C. Batini & M. Scannapieco
Data and Information Quality Book Figures

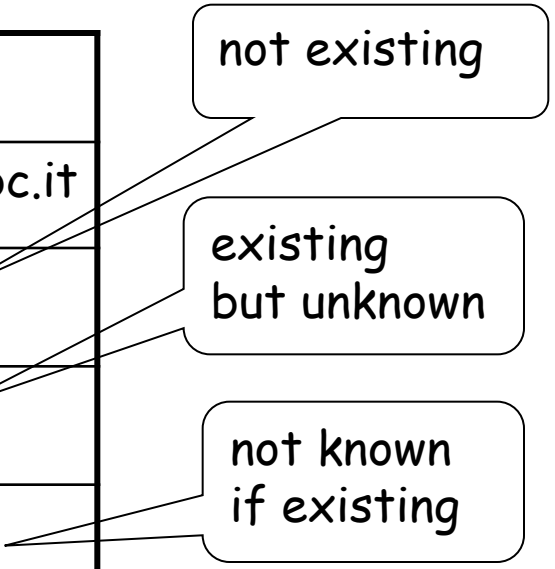
Chapter 2: Data Quality Dimensions

A relation Movies

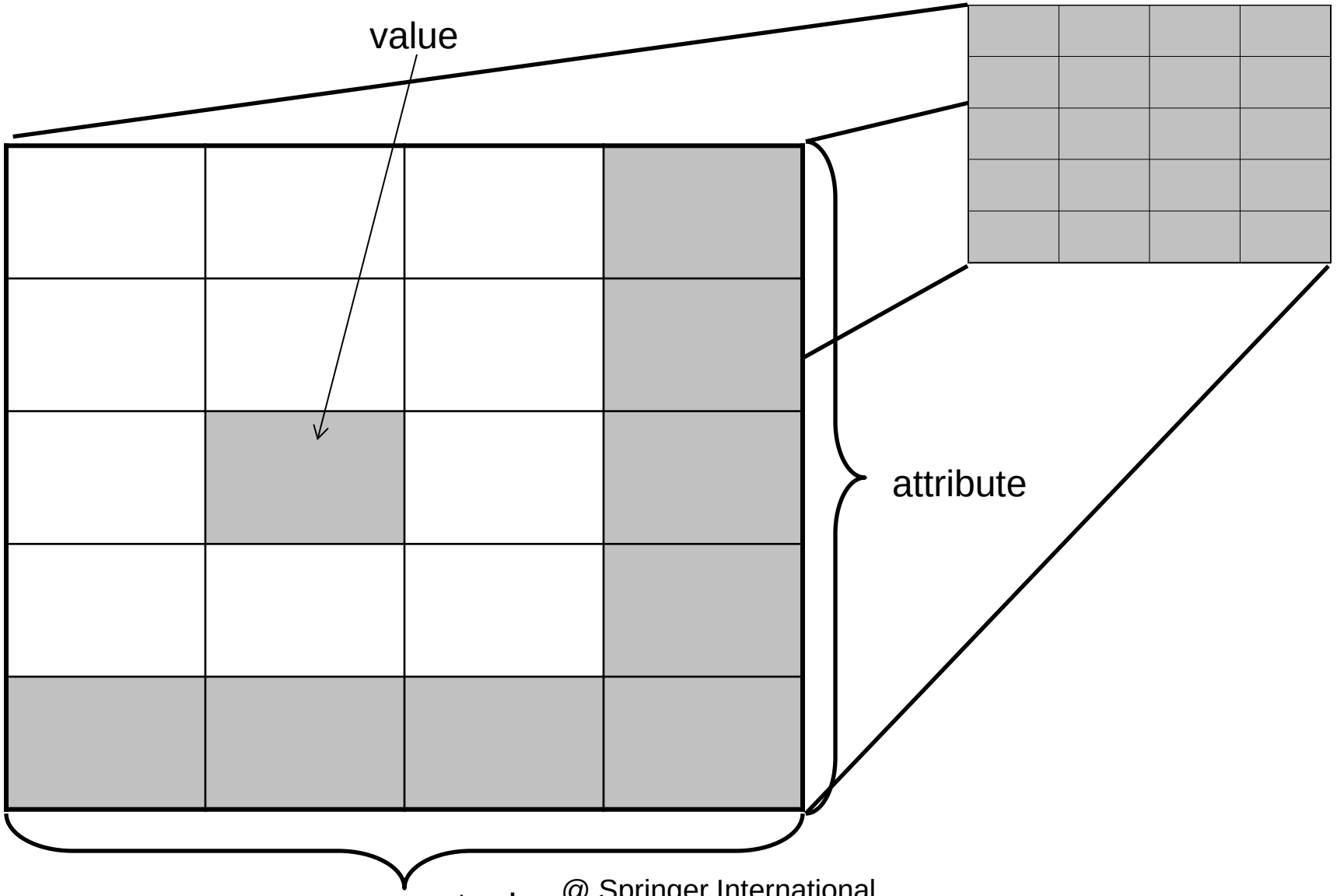
Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

The Person relation, with different null value meanings for the e-mail attribute

ID	Name	Surname	BirthDate	Email
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	NULL
3	Anthony	White	01/01/1936	NULL
4	Marianne	Collins	11/20/1955	NULL



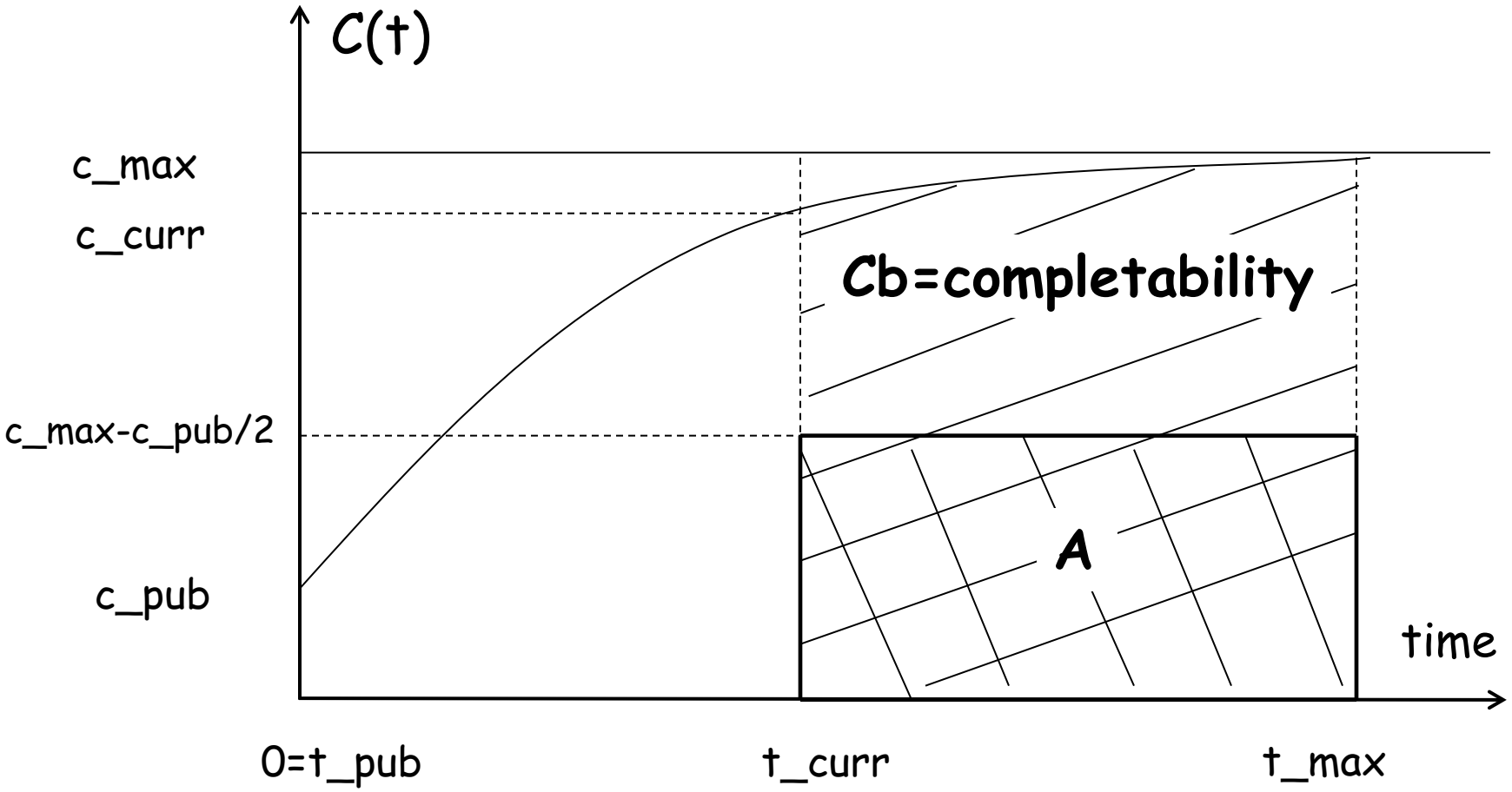
Completeness of different elements in the relational model



Student relation exemplifying the completeness of tuples, attributes and relations

StudentID	Name	Surname	Vote	ExaminationDate
6754	Mike	Collins	29	07/17/2004
8907	Anne	Herbert	18	07/17/2004
6578	Julianne	Merrals	NULL	07/17/2004
0987	Robert	Archer	NULL	NULL
1243	Mark	Taylor	26	09/30/2004
2134	Bridget	Abbott	30	09/30/2004
6784	John	Miller	30	NULL
0098	Carl	Adams	25	09/30/2004
1111	John	Smith	28	09/30/2004
2564	Edward	Monroe	NULL	NULL
8976	Anthony	White	21	NULL
8973	Marianne	Collins	30	10/15/2004

A graphical representation of completability



Example of functional dependencies

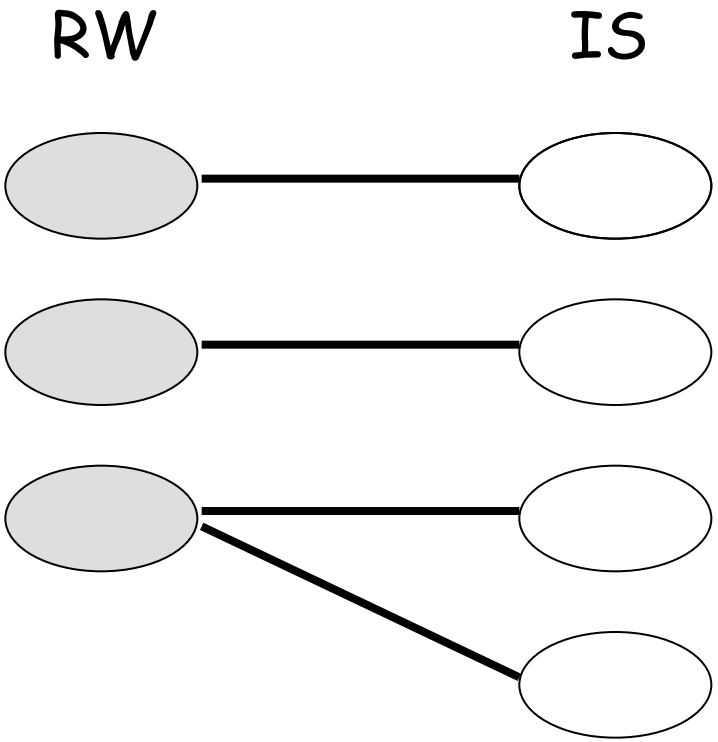
A	B	C	D
a_1	b_1	c_1	d_1
a_1	b_1	c_1	d_2
a_1	b_2	c_3	d_3

r_1

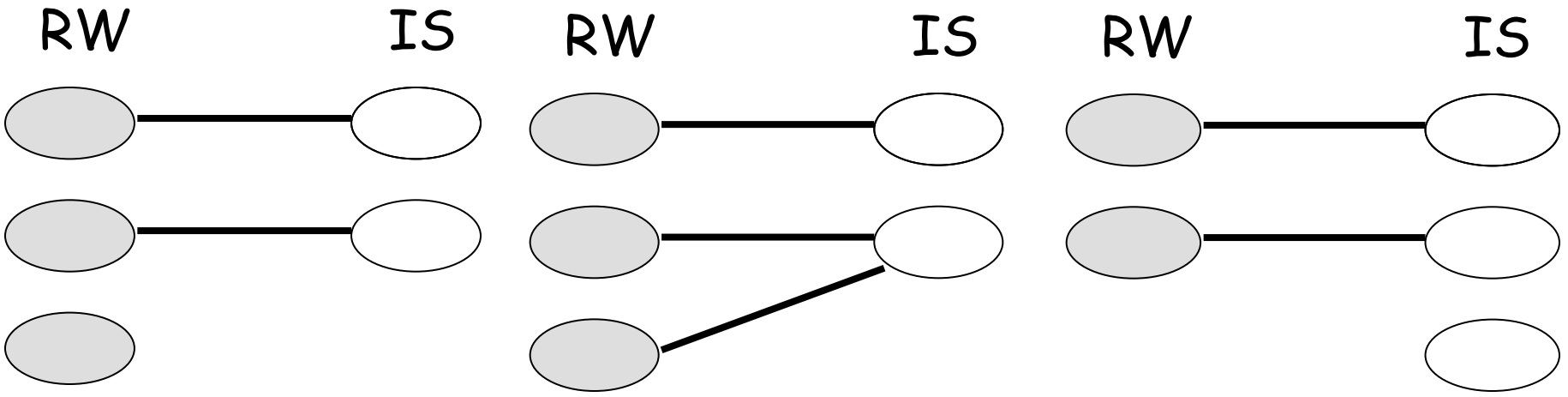
A	B	C	D
a_1	b_1	c_2	d_1
a_1	b_1	c_1	d_2
a_1	b_2	c_3	d_3

r_2

Proper representation of the real world system in the theoretical approach

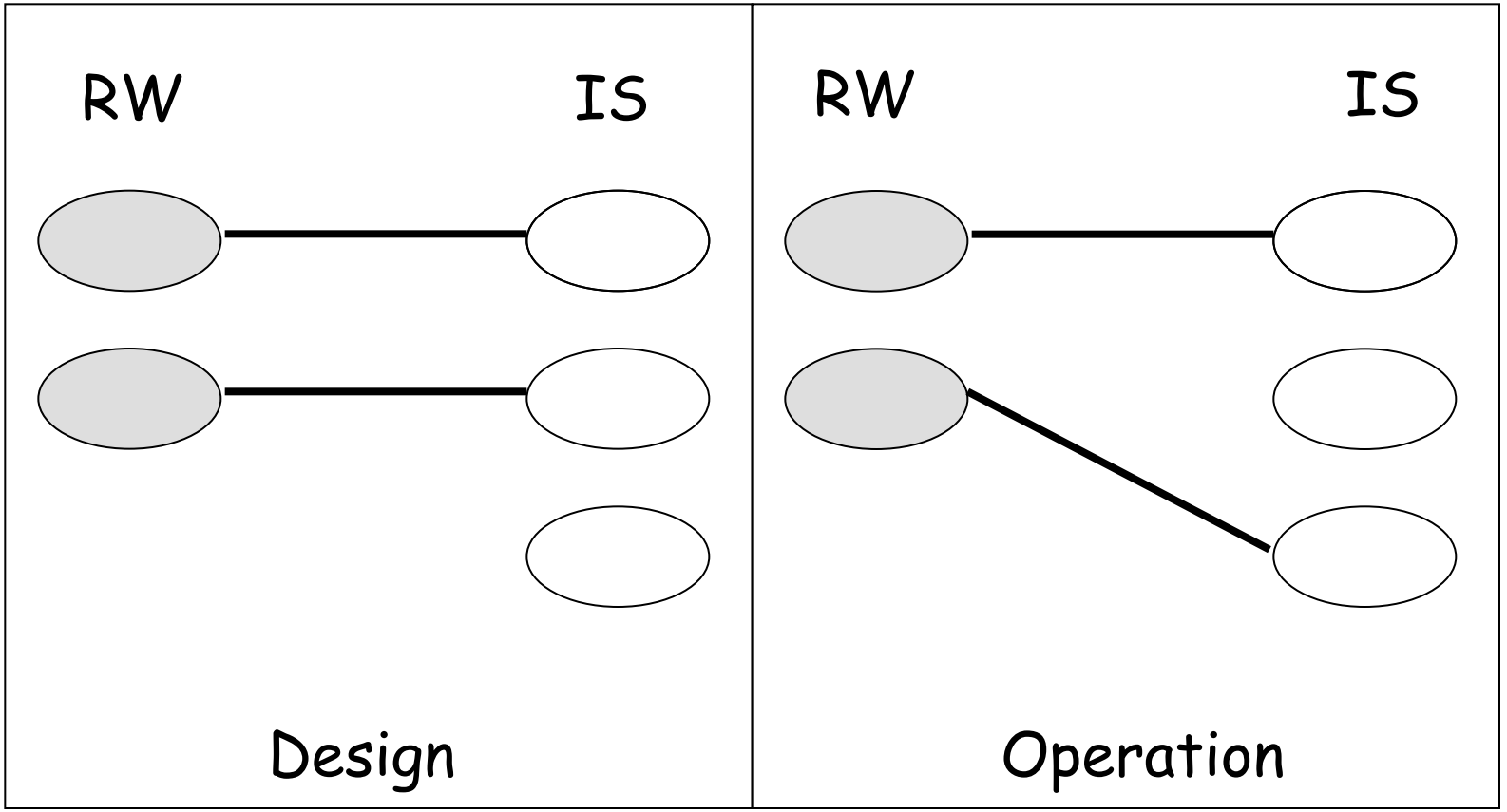


Incomplete, ambiguous, and meaningless representations of the real world system in the theoretical approach

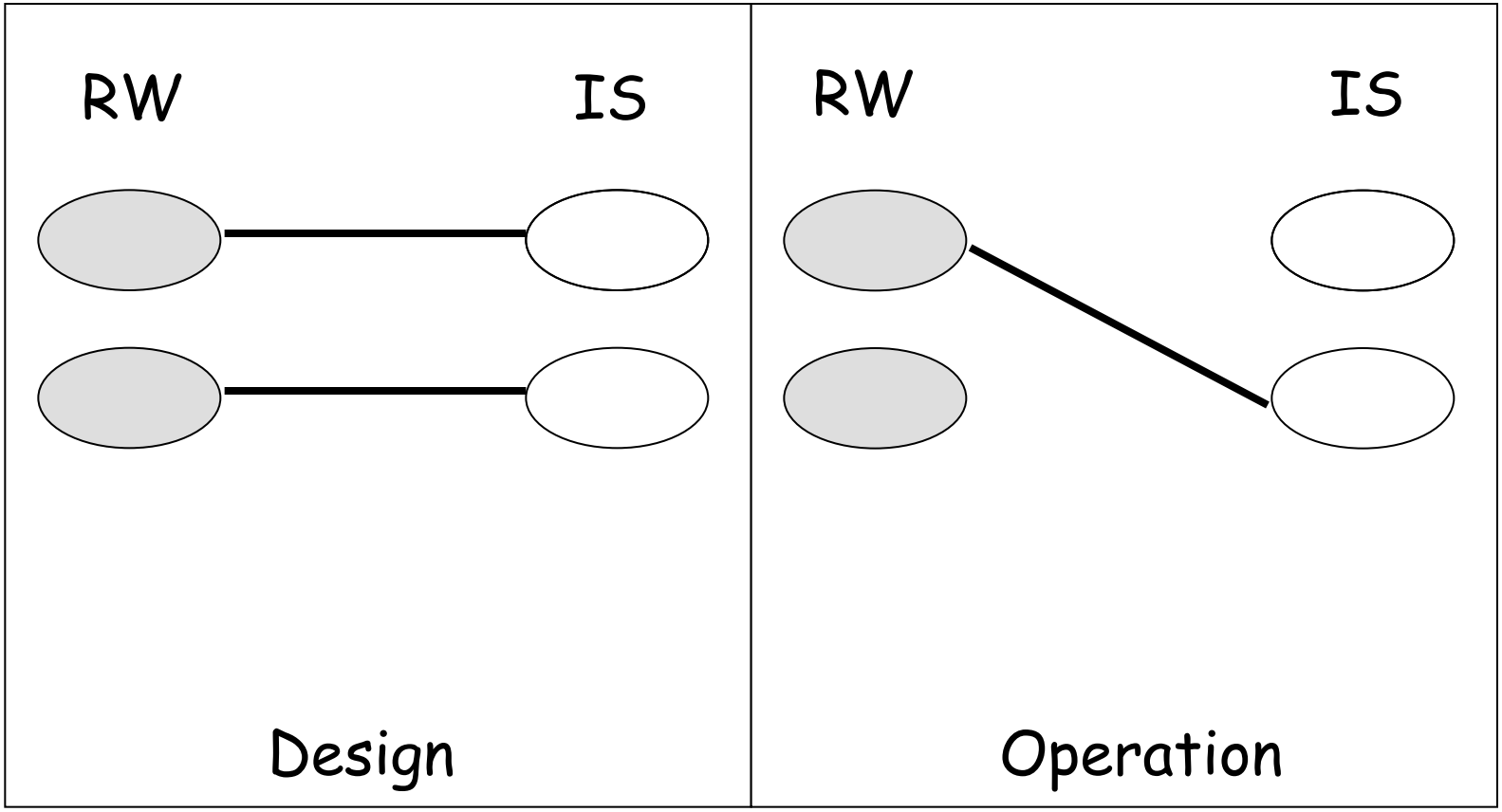


(a) Incomplete (b) Ambiguous (c) Meaningless

Garbling representations of the real world system from [640] – not meaningful



Garbling representations of the real world system from [640] - meaningful



Dimensions proposed in the empirical approach

Category	Dimension	Definition: the extent to which ...
Intrinsic	Beleivability	data are accepted or regardedas true, real and credible
	Accuracy	data are correct, reliable and certified free of error
	Objectivity	data are unbiased and impartial
	Reputation	data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	data are beneficial and provide advantages for their use
	Relevancy	data are applicable and useful for the task at hand
	Timeliness	the age of the data is appropriate for the task at hand
	Completeness	data are of sufficient depth, breadth, and scope for the task at hand
	Appropriate amount of data	the quantity or volume of available data is appropriate
Representational	Intepretability	data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	data are clear without ambiguity and easily comprehended
	Representational consistency	data are always presented in the same format and are compatible with the previous data
	Concise representation	data are compactly represented without behing overwhelmed
Accessibility	Accessibility	data are available or easily and quickly retrieved
	Access security	access to data can be restricted and hence kept secure

Dimensions proposed in the intuitive approach [520]

Dimension Name	Type of dimension	Definition
Accuracy	data value	Distance between v and v', considered as correct
Completeness	data value	Degree to which values are present in a data collection
Currency	data value	Degree to which a datum is up-to-date
Consistency	data value	Coherence of the same datum, represented in multiple copies, or different data to respect integrity constraints and rules
Appropriateness	data format	One format is more appropriate than another if it is more suited to user needs
Interpretability	data format	Ability of the user to interpret correctly values from their format
Portability	data format	The format can be applied to as a wide set of situations as possible
Format precision	data format	Ability to distinguish between elements in the domain that must be distinguished by users
Format flexibility	data format	Changes in user needs and recording medium can be easily accommodated
Ability to represent null values	data format	Ability to distinguish neatly (without ambiguities) null and default values from applicable values of the domain
Efficient use of memory	data format	Efficiency in the physical representation. An icon is less efficient than a code
Representation consistency	data format	Coherence of physical instances of data with their formats

Time-related dimensions definitions

Reference	Definition
Wand 1996	Timeliness refers only to the delay between a change of a real world state and the resulting modification of the information system state
Wang 1996	Timeliness is the extent to which age of the data is appropriate for the task at hand
Redman 1996	Currency is the degree to which a datum is up-to-date. A datum value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value
Jarke 1999	Currency describes when the information was entered in the sources and/or the data warehouse. Volatility describes the time period for which information is valid in the real world
Bovee 2001	Timeliness has two componenents: age and volatility. Age or currency is a measure of how old the information is, based on how long ago it was recorded. Volatility is a measure of information instability-the frequency of change of the value for an entity attribute
Naumann 2002	Timeliness is the average age of the data in a source
Liu 2002	Timeliness is the extent to which data are sufficiently up-to-date for a task

Completeness dimensions definitions

Reference	Definition
Wand 1996	The ability of an information system to represent every meaningful state of the represented real world system.
Wang 1996	The extent to which data are of sufficient breadth, depth and scope for the task at hand
Redman 1996	The degree to which values are present in a data collection
Jarke 1999	Percentage of the real-world information entered in the sources and/or the data warehouse
Bovee 2001	Deals with information having all required parts of an entity's information present
Naumann 2002	It is the quotient of the number of non-null values in a source and the size of the universal relation
Liu 2002	All values that are supposed to be collected as per a collection theory

Two ways of modeling residence addresses

Person

ID	Name	Surname
1	John	Smith
2	Mark	Bauer
3	Ann	Swenson

Address

ID	StreetPrefix	StreetName	Number	City
A11	Avenue	Sunset	113	Chicago
A12	Street	4 Heroes	null	Denver

ResidenceAddress

PersonID	AddressID
1	A11
2	A11
3	A12

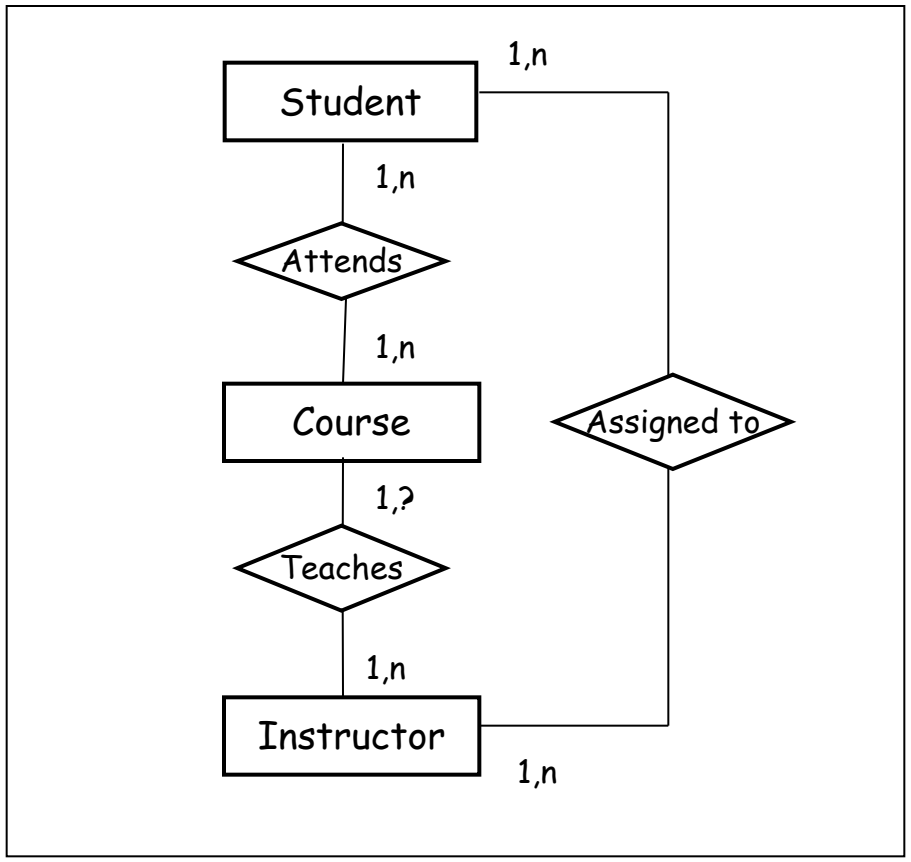
(b)

Person

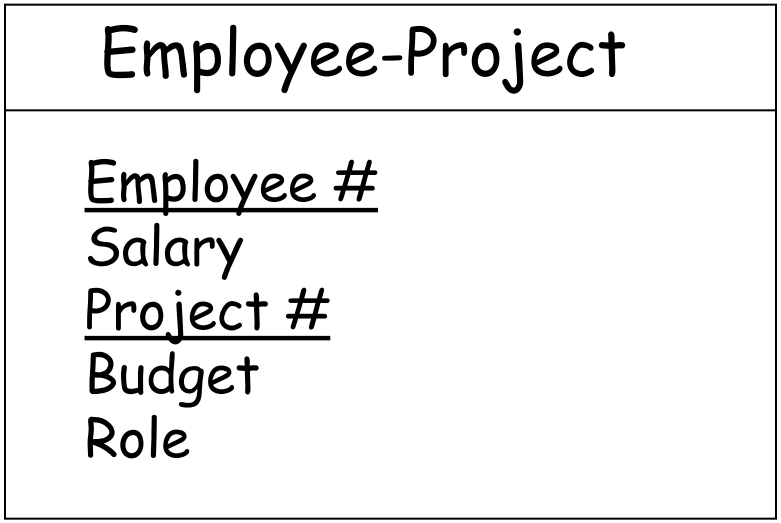
ID	Name	Surname	Address
1	John	Smith	113 Sunset Avenue 60601 Chicago
2	Mark	Bauer	113 Sunset Avenue 60601 Chicago
3	Ann	Swenson	4 Heroes Street Denver

(a)

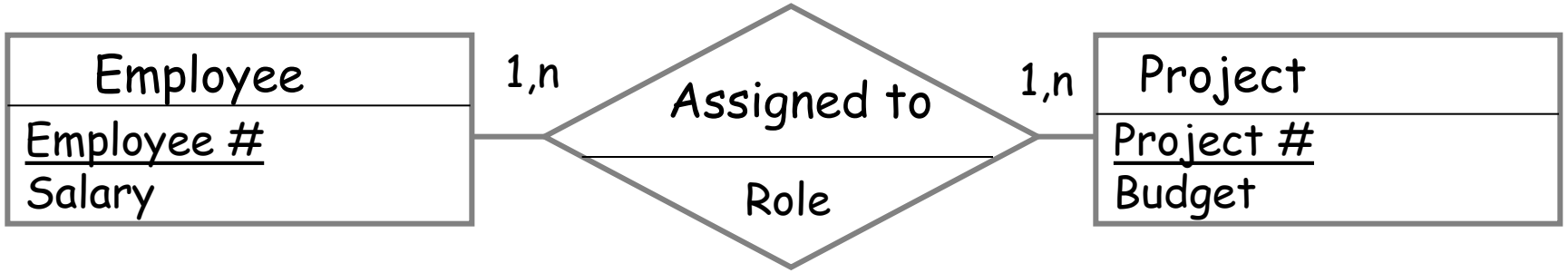
A possibly redundant schema



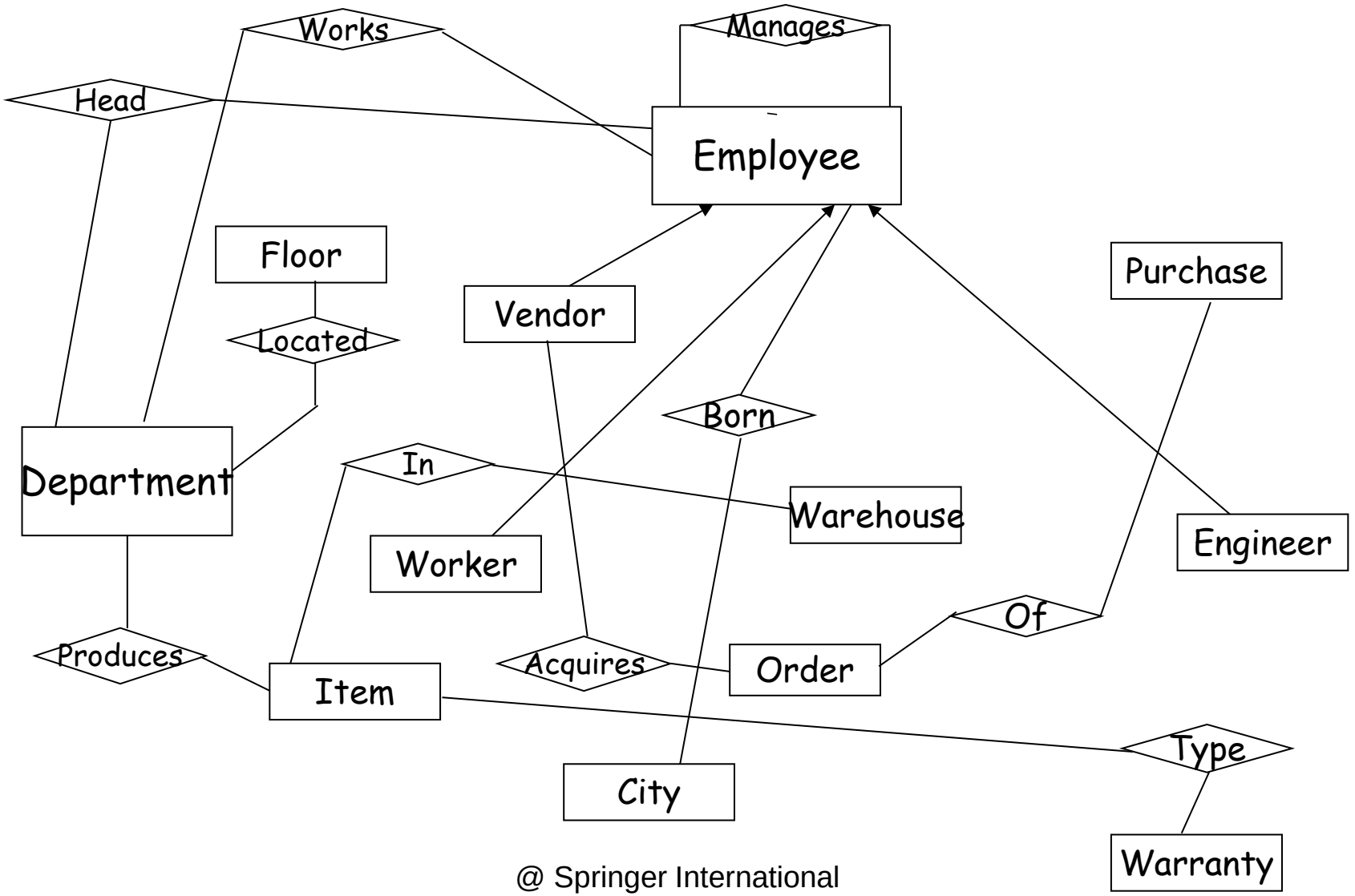
An unnormalized Entity Relationship schema



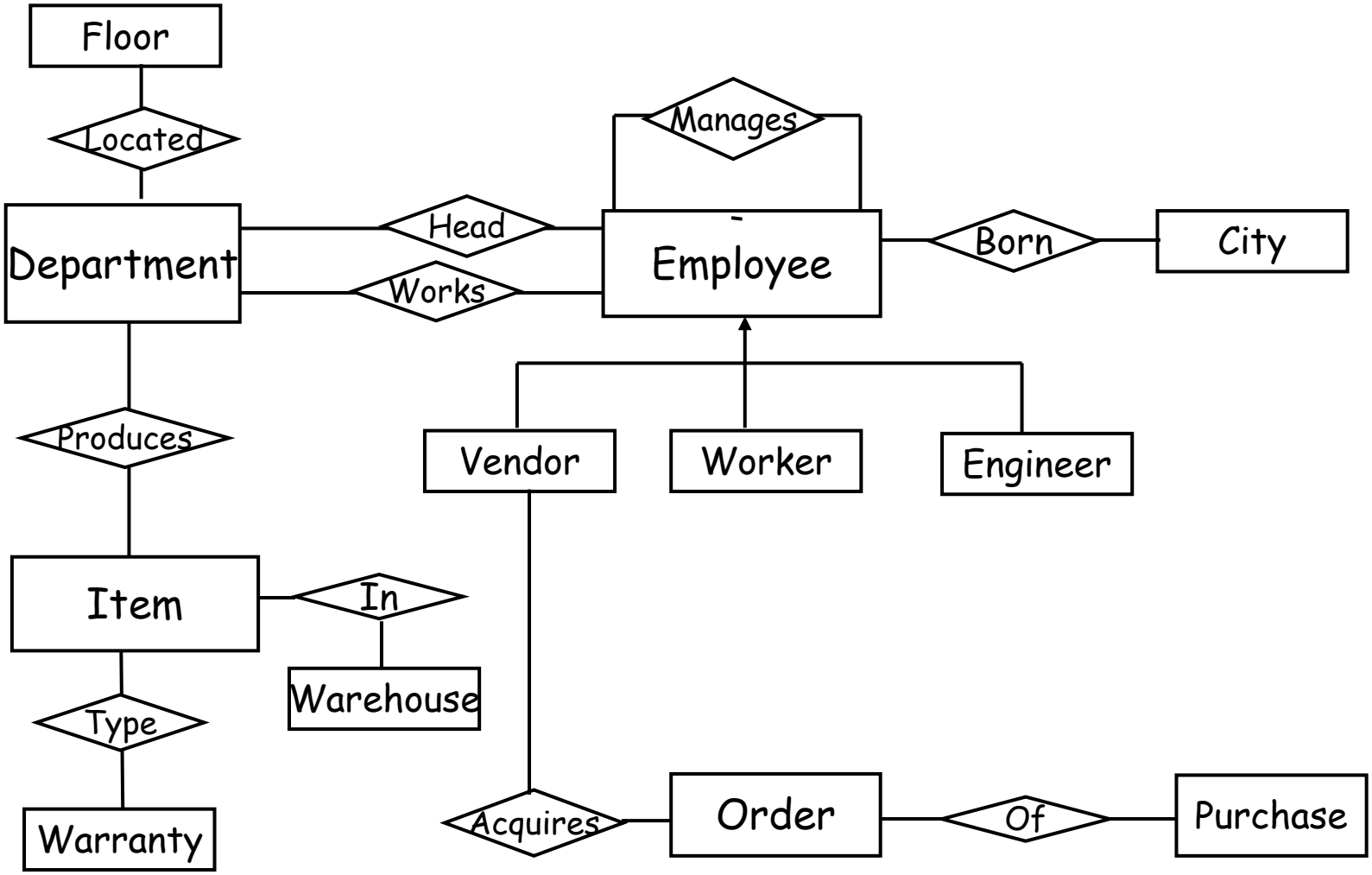
A normalized schema



"Spaghetti style" Entity Relationship diagram"



An equivalent readable schema



A schema transformation that improves compactness