuitiven Gedanken über wie viel wir von einem Nachricht lernen?

-Heute ist -wie fast immer - wieder Stau.

-Gestern hat geregnet.

-Ich habe den Hauptgewinn bekommen!

Informationsübertragung – Informationsgehalt

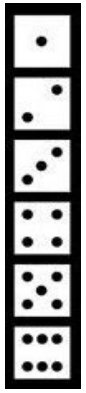
Wie verbindet sich das mit der intuitiven Gedanken über wie viel wir von einem Nachricht lernen?

	p	q=1-p	
-Kopf oder Zahl?	½	½	← Keine Ahnung
-Heute ist -wie fast immer - wieder Stau.	¾	¼	
-Gestern hat geregnet.	10%	70%	
-Ich habe den Hauptgewinn bekommen!	1/13,983,816	0.999...	← Fast sicherlich nicht gewonnen

bekommene Information ist umgekehrt proportional zu der Wahrscheinlichkeit (p)

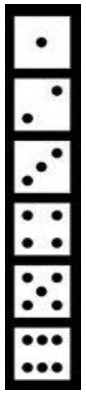
Ganz intuitiv: wenn etwas sehr selten vorkommt, (also p ist klein), dann, wenn es doch irgendwie passiert dann ist das ein großer Nachricht, als wenn etwas häufiger vorkommt, und so passiert es wieder mal.

Informationsübertragung – Informationsmaß

Fair	P_i	Wahrscheinlichkeit	Kodierungs beispiel	benötigte Bits	p*(Anzahl der Bits)
	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
			Erwartungswert der Bitmenge		**3**

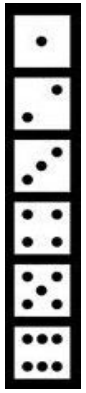
Falsch P_i

Wir können in diesem Fall besser Kodieren:

	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
			Erwartungswert der Bitmenge		**1.94**

Eine bessere Kodierung Nutzt aus, das manche Ereignisse kommen selten vor, aber andere öfter.
Die seltene Ereignisse kodieren wir mit längeren „Wörter", dafür aber die often vorkommende mit kürzeren. So wird der durchschnittlicher Anzahl der Bits kleiner.

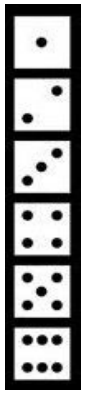
Informationsübertragung – Informationsmaß

Fair	P_i	Wahrscheinlichkeit	Kodierungs beispiel	benötigte Bits	p*(Anzahl der Bits)
	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
			Erwartungswert der Bitmenge		**3**

Hier wissen wir nichts im voraus, alles ist gleich möglich

Falsch P_i

Wir können in diesem Fall besser Kodieren:

	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
			Erwartungswert der Bitmenge		**1.94**

Hier „lernen" wir weniger, denn wir wissen schon im voraus, das 1 und 2 sehr häufig vorkommen.

Informationsübertragung – Informationsmaß

Shannon : definieren wir H mit $$H = p \cdot \log_2\left(\frac{1}{p}\right)$$

Man kann auch den Informationsgehalt eines einzigen Ereignisses ausrechnen:

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

Also H= p*I, somit ist H ein Erwartungswert, oder Mittelwert.
(wenn wir über alle Möglichkeiten summieren)

$\log_2$: 2-Basis logaritmus

Beispiele:

$\log_2(2) = 1$
$\log_2(4) = 2$
$\log_2(8) = 3$

Informationsübertragung – Informationsmaß

Shannon

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

Wenn wir mehrere Möglichkeiten haben, dann ist H eine Summe:

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

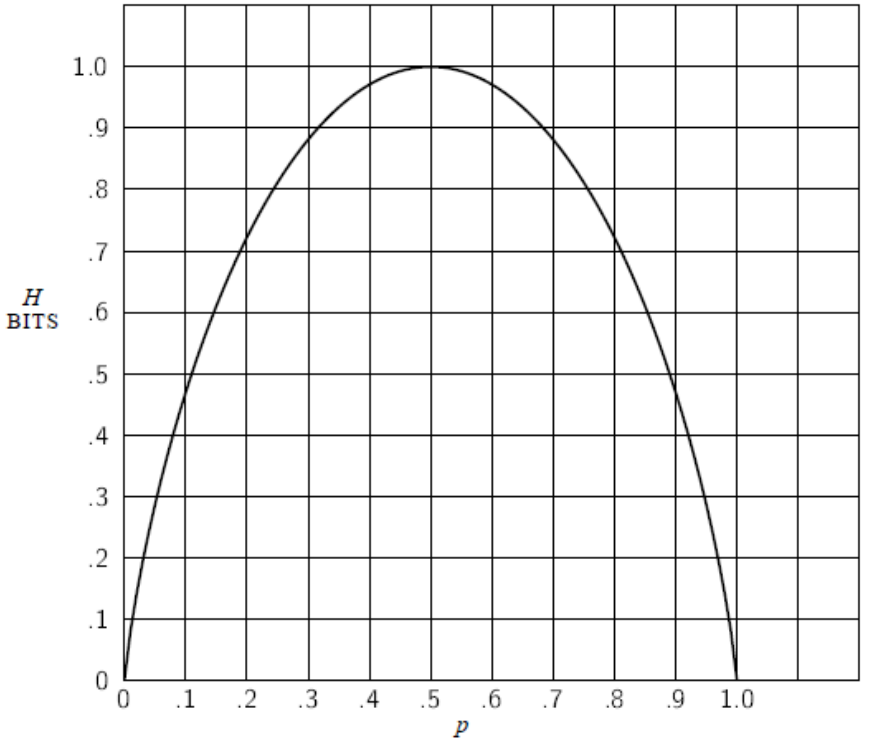
andere log-basis Möglichkeiten:
$\log_e$ (ln) : [nat]
$\log_{10}$ (lg) : [ban]

Informationsmaß - Entropie

Kopf oder Zahl?

p

q = 1-p



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

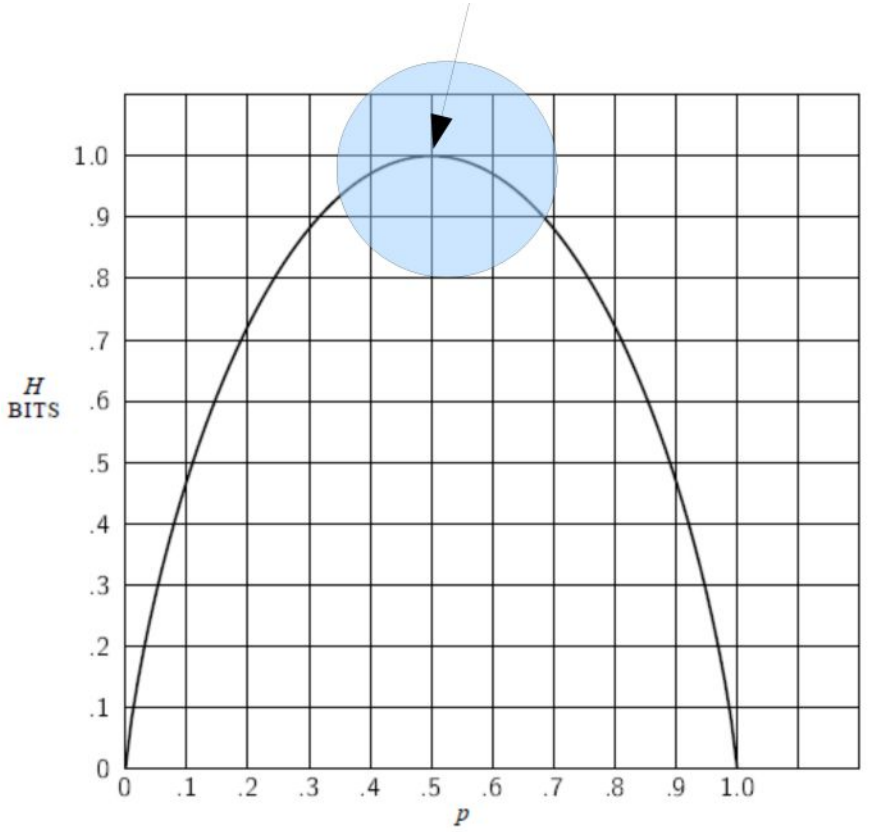
Informationsmaß - Entropie

faire Münze: p = ½

Keine Erwartungen
maximale Unsicherheit

Kopf oder Zahl?

p q = 1-p



H deshalb wird auch benannt als:
Shannon-entropy

Wenn wir nichts im voraus wissen, und so alle Ereignisse sind gleich möglich, dann ist der Anzahl der erwarteten Ereignisse maximal.

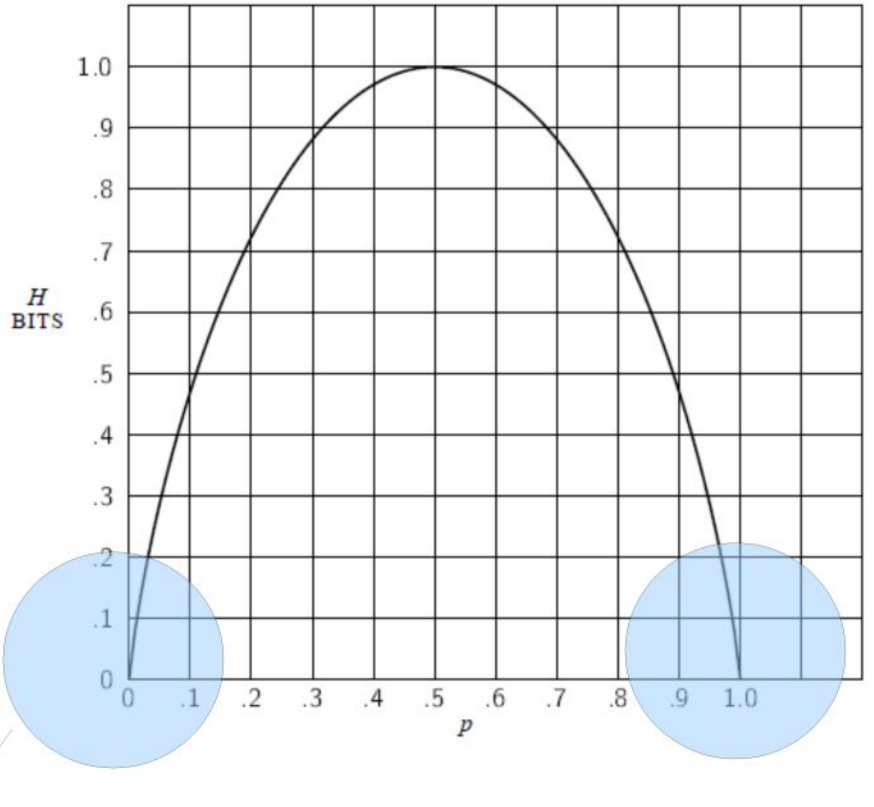
In der Physik ist Entropie dann maximal, wenn der Anzahl der Mikrozustände maximal ist.

Informationsmaß - Entropie

Kopf oder Zahl?

p

q = 1-p



H deshalb wird auch benannt als:

Shannon-entropy

H ist nur dann 0, wenn es nur eine Möglichkeit gibt: p=0 oder p=1

Physikalische Entropie (S) ist nur dann 0 wenn nur 1 Mikrozustand möglich ist.

Datenbasen

Datenbasen speichern Information.

Speichern ist nicht genug, wir möchten auch aufsuchen, wiedergeben, etc.

Datenbasen

Auf dem Papier ist alles sehr Aufwändig.

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					

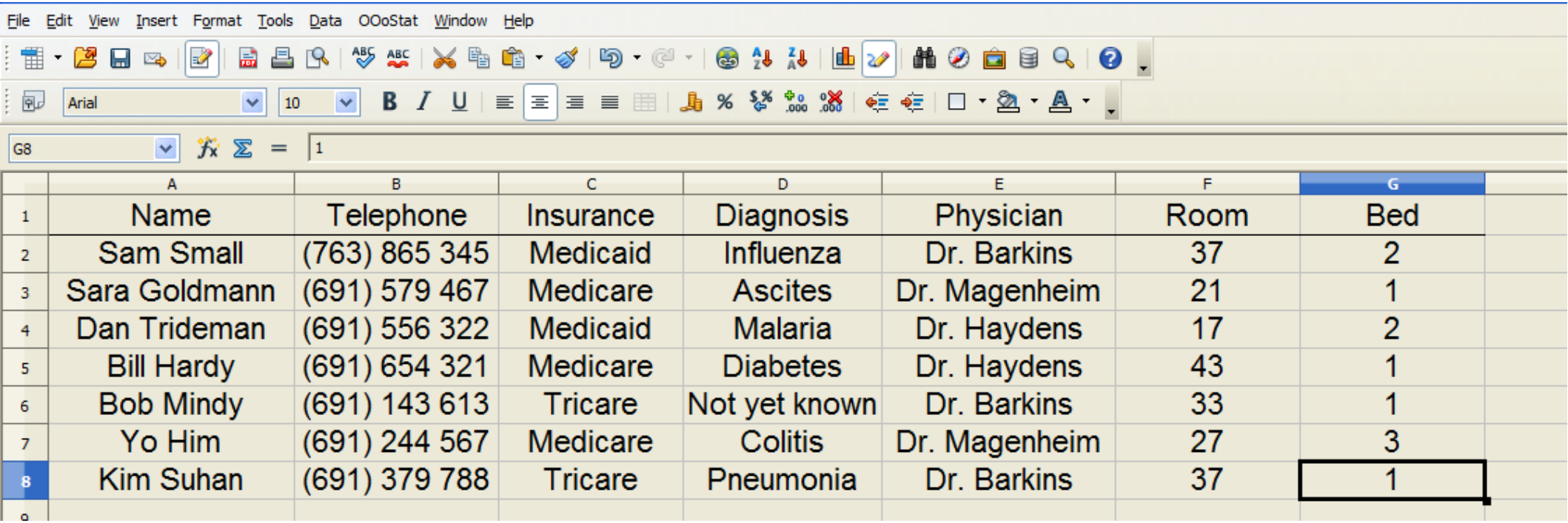
FAMILY HISTORY

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

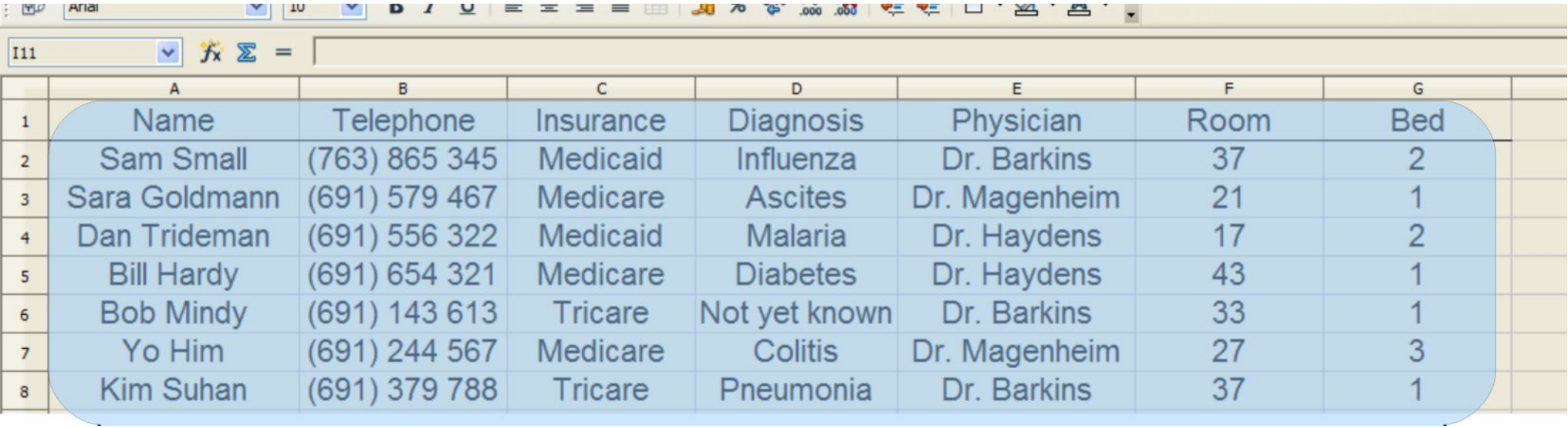
PERSONAL EYE INFORMATION

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



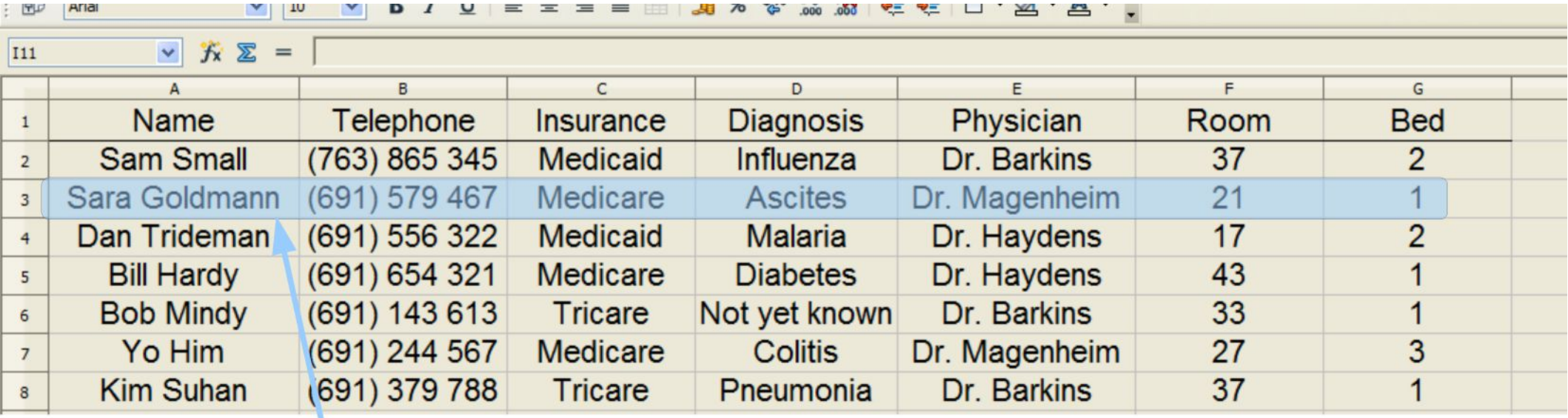
	A	B	C	D	E	F	G
1	Name	Telephone	Insurance	Diagnosis	Physician	Room	Bed
2	Sam Small	(763) 865 345	Medicaid	Influenza	Dr. Barkins	37	2
3	Sara Goldmann	(691) 579 467	Medicare	Ascites	Dr. Magenheim	21	1
4	Dan Trideman	(691) 556 322	Medicaid	Malaria	Dr. Haydens	17	2
5	Bill Hardy	(691) 654 321	Medicare	Diabetes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



	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

Tabelle : geordnete Datenmenge (Informationen)

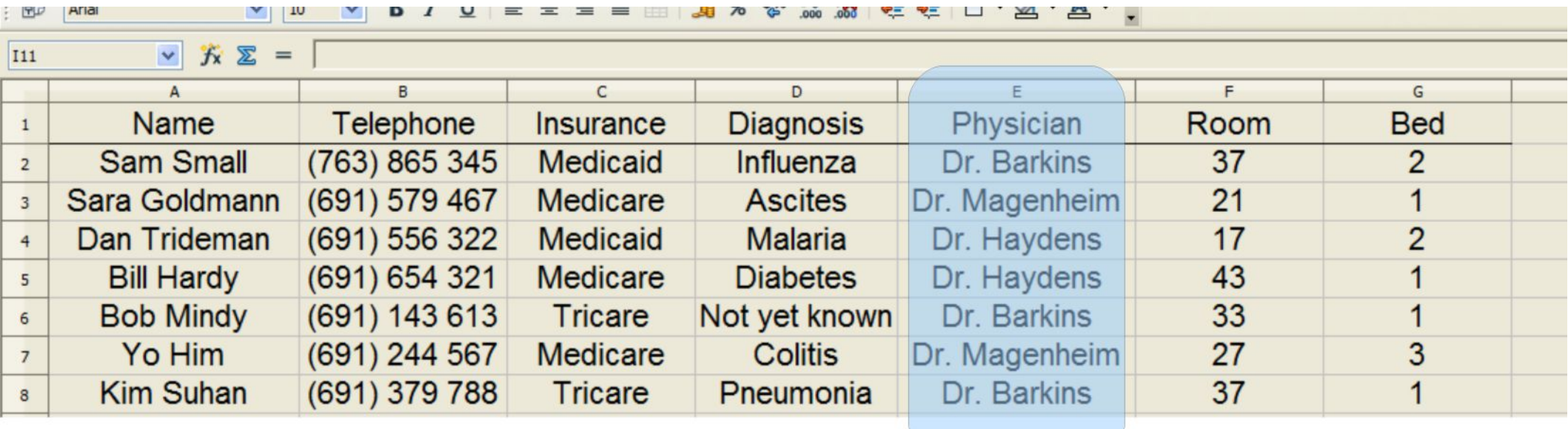
Databases – storing information



	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

„Tupel” oder Datensatz (record) :

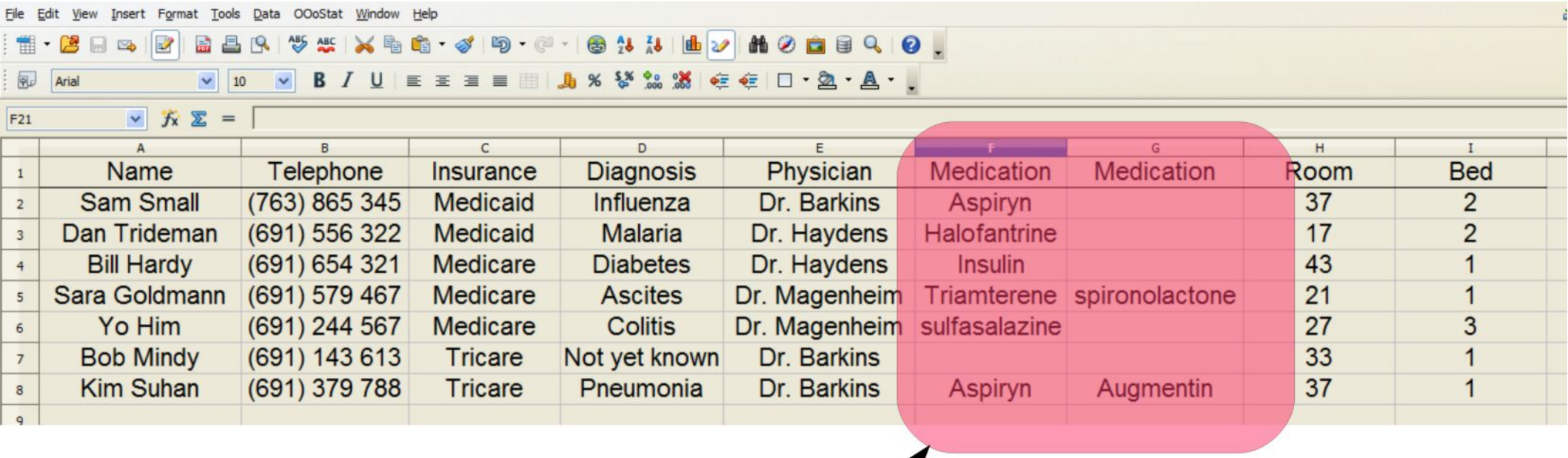
Eine Zeile in der Tabelle, ist ein Datensatz.
Jeder Datensatz hat dieselbe Struktur



	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

Spalte: Datentyp

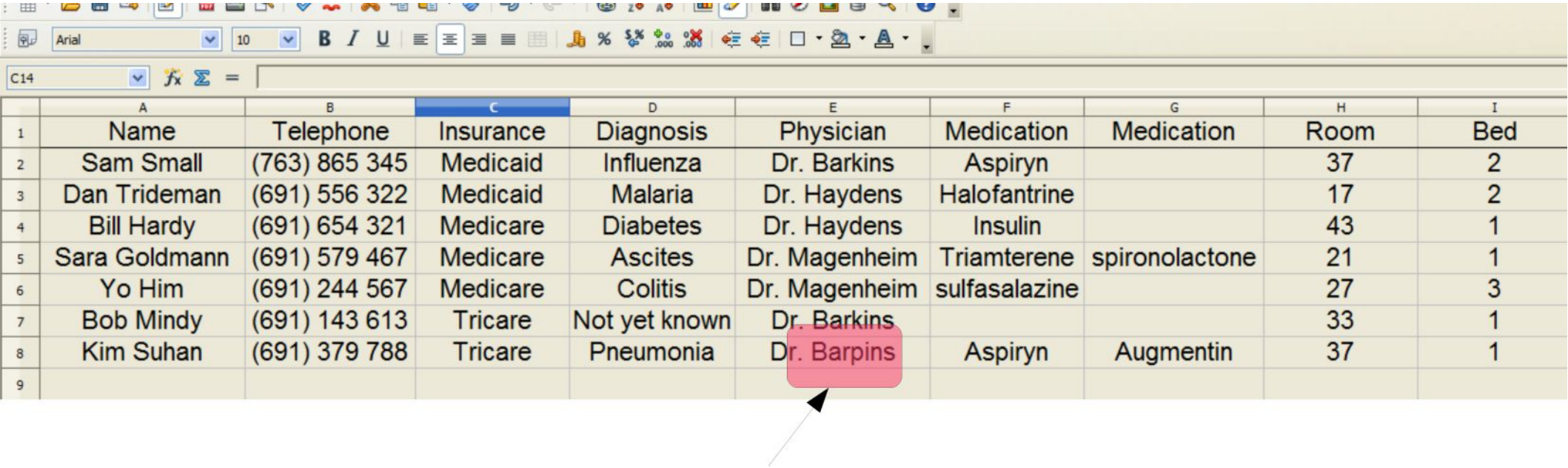
Es gibt aber probleme mit solchen Tabellen



	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									

Datensätze können unterschiedliche größe haben

Ungünstig
Wenn etwas leer ist: absichtlich, oder fehlerhaft?



	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									

Dieselben Daten sind mehrmals gespeichert:

Typos
Später können wir nicht mehr alle ändern (einfach zu viele)

Databases – SQL

Structured Query Language:

Strukturierte Aufsuchssprache.
Relationelle Datenbanke.

Die Relationsschemen legen die Verbindungen fest.

Ein Information ist nur EINMAL gespeichert!

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