CS213
Principles of Database Systems(H)

Chapter 1

Shiqi YU 于仕琪

yusq@sustech.edu.cn

Most contents are from Stéphane Faroult's slides

About Me

Shiqi Yu 于仕琪

*** Office: Room 312, South Tower, CoE Building(工学院南楼312)**

*** Email: yusq@sustech.edu.cn**

The Course

Textbook: No textbook required

- You can search online for your questions.
- Most points are included in the slides.

The Course

Grading

- Lab attendance: 10%
- Assignments: 20%
- Projects: 30% (two projects)
- Final exam: 40%

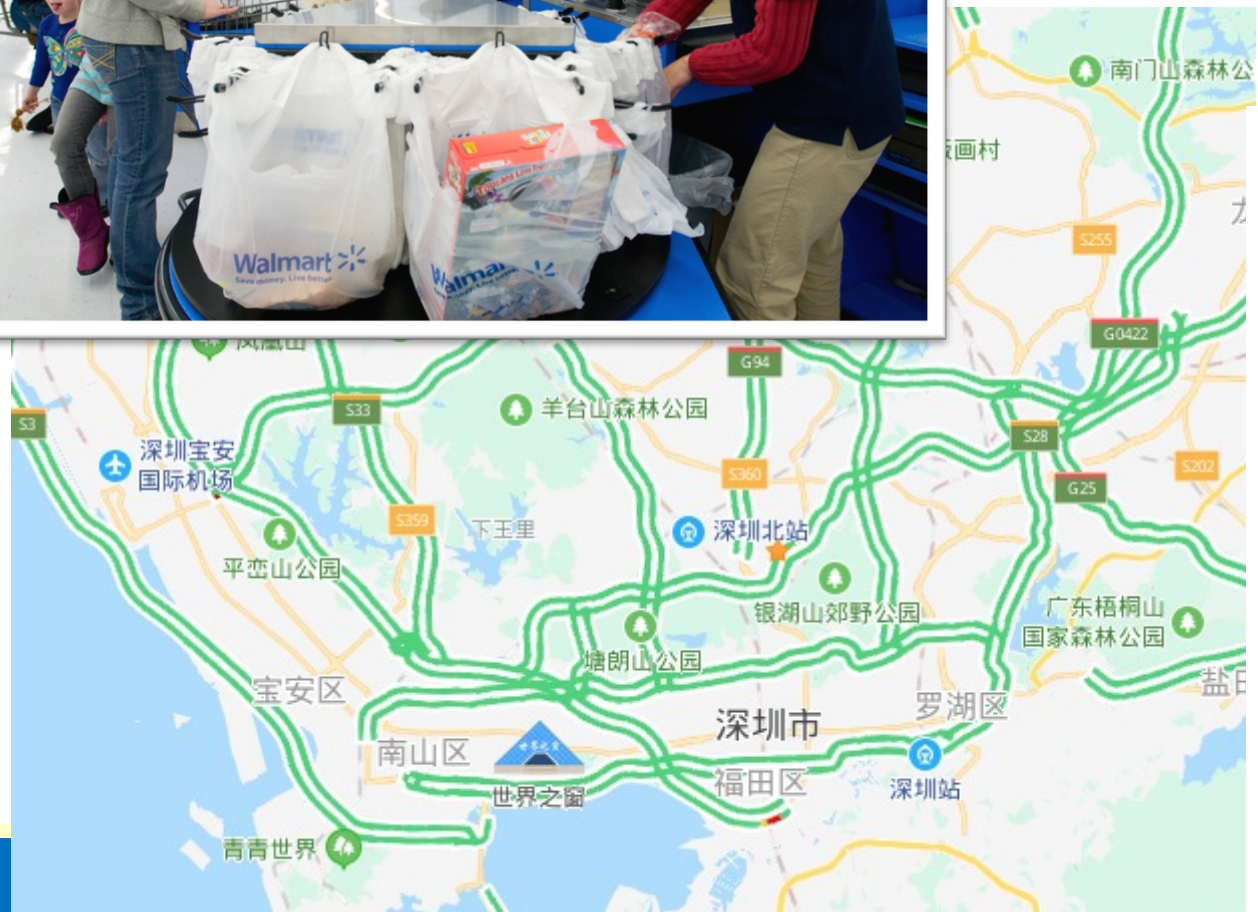
1. Introduction to Databases

Shiqi Yu 于仕琪

yusq@sustech.edu.cn

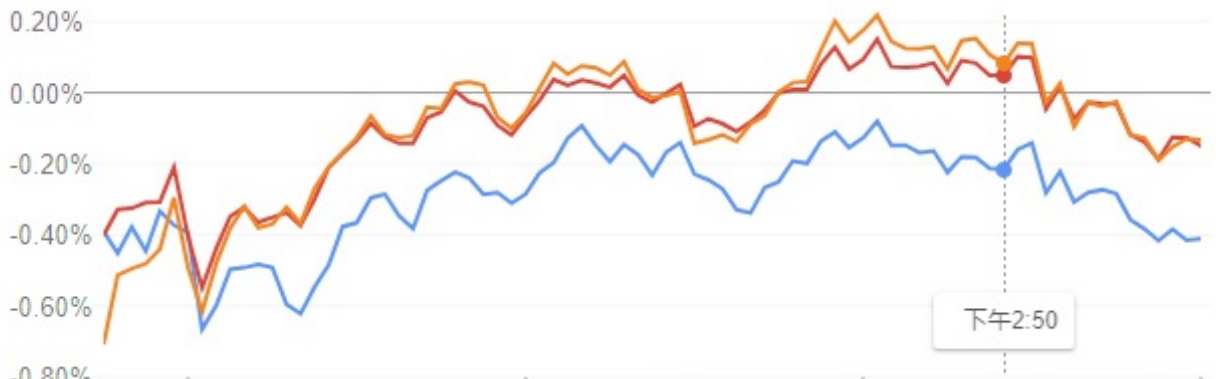
Most contents are from Stéphane Faroult's slides

What are in Databases?

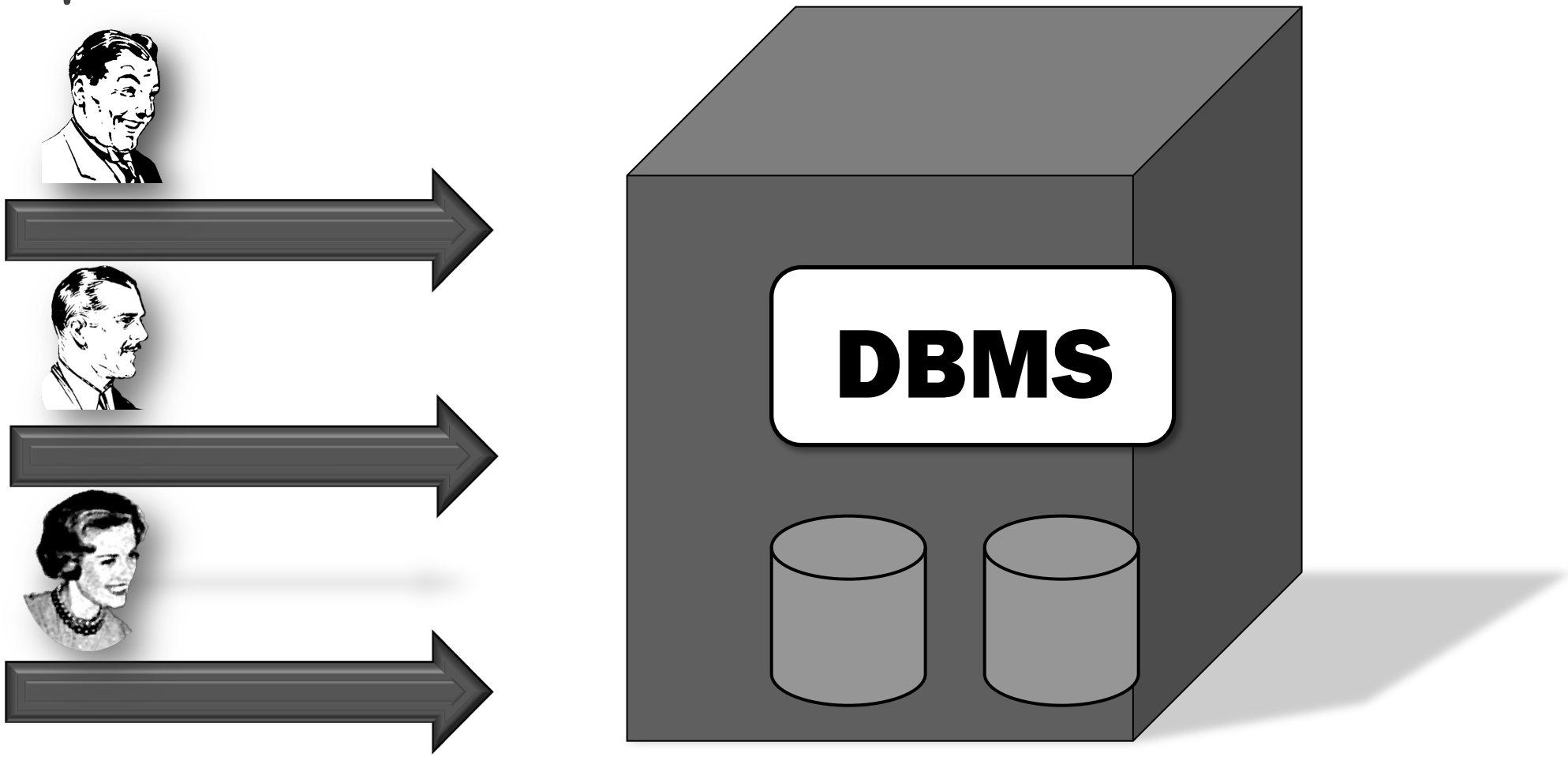


Dow Jones Industrial Average	.DJI	29,487.71	-63.71 (-0.22%)
S&P 500 Index	.INX	3,381.11	+1.66 (0.05%)
Nasdaq Composite	.IXIC	9,733.96	+7.99 (0.08%)

1天 5 天 1个月 6 个月 YTD 1年 5 年 最大



Databases are everywhere today
but the concept is old.



The idea was to have one system doing once and for all the
boring data storage/retrieval part.

Edgar F. Codd

1923 - 2003

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

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.

E. F. Codd, A Relational Model of Data for Large Shared Data Banks, Information Retrieval, June, 1970

ORACLE

Larry Ellison, Bob Miner and Ed Oates founded Oracle Corporation in 1977 and under the name Software Development Laboratories (SDL).

Ellison took inspiration from Edgar Codd's paper on relational databases.

Ellison was listed by Forbes magazine as the fourth-wealthiest person in the United States and as the sixth-wealthiest in the world, with a fortune of $69.1 billion in 2019.

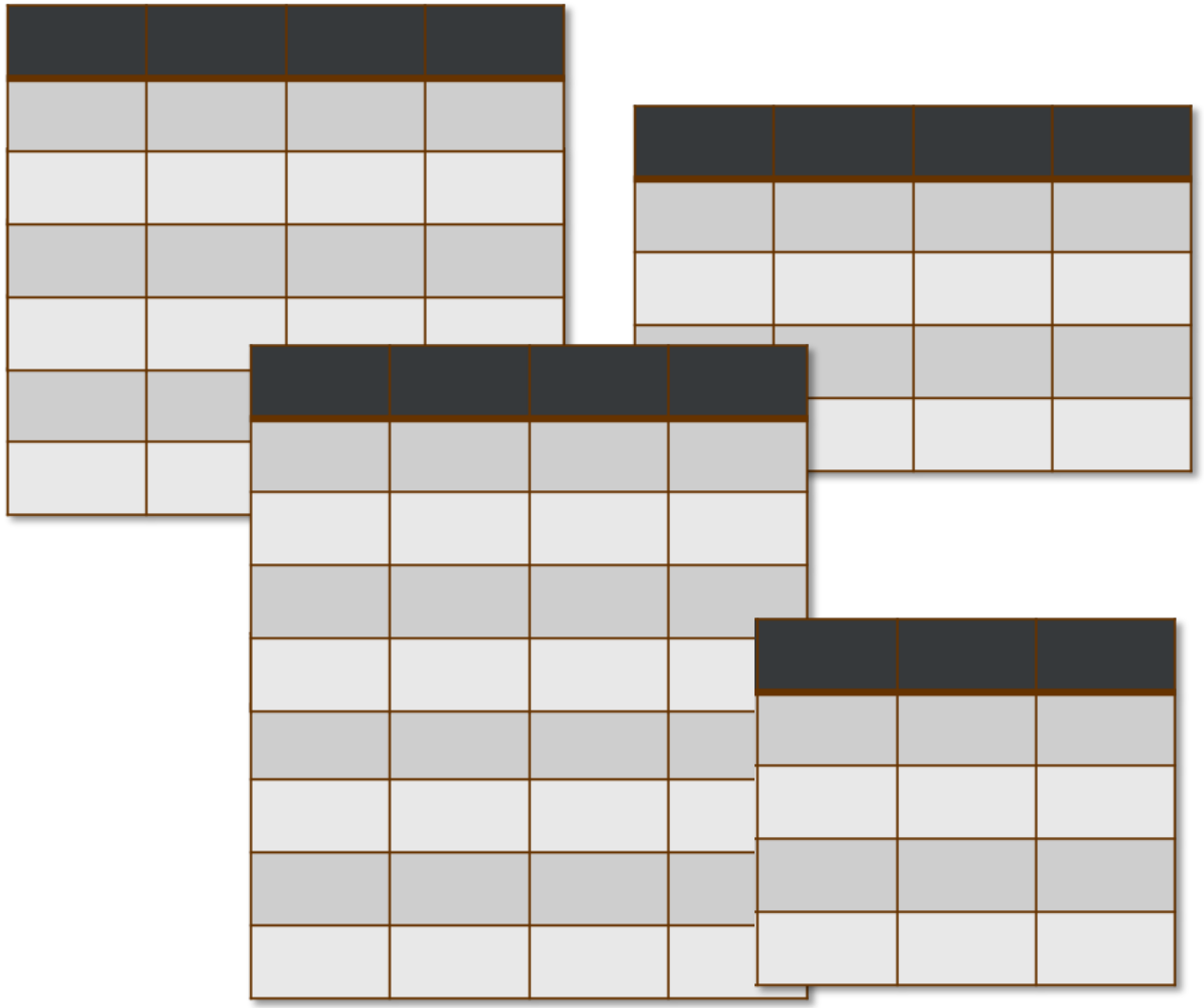
Larry Ellison

https://en.wikipedia.org/wiki/Oracle_Corporation

Relational Database

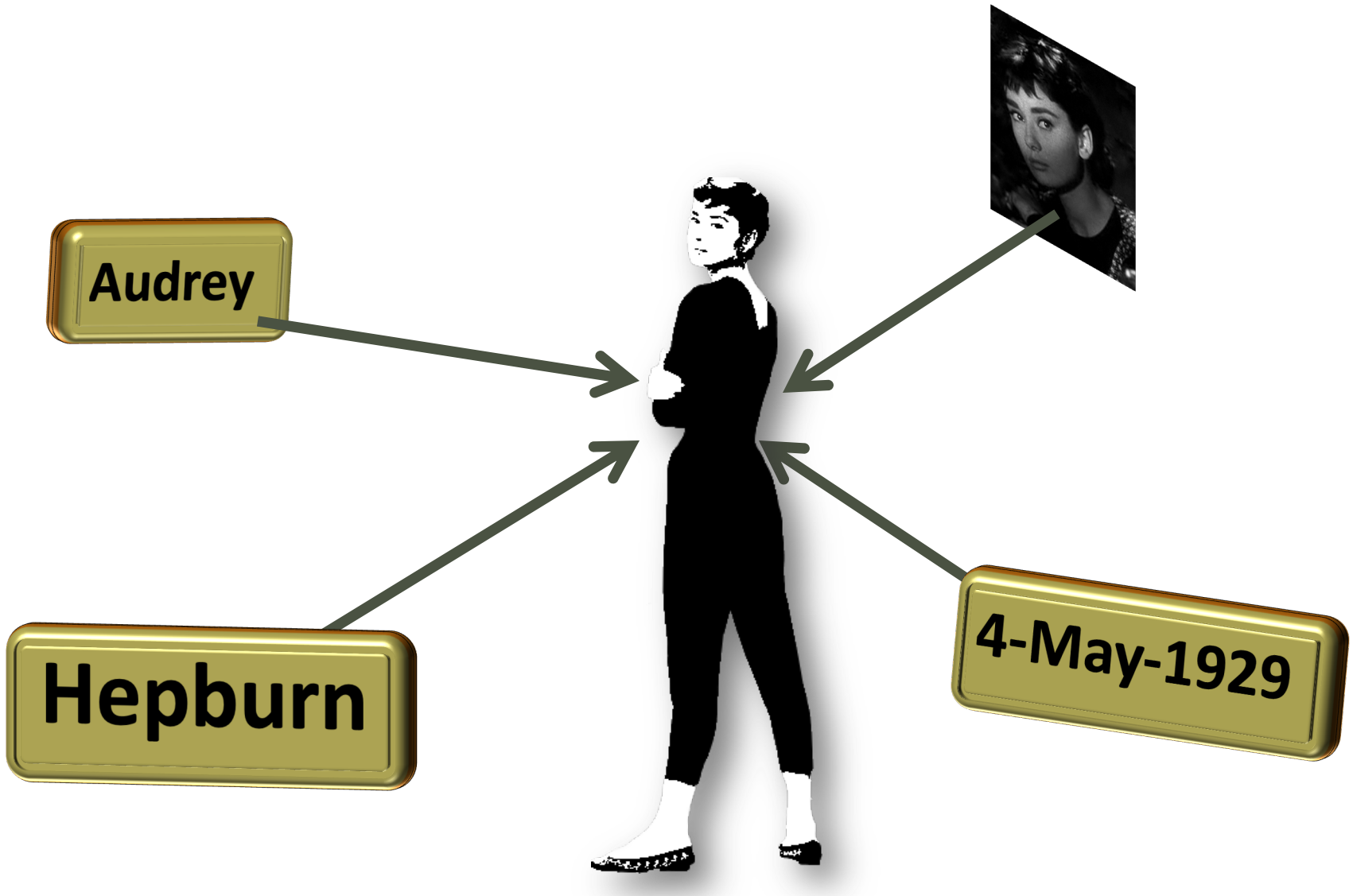
Based on the relational model of data

- Organizes data into one or more tables
- Rows are also called records or tuples
- Columns are also called attributes

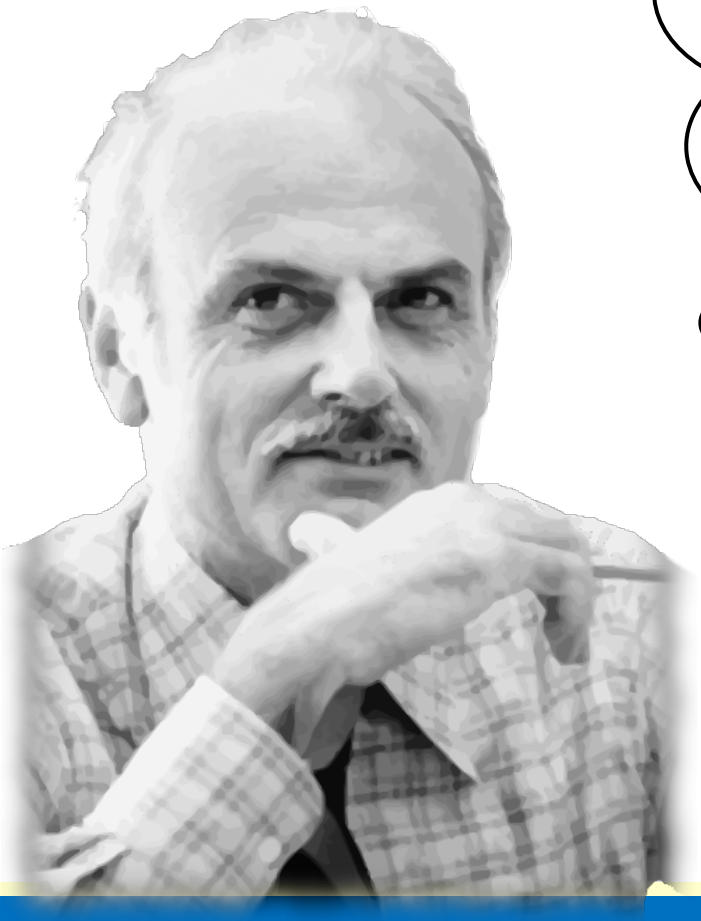


Surname	*Firstname*	*Birthdate*	*Picture*
Hepburn	**Audrey**	**4-May-1929**	

Each column in the table stores a piece of data, and one row represents a "known fact": Audrey Hepburn was born on 1929/05/04 and looked like this.



All the pieces of data in a row are related, hence "relational".

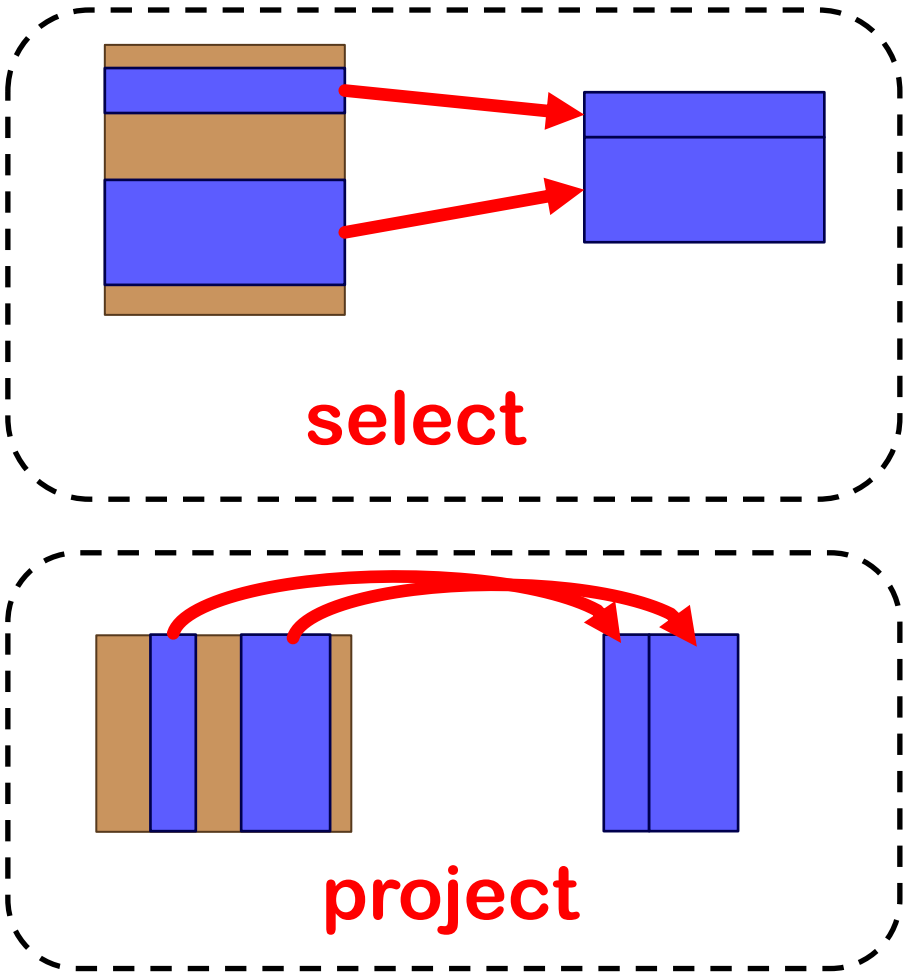


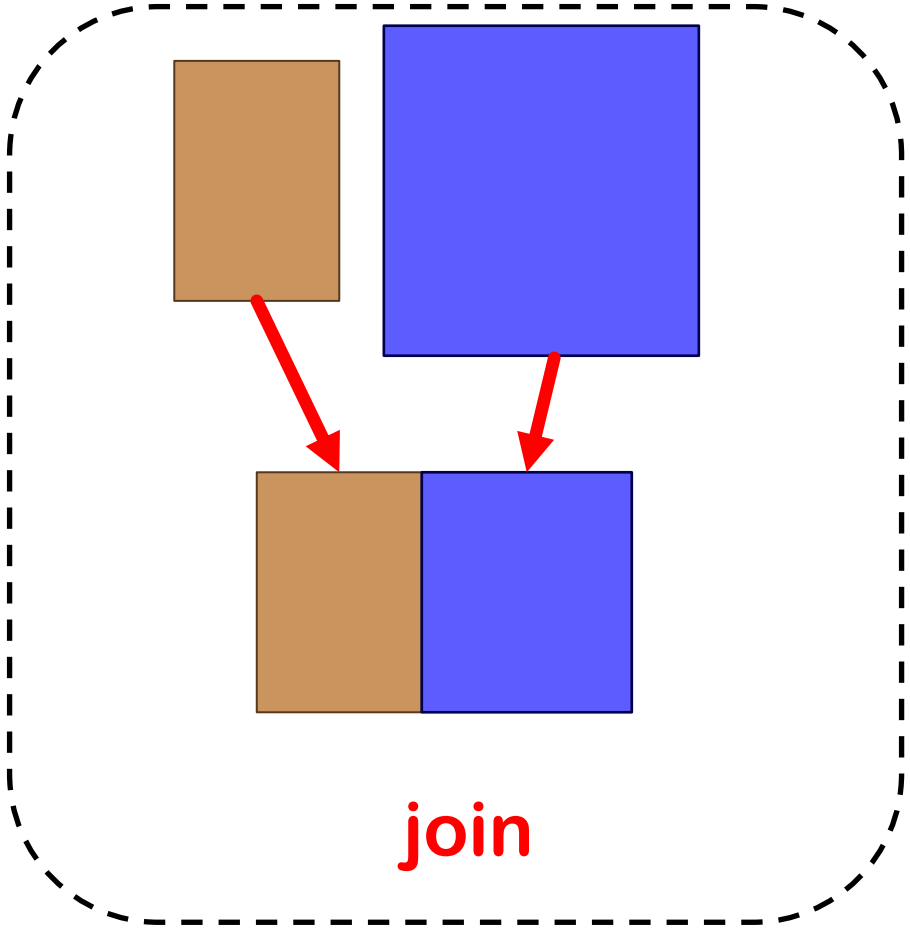
OPERATE

on relations

But Codd's big idea was that as relations are well known mathematical sets, you could operate on them, and get new sets.

Operations





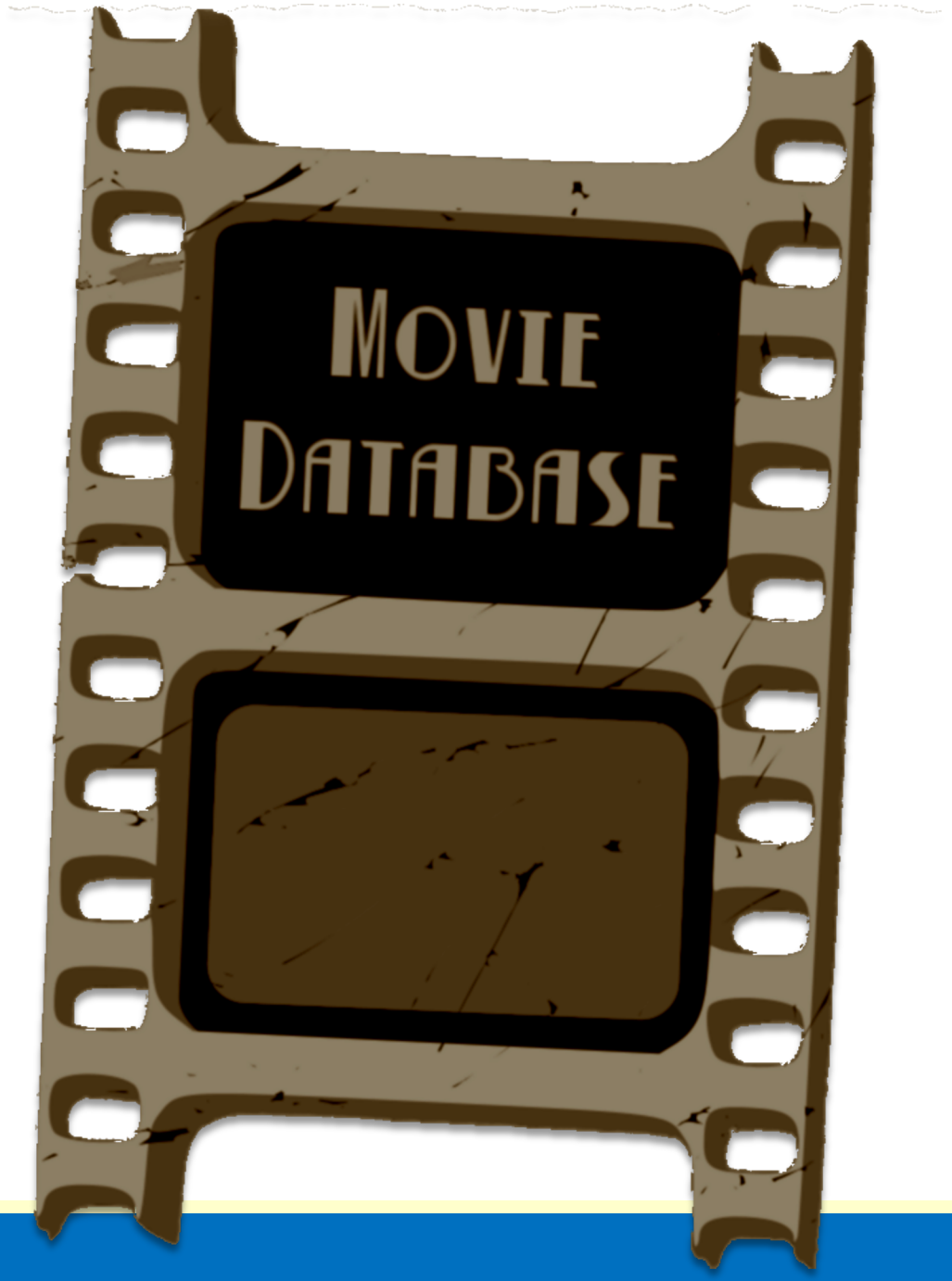
2. Key

Shiqi Yu 于仕琪

yusq@sustech.edu.cn

Most contents are from Stéphane Faroult's slides

We are going to illustrate the design process with a film database.



	A	B	C	D	E	F
1	**Movie Title**	**Country**	**Year**	**Director**	**Starring**	
2	Citizen Kane	US	1941	welles, o.	Orson Welles, joseph Cotten	
3	La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir	
4	North by Northwest	US	1959	HITCHCOCK, A.	Cary GRANT, Eva Marie SAINT, James MASON	
5	Singin' In the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor	
6	Rear Window	US	1954	HITCHCOCK, A.	James STEWART, Grace KELLY	
7	City Lights	US	1931	CHAPLIN, C.	Charlie CHAPLIN, Virginia CHERRILL	
8	The Third Man	GB	1949	Reed, C.	Joseph Cotten, Alida Valli, Orson Welles	
9	The Searchers	US	1956	Ford, J.	John Wayne, Jeffrey Hunter, Natalie Wood	
10	Ladri di biciclette	IT	1949	DeSica, V.	Lamberto Maggiorani, Enzo Staiola	
11	Annie Hall	US	1977	Allen, W.	Woody Allen, Diane Keaton	
12	On the Waterfront	US	1954	Kazan, E.	Marlon Brando, Eva Marie Saint, Karl Malden	
13	All about Eve	US	1950	Mankiewicz, J.	Bette Davis, Anne Baxter, George Sanders	
14	Casablanca	US	1942	Curtiz, M.	Humphrey Bogart,Ingrid Bergman, Claude Rains	
15	The Treasure of the Sierra Madre	US	1948	HUSTON, J.	Humphrey BOGART, Walter HUSTON, Tim HOLT	
16	High Noon	US	1952	Zinnemann, F.	Gary Cooper, Grace Kelly	
17	Some Like It Hot	US	1959	Wilder, B.	Tony Curtis, Jack Lemmon, Marilyn Monroe	
18						

It's easy to find on the web lists such as "The 100 greatest films ever", sometimes as a .csv file that you can load into a spreadsheet.

Movie Title	Country	Year	Director	Starring
Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
Rear Window	US	1954	HITCHCOCK, A.	James Stewart, Grace Kelly

First of all, order of rows and columns doesn't matter. What matters is that all the data in one row relates to the same film, and that data in a column matches the corresponding header.

Movie Title	Country	Year	Director	Starring
Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
Rear Window	US	1954	HITCHCOCK, A.	James Stewart, Grace Kelly
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason

You don't want, either, to have twice the same row in a table. It would tell you nothing new, and there would be the risk of getting wrong results when counting films by country for example. Duplicates are forbidden in relational tables.

Key

But if you have no duplicates, you need to identify what allows you to differentiate one row from another. It may be one column, or one set of columns, known collectively as a "key".

Flickr: Micky

Surname	*Birthdate*	*Picture*
Hepburn	**4-May-1929**	
Katharine		

Movie Title	Director	Country	Year	Starring
Singin' in the Rain	Donen/Kelly	US	1952	Gene Kelly, Debbie Reynolds, Donald O'Connor
Citizen Kane	welles, o.	US	1941	Orson Welles, Joseph Cotten
Rear Window	HITCHCOCK, A.	US	1954	James Stewart, Grace Kelly
La règle du jeu	Renoir, J.	FR	1939	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	HITCHCOCK, A.	US	1959	Cary Grant, Eva Marie Saint, James Mason

What is the "key" with films? The obvious answer, which looks correct, is to say "the title".

Except that you have remakes: same title, different films.

Movie Title	Director	Country	Year	Starring
Singin' in the Rain	Donen/Kelly	US	1952	Gene Kelly, Debbie Reynolds, Donald O'Connor
Citizen Kane	welles, o.	US	1941	Orson Welles, Joseph Cotten
Rear Window	HITCHCOCK, A.	US	1954	James Stewart, Grace Kelly
La règle du jeu	Renoir, J.	FR	1939	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	HITCHCOCK, A.	US	1959	Cary Grant, Eva Marie Saint, James Mason

Then you might say: title and director. A director wouldn't turn a remake a one of his films, would he?

Sorry to disappoint you, Hitchcock did it …

Movie Title	Director	Country	Year	Starring
Singin' in the Rain	Donen/Kelly	US	1952	Gene Kelly, Debbie Reynolds, Donald O'Connor
Citizen Kane	welles, o.	US	1941	Orson Welles, Joseph Cotten
Rear Window	HITCHCOCK, A.	US	1954	James Stewart, Grace Kelly
La règle du jeu	Renoir, J.	FR	1939	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	HITCHCOCK, A.	US	1959	Cary Grant, Eva Marie Saint, James Mason

the combination title/director/year will be unique

There is another problem, which is that there may be several directors.

• Directed by GENE KELLY and STANLEY DONEN •

Tough in a program to recognize that the film by Kelly, the one by Donen, the one by Kelly and Donen and the one by Donen and Kelly are the same one.

Movie Title	Director	Country	Year	Starring
Singin' in the Rain	Donen/Kelly	US	1952	Gene Kelly, Debbie Reynolds, Donald O'Connor
Citizen Kane	welles, o.	US	1941	Orson Welles, Joseph Cotten
Rear Window	HITCHCOCK, A.	US	1954	James Stewart, Grace Kelly
La règle du jeu	Renoir, J.	FR	1939	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	HITCHCOCK, A.	US	1959	Cary Grant, Eva Marie Saint, James Mason

For purely commercial reasons, though, it may be assumed that the combination title/country/year will be unique. You would confuse people by simultaneously releasing two films with the same title.

It may happen that several different keys are available (for instance StudentId, email address, ID number ... you share none of them). One of them is (arbitrarily) singled out and called the primary key. You usually choose the simplest one.

Primary Key

Flickr: Micky

3. Normalization

Shiqi Yu 于仕琪

yusq@sustech.edu.cn

Most contents are from Stéphane Faroult's slides

Normalization

Movie Title	Director	Country	Year	Starring
North By Northwest	HITCHCOCK, A.	US	1959	Cary Grant, Eva Marie Saint, James Mason
Rear Window	HITCHCOCK, A.	US	1954	James Stewart, Grace Kelly
Strangers on a Train	Alfred Hitchcock	US	1951	Farley Granger, Robert Walker

One common problem with databases is that data may be entered at different time by different people. If you let them enter data the way they want, it will make data retrieval very difficult because when they search, computers merely compare bytes. You need to standardize your data, a process also known as **normalization**.

Director _Firstname	Director_Surname
Alfred	Hitchcock
Orson	Welles

The first rule of normalization (also known as "First Normal Form" or 1NF for short) is that **each column should only contain ONE piece of information**.

Thus, surname and first name should be stored in different columns. Additionally (this can be done by programs storing data) some kind of "standard capitalization" should be used. The standard itself doesn't really matter (you could choose to have surnames in uppercase), what counts is respecting one standard.

How about Chinese names?

Do you prefer to separate the surname and the given name?

于仕琪

or

于, 仕琪

欧阳风

or

欧阳, 风

or

欧, 阳风

Directors

Id	Firstname	Surname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985

We could also store additional information about each director.

In the world of films we can consider that first name and surname (which may not be the real ones) uniquely identify someone; they are “brands” more than names.

We can also add a numerical id, and make it the primary key.

Movies

Movie Title	Directorid	Country	Year	Starring
North By Northwest	1	US	1959	Cary Grant, Eva Marie Saint, James Mason
Rear Window	1	US	1954	James Stewart, Grace Kelly
Strangers on a Train	1	US	1951	Farley Granger, Robert Walker
Citizen Kane	2	US	1941	Orson Welles, Joseph Cotten
The Magnificent Ambersons	2	US	1942	Joseph Cotten, Dolores Costello, Tim Holt

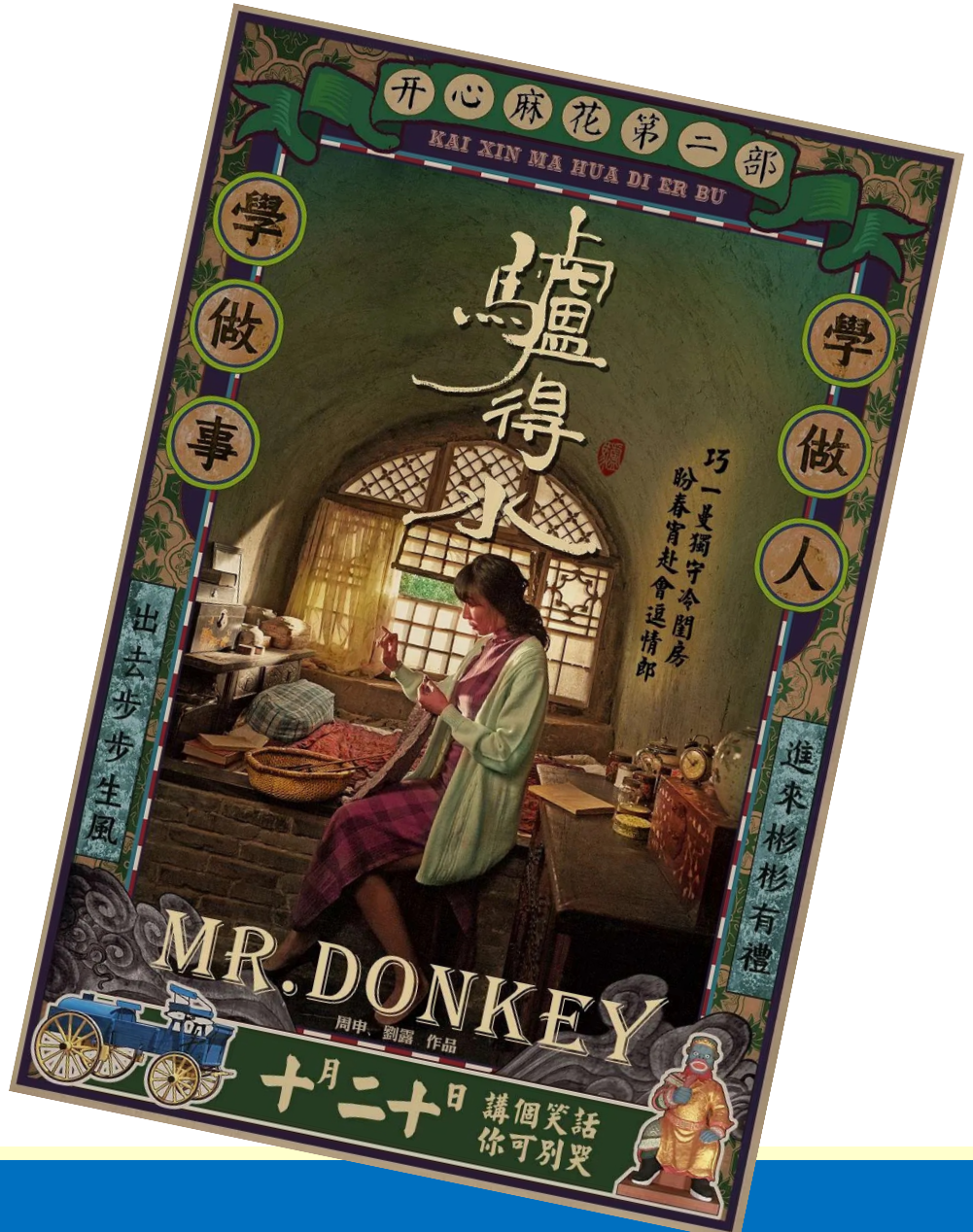
The director name in table Movies is replaced with the corresponding identifier.

Directors

Id	Firstname	Surname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985

Except that "Singin' In The Rain" still has two directors.

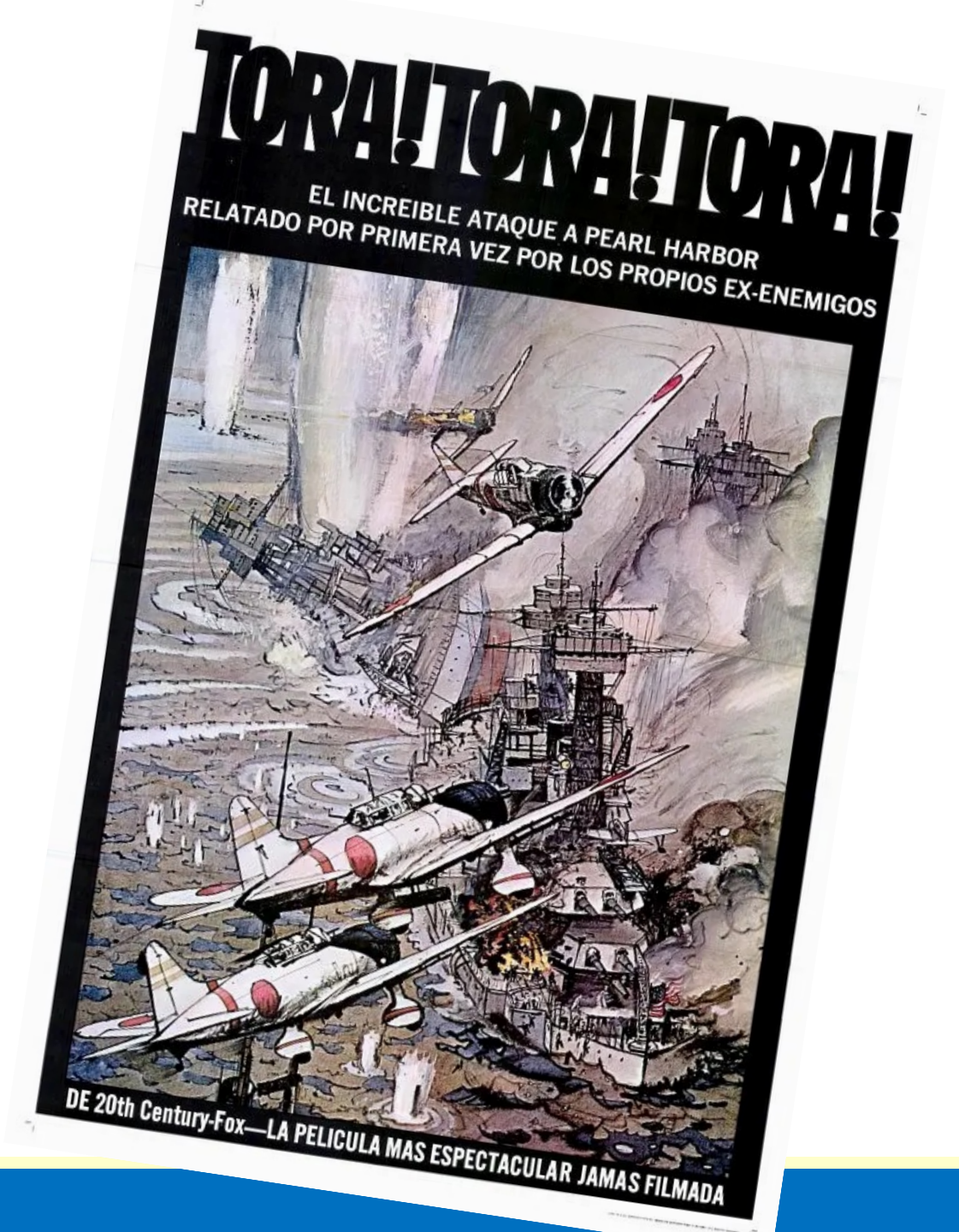
It's not an isolated case.



Movie Title	Directorid1	Directorid2	Country	...

Should we add one column per possible director? Bad, bad idea. First of all, we would have to decide on the maximum number of directors a film may have, and if we have it wrong we may have problems one day.

You may have 3.



An Italian film even is with 5 directors!

...	Year	Actorid1	Actorid2	Actorid3

The problem is the same, only worse, with actors.

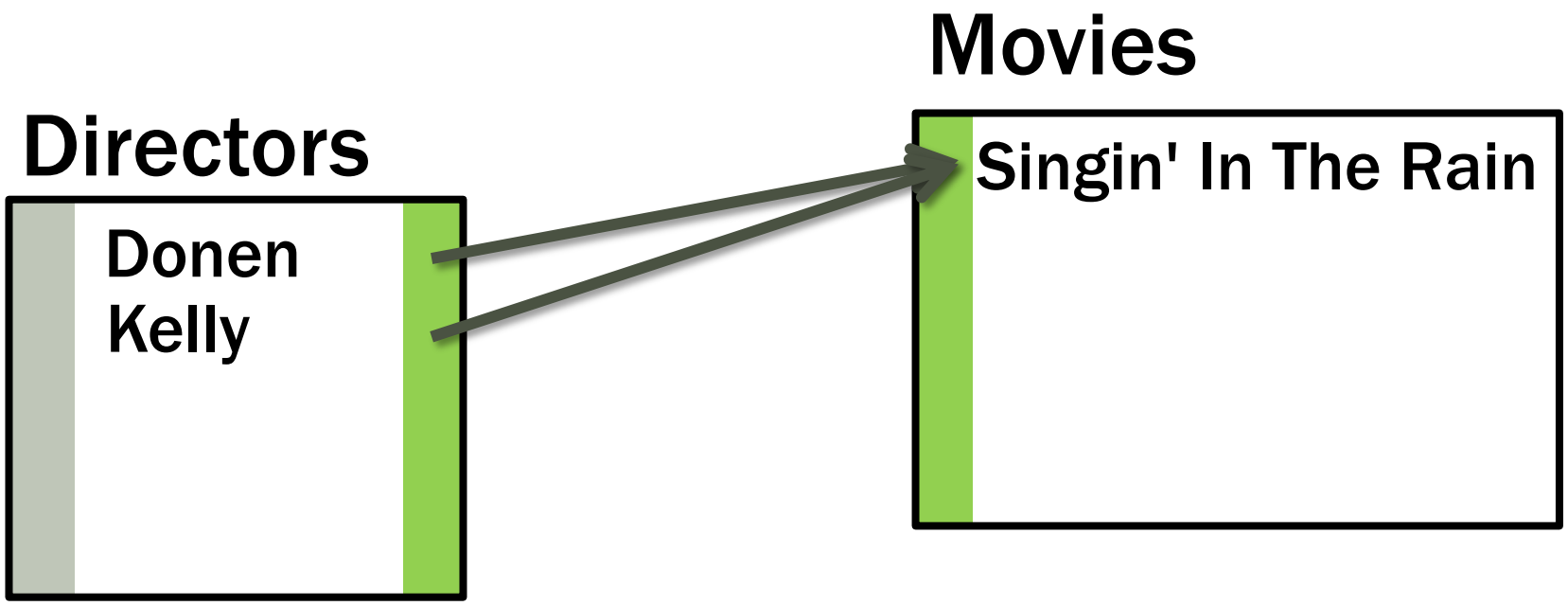
We might say "we only keep the three more important ones", but perhaps one day you'd be curious to see the first film where somebody now really famous had a small role, which would require storing the full cast.

Or you may simply have a lot of trouble deciding who are the three most important actors.

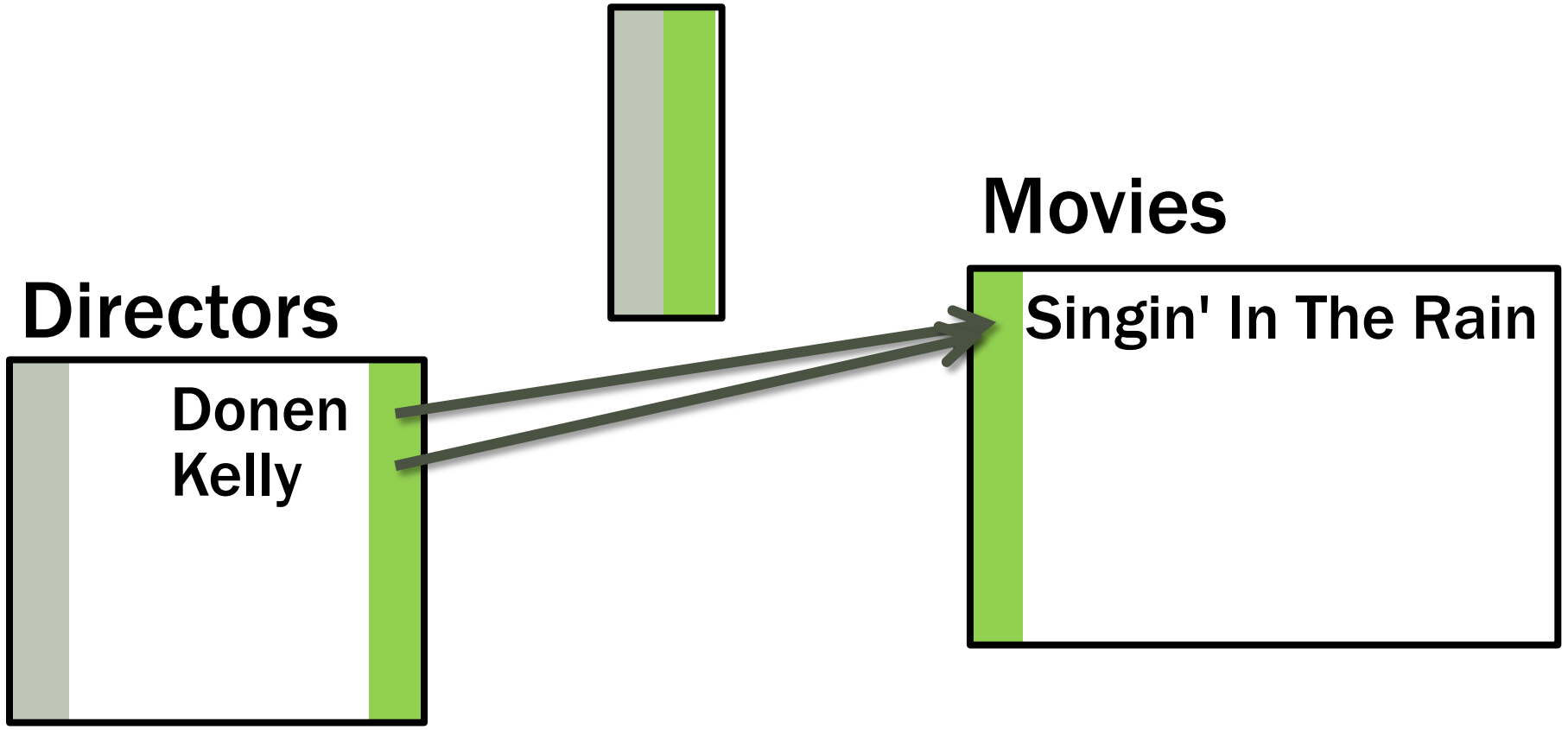
Movies

MovieId	Movie Title	Country	Year
1	North By Northwest	US	1959
2	Rear Window	US	1954
3	Strangers on a Train	US	1951
4	Citizen Kane	US	1941
5	The Magnificent Ambersons	US	1942
6	Singin' in the Rain	US	1952

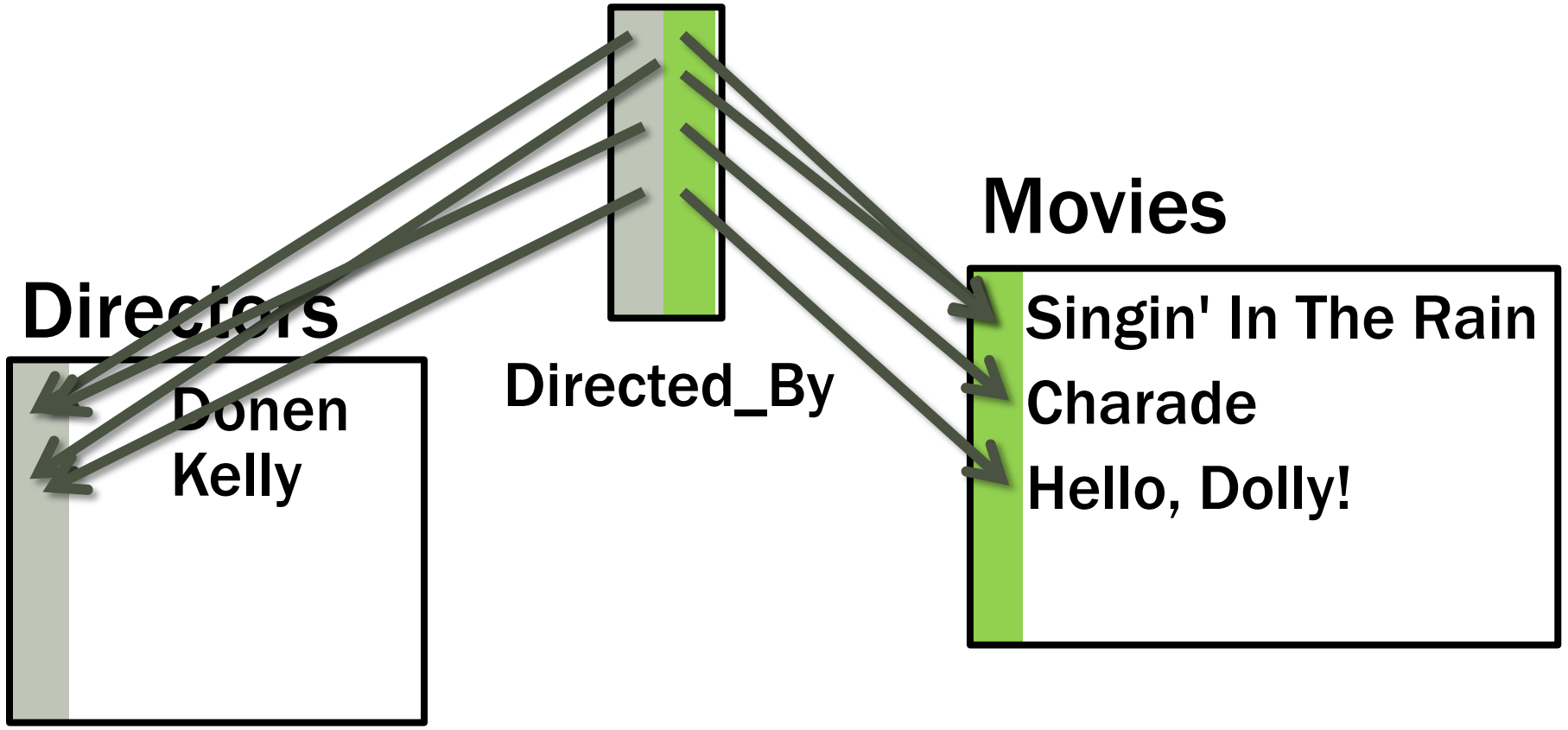
So we must devise something else. First of all let's remove director and actor information from the table of Movies.



We could think of associating a film with a director, instead of the reverse, but it's just shifting the problem. Some directors have directed over 100 films ...



The proper way to solve such a problem is to create an additional table that associates ONE director with ONE film, with as many rows per director as he has directed films, and as many rows per film as there are directors.



The proper way to solve such a problem is to create an additional table that associates ONE director with ONE film, with as many rows per director as he has directed films, and as many rows per film as there are directors.

Movies

Movie Title	Directorid	Country	Year	Actorid1	Actorid2	Actorid3
North By Northwest	1	US	1959	11	14	16
Rear Window	1	US	1954	17	15	
Strangers on a Train	1	US	1951	9		
Citizen Kane	2	US	1941	19	18	
The Magnificent Ambersons	2	US	1942	18	12	10

You can do the same with actors.

Directed_By

MovieId	DirectorId
1	1
2	1
3	1
4	2
5	2

Played_In

MovieId	ActorId
7	1
8	1

Movies

MovieId	Movie Title	Country	Year
1	North By Northwest	US	1959
2	Rear Window	US	1954
3	Strangers on a Train	US	1951
4	Citizen Kane	US	1941
5	The Magnificent Ambersons	US	1942

Directors

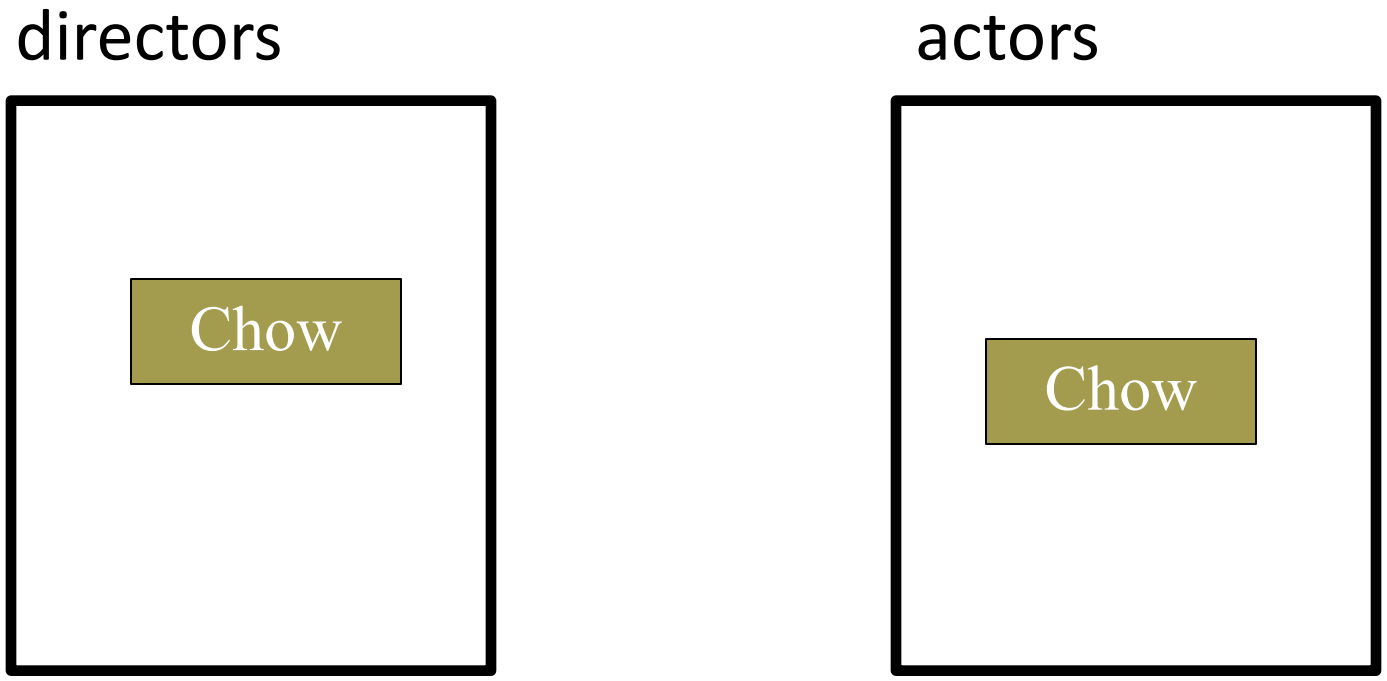
Id	Firstname	Surname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985

Actors

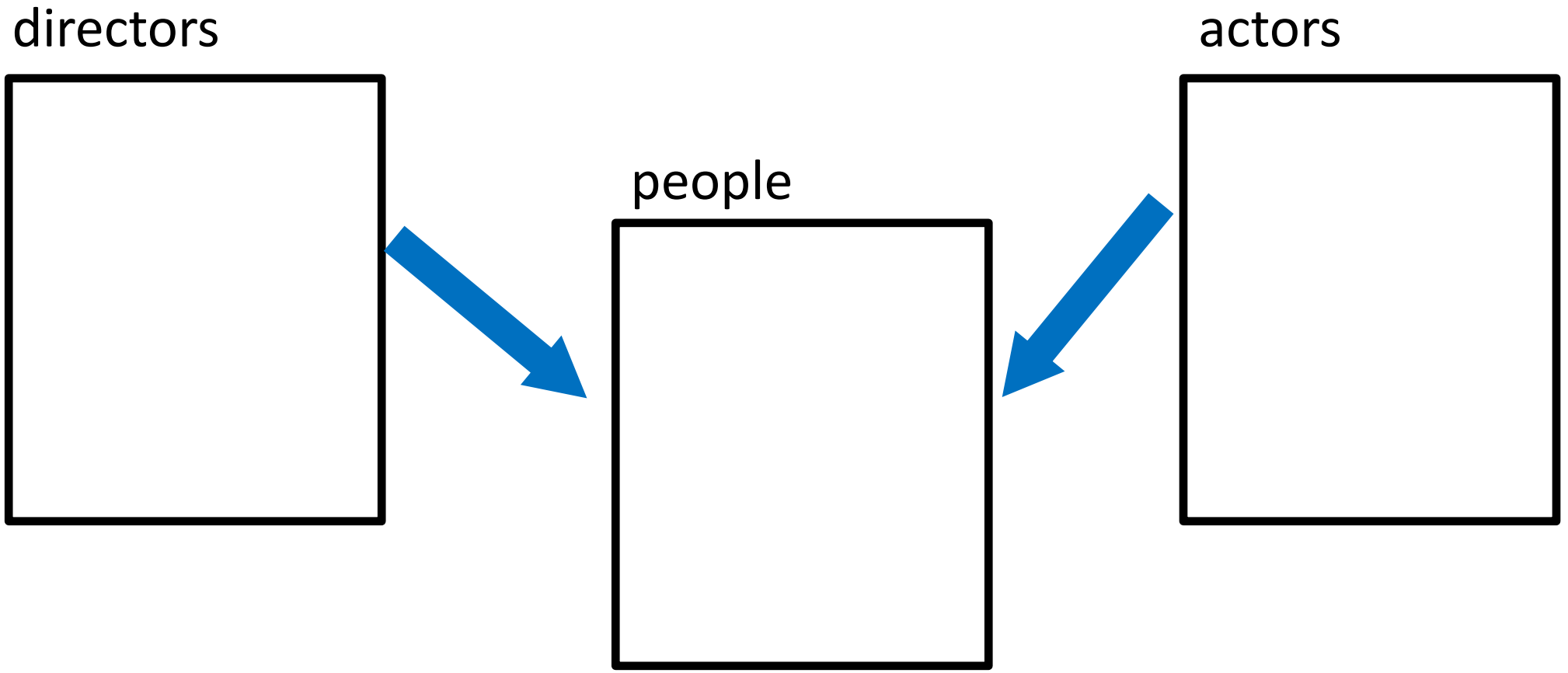
Id	Firstname	Surname	Born	Died
1	Audrey	Hepburn	1929	1993

The intermediate table stores the link.

Some directors have the unpleasant habit of starring in their films.



If we store the same information about the director Chow and the actor Chow, the day when we need to update the database. We risk missing one occurrence and having inconsistencies. We'll no longer know what is true, and lose information.



So the reasonable thing to do is to merge directors and actors into a single "people" table that stores the information only once for each unique individual.

Movies

MovieId	Movie Title	Country	Year
2	Rear Window	US	1954
4	Citizen Kane	US	1941

Credits

MovieId	PersonId	Credited
2	1	D
2	3	A
2	4	A
4	2	D
4	2	A
4	5	A

People

Id	Firstname	Surname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985
3	James	Stewart	1908	1997
4	Grace	Kelly	1929	1982
5	Joseph	Cotten	1905	1994

This requires an additional column to the table linking people and movies, to tell in which capacity a person intervened in a film. This allows to only have different rows and know than Welles directed and played in "Citizen Kane".

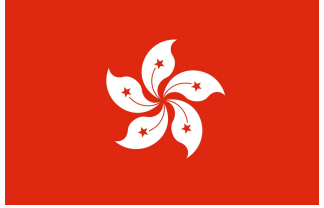
Unfortunately, information that you think was simple to manage sometimes proves to be elusive.

Let's dig deeper

Country?

"Country" looks simple enough. But since the invention of cinema (1895), the political map of the world has been redrawn several times.

Hong Kong films?



A lot of famous films produced in Hong Kong before 1997.

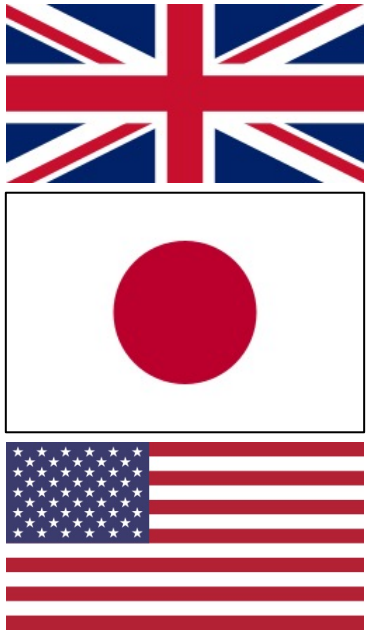
But HK has never been a country.

HK films are parallel with Mainland China films most of the time.

Soviet Union films?

SU isn't on the current world map.

Should we convert the SU films produced before 1991 to Russian?



And there are international co-productions. What is the country of the film then?

Associate a "main country" with each film? Money? Theme? Director?

Table associating movieid with one or several countries?

You often have several options and which one is the best one depends. You may opt for simplicity, but must be aware that you may lose some information.

Let's dig deeper

Release date?

A film can be released at different dates in different countries.

Add one more table for release dates?

Let's dig deeper

runtime?

March 3: Logan

- 137 minutes, USA
- 123 minutes, China

Everything is vague to a degree you do not realize till you have tried to make it precise.

------Bertrand Russell
(1892-1970)

	A	B	C	D	E	F
1	**Movie Title**	**Country**	**Year**	**Director**	**Starring**	
2	Citizen Kane	US	1941	welles, o.	Orson Welles, joseph Cotten	
3	La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir	
4	North by Northwest	US	1959	HITCHCOCK, A.	Cary GRANT, Eva Marie SAINT, James MASON	
5	Singin' In the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor	
6	Rear Window	US	1954	HITCHCOCK, A.	James STEWART, Grace KELLY	
7	City Lights	US	1931	CHAPLIN, C.	Charlie CHAPLIN, Virginia CHERRILL	
8	The Third Man	GB	1949	Reed, C.	Joseph Cotten, Alida Valli, Orson Welles	
9	The Searchers	US	1956	Ford, J.	John Wayne, Jeffrey Hunter, Natalie Wood	
10	Ladri di biciclette	IT	1949	DeSica, V.	Lamberto Maggiorani, Enzo Staiola	
11	Annie Hall	US	1977	Allen, W.	Woody Allen, Diane Keaton	
12	On the Waterfront	US	1954	Kazan, E.	Marlon Brando, Eva Marie Saint, Karl Malden	
13	All about Eve	US	1950	Mankiewicz, J.	Bette Davis, Anne Baxter, George Sanders	
14	Casablanca	US	1942	Curtiz, M.	Humphrey Bogart,Ingrid Bergman, Claude Rains	
15	The Treasure of the Sierra Madre	US	1948	HUSTON, J.	Humphrey BOGART, Walter HUSTON, Tim HOLT	
16	High Noon	US	1952	Zinnemann, F.	Gary Cooper, Grace Kelly	
17	Some Like It Hot	US	1959	Wilder, B.	Tony Curtis, Jack Lemmon, Marilyn Monroe	
18						

So you have seen how we went from a raw spreadsheet...

Movies

MovieId	Movie Title	Country	Year
2	Rear Window	US	1954
4	Citizen Kane	US	1941

Credits

MovieId	PersonId	Credited
2	1	D
2	3	A
2	4	A
4	2	D
4	2	A
4	5	A

People

Id	Firstname	Surname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985
3	James	Stewart	1908	1997
4	Grace	Kelly	1929	1982
5	Joseph	Cotten	1905	1994

... to a relatively clean database model allowing us to handle weird cases. This process is known as **normalization**.

Normalization

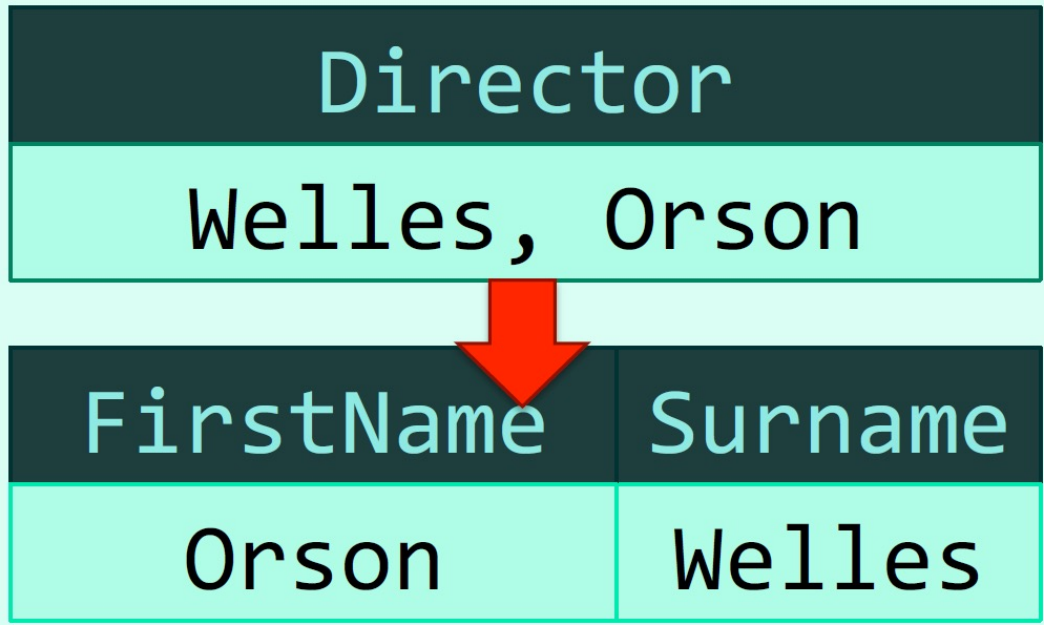
Three simple rules to design a database by Codd in 1971:

- First Normal Form (1NF)
- Second Normal Form (2NF)
- Third Normal Form (3NF)

Codd, E. F. "Further Normalization of the Data Base Relational Model". IBM Research Report RJ909 (August 31, 1971)

Normalization

Simple attributes



1

Normalization

attributes depend on the full key

Title	Year	Country	Continent
Good Bye, Lenin!	2003	Germany	Europe

2

Wrong. I don't want to repeat it for every German film. It should be said once in a table of countries that Germany is in Europe.

Normalization

non-key attributes not depend on each other

Country	Name	Continent	Currency	Symbol
de	Germany	Europe	Euro	€

3

Wrong. I don't want to repeat it for every country in the Euro zone. I should have it in a currency table, the key of which would be the currency.

Every non key attribute must provide a fact about the key, the whole key, and nothing but the key.

William Kent (1936 – 2005)

William Kent. "A Simple Guide to Five Normal Forms in Relational Database Theory", Communications of the ACM 26 (2), Feb. 1983, pp. 120–125.

Thank You!