

Databases

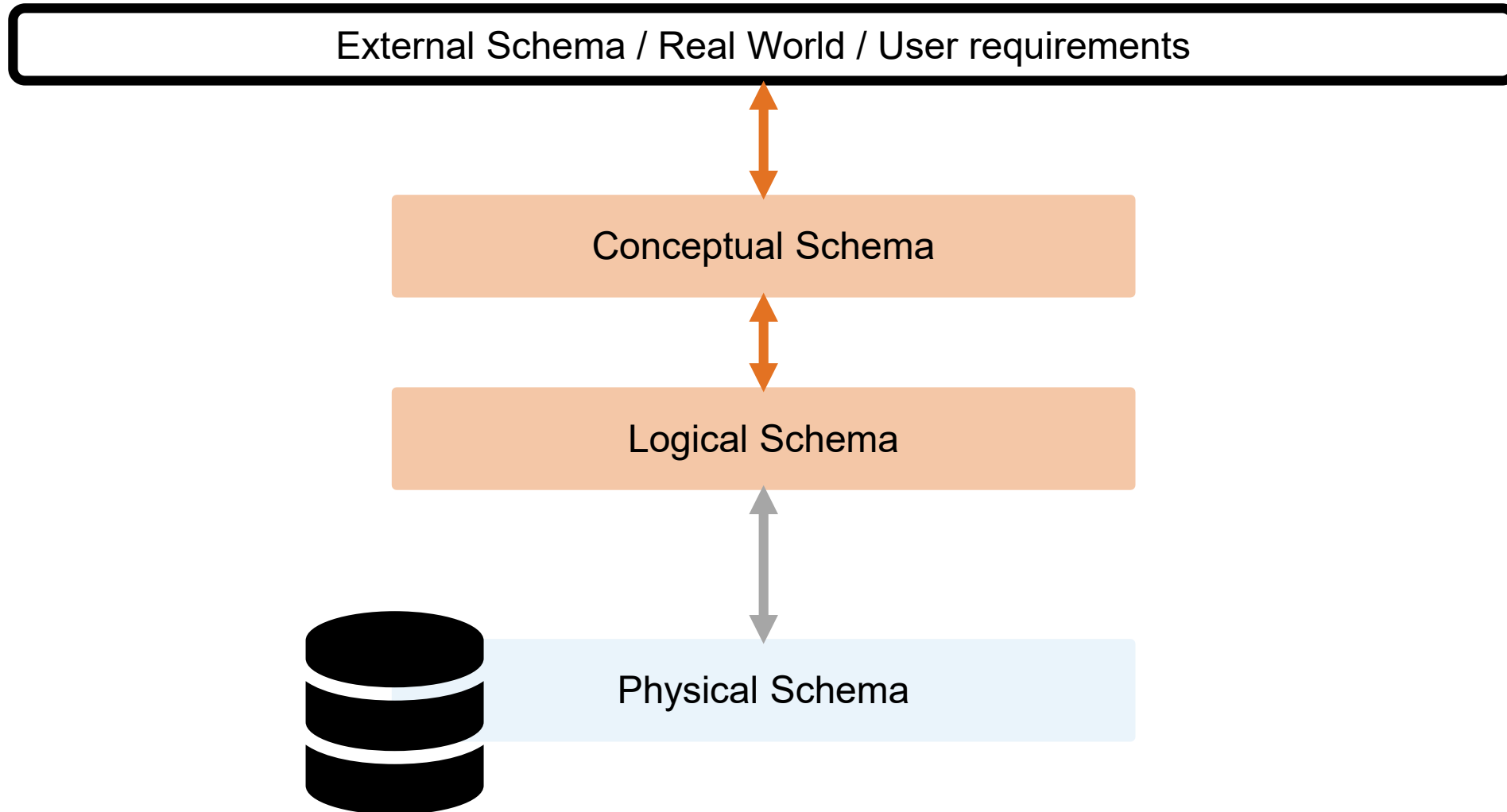
Mirco Schönfeld
University of Bayreuth

mirco.schoenfeld@uni-bayreuth.de
[@TWIlyY29](#)



https://commons.wikimedia.org/wiki/File:Microsoft_Bing_Maps%27_datacenter_-_Flickr_-_Robert_Scoble.jpg
https://commons.wikimedia.org/wiki/File:Virginia_Tech_-_data_center.jpg
[https://commons.wikimedia.org/wiki/File:123Net_Data_Center_\(DC2\).jpg](https://commons.wikimedia.org/wiki/File:123Net_Data_Center_(DC2).jpg)

Levels of Abstraction



Levels of Abstraction

External Schema / Real World / User requirements

First step of modeling:
Identify relevant entities,
attributes and relations

Conceptual Schema

Purposeful mapping of
real-world information
to a finite set of
modeled elements.

Logical Schema

Physical Schema



Levels of Abstraction

External Schema / Real World / User requirements

First step of modeling:
Identify relevant entities,
attributes and relations

Conceptual Schema

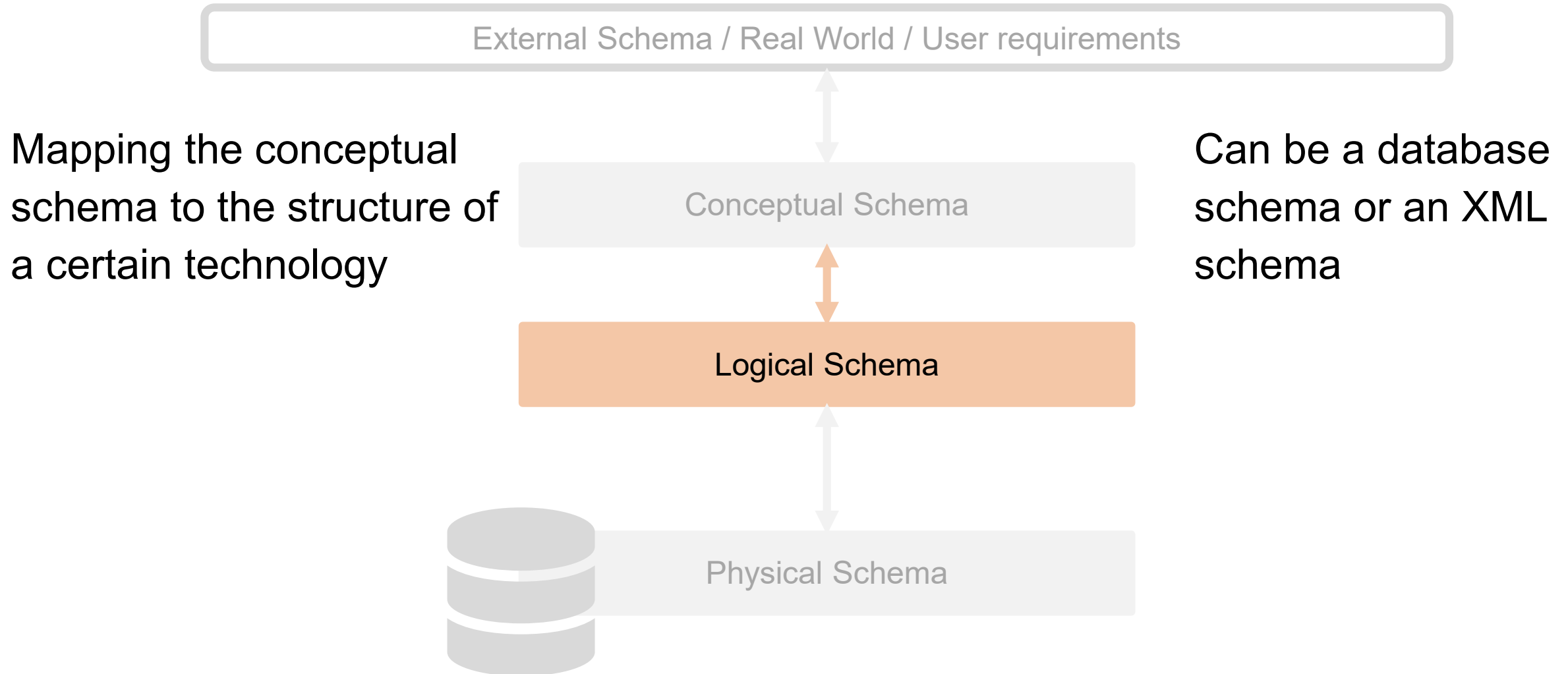
Purposeful mapping of
real-world information
to a finite set of
modeled elements.

This mapping is always a
selection and construction of
someone.

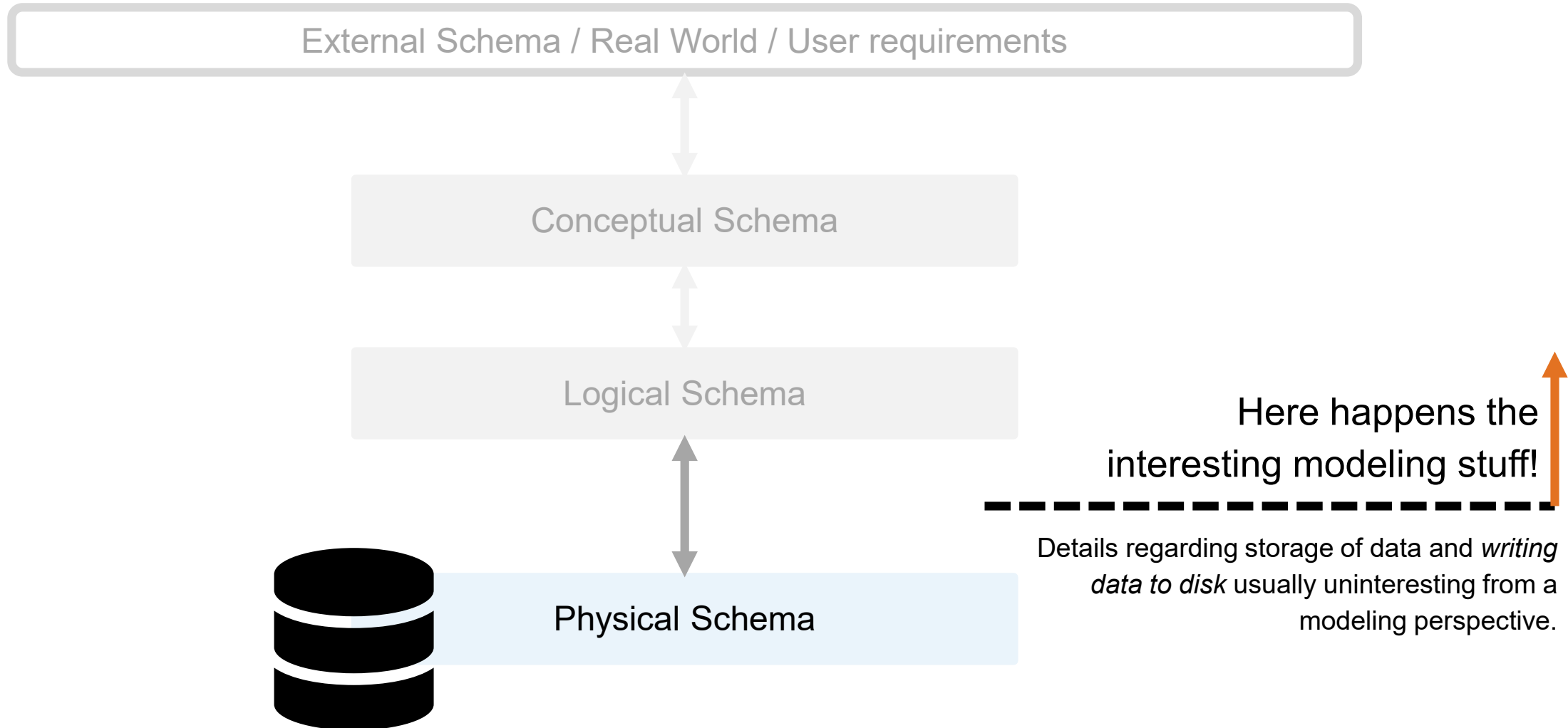
Physical Schema



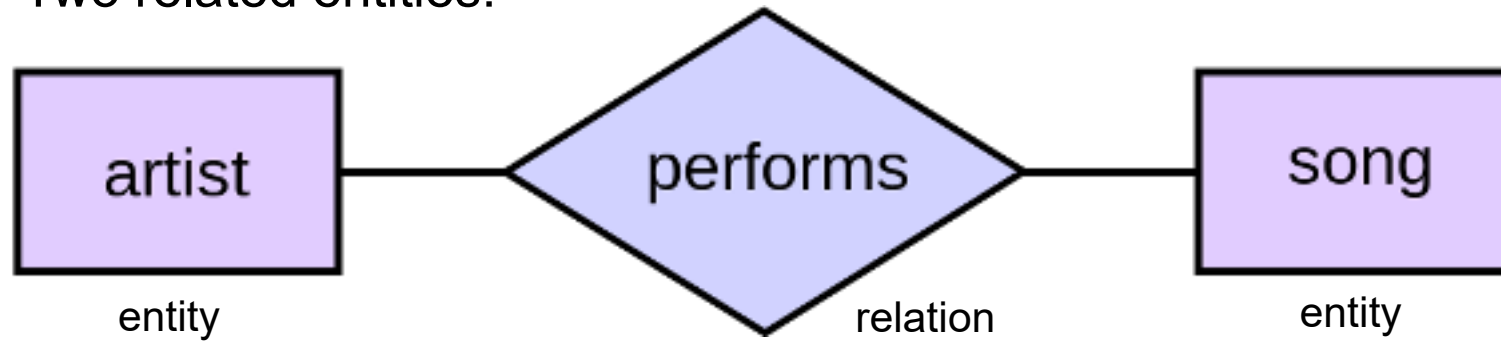
Levels of Abstraction



Levels of Abstraction



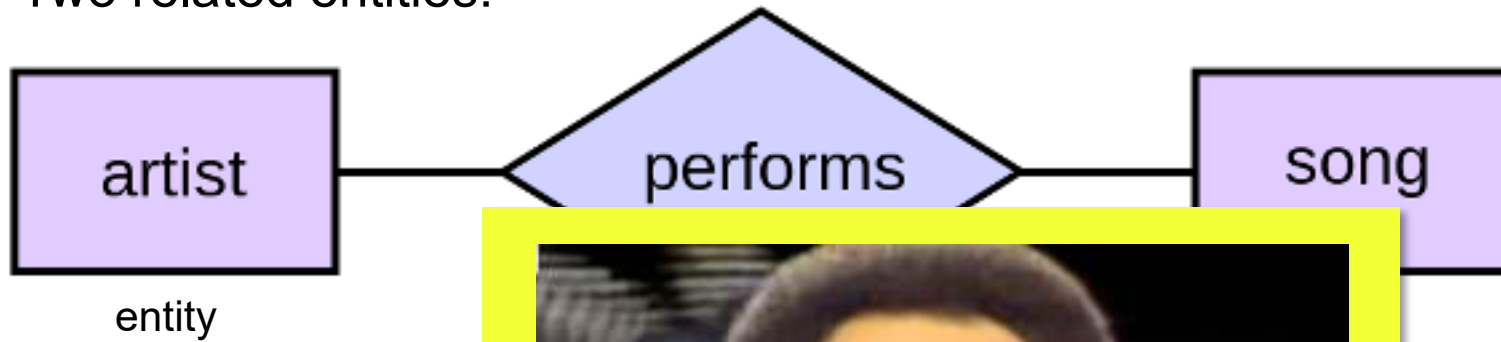
Two related entities:



Entities and relations with attributes:



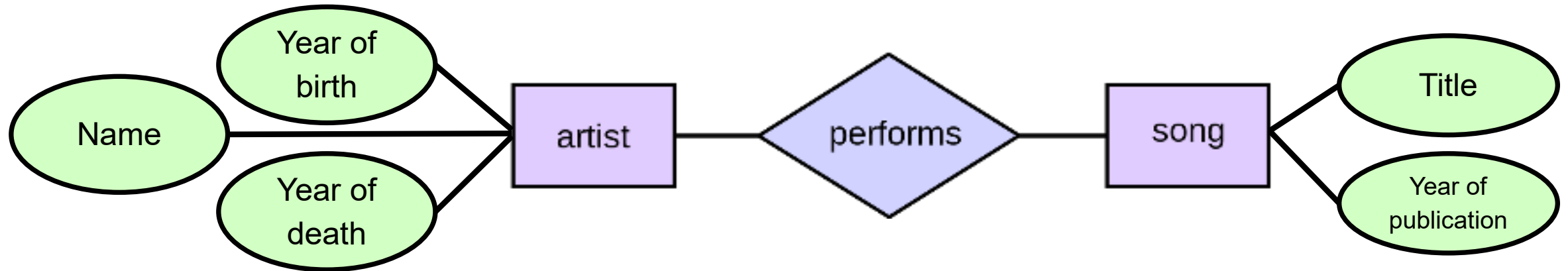
Two related entities:



Entities and relations with attributes



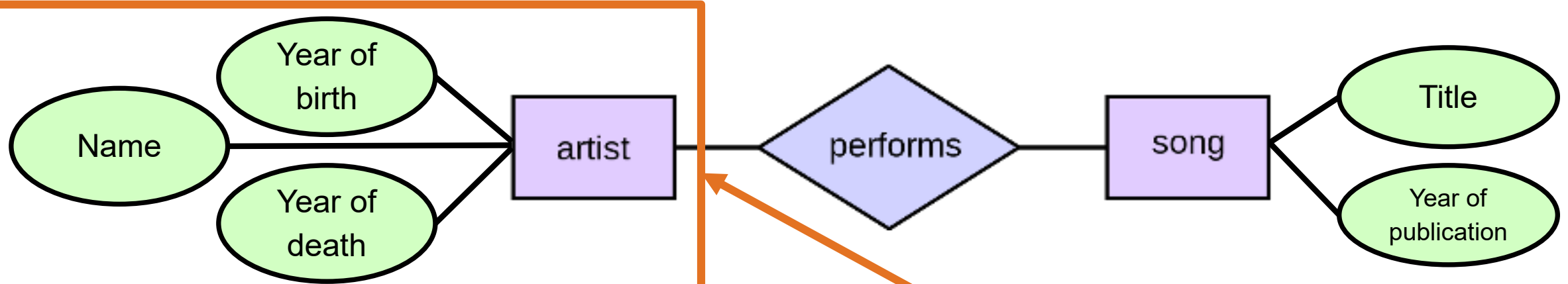
From ER to Relational Models



artist

Name	Year of Birth	Year of Death
Jimi Hendrix	1942	1970
Kurt Cobain	1967	1994
Amy Winehouse	1983	2011

From ER to Relational Models

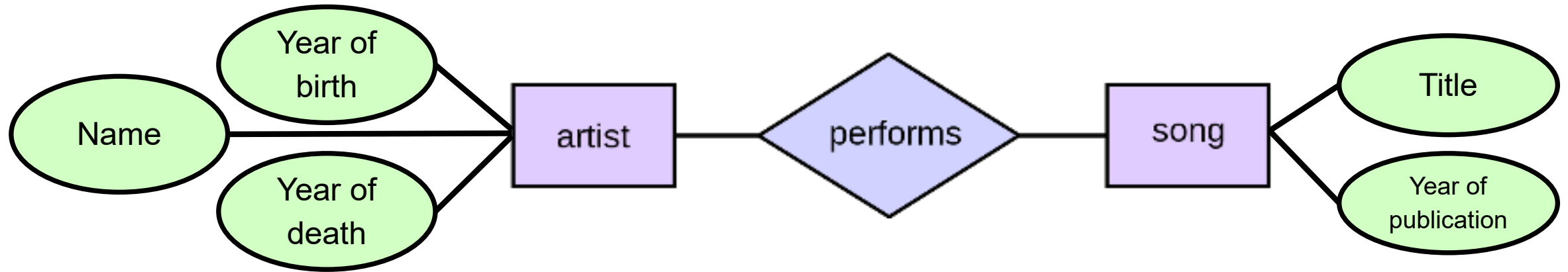


artist

Name	Year of Birth	Year of Death
Jimi Hendrix	1942	1970
Kurt Cobain	1967	1994
Amy Winehouse	1983	2011

Entity type

From ER to Relational Models

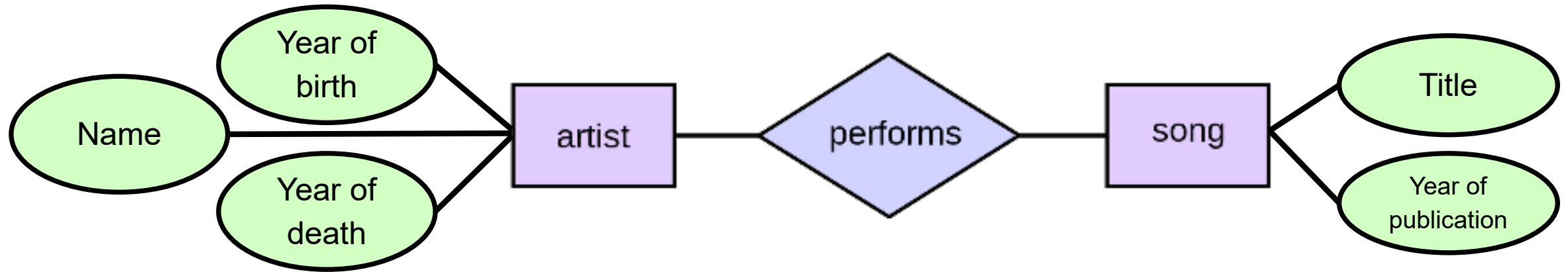


artist

Name	Year of Birth	Year of Death
Jimi Hendrix	1942	1970
Kurt Cobain	1967	1994
Amy Winehouse	1983	2011

Entity occurrence
/ instance

From ER to Relational Models



artist

Name	Year of Birth	Year of Death
Jimi Hendrix	1942	1970
Kurt Cobain	1967	1994
Amy Winehouse	1983	2011

song

Title	Year of publication
All Along the Watchtower	1968
Rehab	2006
Back to Black	2006
Come as you are	1991



From ER to Relational Models

ER models are readily used to represent relational database structures

ER models are used to develop a database schema

Entities and relations in ER terminology are both (!) relations in the database world

Relational algebra helps structuring and optimizing databases and tables



Relational Model

A relational algebra defines operations that can be applied to a set of relations.

For example, relations can be filtered, linked or aggregated.

The results of all operations are also relations.

Relations are basically tables with unique entries.

An entry of a relation (i.e. a row of a table) is called a tuple.

It helps thinking of the rows of a table as being elements of a set.

Adding the same tuple twice is pointless, because the tuples cannot be logically distinguished from each other.

Implicit consequence: introduction of identifiers of elements of the set

Information Retrieval

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

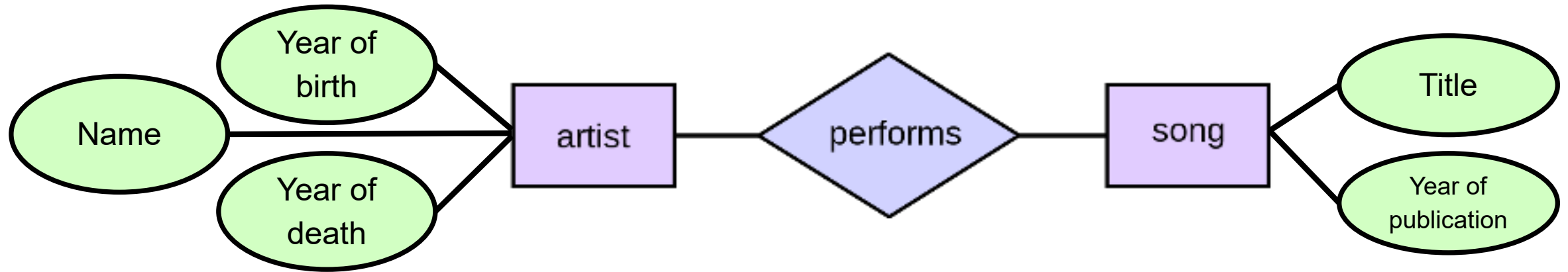
1970: Edgar Codd proposes the relational model

The one and only Peter Müller

Peter Muller, Peter Müller or **Peter Mueller** may refer to:

- [Peter Müller \(ice hockey\)](#) (1896-?), Swiss ice hockey player
- [Peter Muller \(architect\)](#) (born 1927), architect with works in Bali, Sydney, South Australia and Melbourne
- [Peter Müller \(boxer\)](#) (born 1928), Swiss boxer
- [Peter Müller \(footballer, born 1946\)](#), East German footballer
- [Peter Müller \(footballer, born 1948\)](#), West German footballer
- [Peter Mueller \(Canadian football\)](#) (born 1951), former tight end for the Toronto Argonauts
- [Peter Mueller \(speed skater\)](#) (born 1954), former US speed skater and speed skating coach
- [Peter Müller \(politician\)](#) (born 1955), German politician and judge
- [Peter Müller \(skier\)](#) (born 1957), Swiss alpine skier competing in the 1980s
- [Peter Müller \(co-driver\)](#) (born 1962), Austrian rally co-driver
- [Peter Müller \(footballer, born 1969\)](#), German footballer
- [Pete Muller \(photographer\)](#) (born 1982), news photographer
- [Peter Mueller \(ice hockey\)](#) (born 1988), American ice hockey player, playing in the NLA
- [Peter Muller \(actor\)](#), played Dr. Logan King in the television series *Shortland Street*
- [Pete Muller \(businessman and singer-songwriter\)](#), musician and founder and CEO of PDT Partners

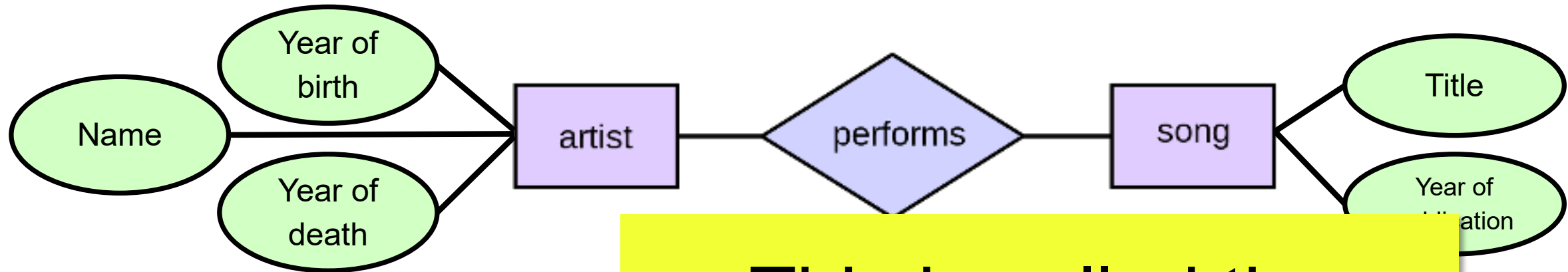
From ER to Relational Models



artist

AID	Name	Year of Birth	Year of Death
1	Jimi Hendrix	1942	1970
2	Kurt Cobain	1967	1994
3	Amy Winehouse	1983	2011

From ER to Relational Models



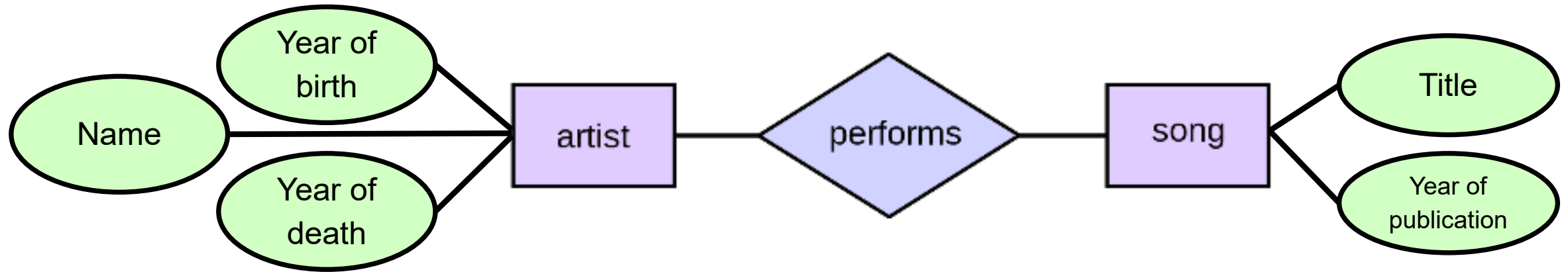
This is called the
primary key

artist

AID	Name	Year of Birth	Year of Death
1	Jimi Hendrix	1942	1970
2	Kurt Cobain	1967	1994
3	Amy Winehouse	1983	2011

Keys have to be unique
in terms of the **semantic**
of the real world!

From ER to Relational Models



artist

AID	Name	Year of Birth	Year of Death
1	Jimi Hendrix	1942	1970
2	Kurt Cobain	1967	1994
3	Amy Winehouse	1983	2011

song

SID	Title	Year of publication
1	All Along the Watchtower	1968
2	Rehab	2006
3	Back to Black	2006
4	Come as you are	1991



Minimizing redundancies

Key strength of the relational algebra: normalization

Normalization means to split columns into relations following certain rules such that redundancies are removed from the database.

Normal forms are classes of quality criteria for databases

Here, we focus on 1. NF, 2. NF, and 3. NF.

Informally, a database is “normalized” if it meets 3. NF

Why normalization?

Updates, insertions, and deletions should have no side-effects that impact database integrity, and these operations should affect as little parts of the database as possible.

Employees' Skills

Employee ID	Employee Address	Skill
426	87 Sycamore Grove	Typing
426	87 Sycamore Grove	Shorthand
519	94 Chestnut Street	Public Speaking
519	96 Walnut Avenue	Carpentry

An **update anomaly**. Employee 519 is shown as having different addresses on different records.

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201
424	Dr. Newsome	29-Mar-2007	?

An **insertion anomaly**. Until the new faculty member, Dr. Newsome, is assigned to teach at least one course, their details cannot be recorded.

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201

↓
DELETE

A **deletion anomaly**. All information about Dr. Giddens is lost if they temporarily ceases to be assigned to any courses.

1. Normal Form

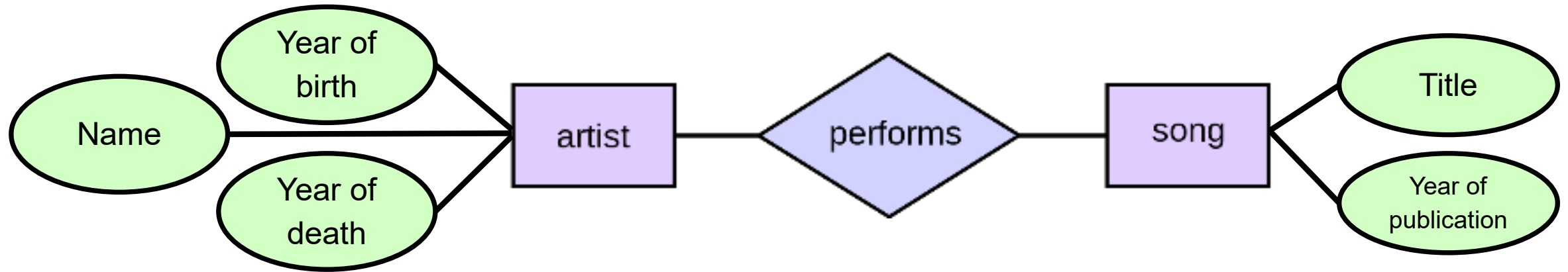
Relations have atomic attributes, i.e. no table-valued attributes and no repeating groups

AID	Name	Year of Birth	Year of Death
1	Jimi Hendrix	1942	1970
2	Kurt Cobain	1967	1994
3	Amy Winehouse	1983	2011



AID	Name	Given name	Year of Birth	Year of Death
1	Hendrix	Jimi	1942	1970
2	Cobain	Kurt	1967	1994
3	Winehouse	Amy	1983	2011

1. Normal Form



artist

AID	Name	Given name	Year of Birth	Year of Death
1	Hendrix	Jimi	1942	1970
2	Cobain	Kurt	1967	1994
3	Winehouse	Amy	1983	2011


song

SID	Title	Year of publication
1	All Along the Watchtower	1968
2	Rehab	2006
3	Back to Black	2006
4	Come as you are	1991

2. Normal Form

A table is in 1. normal form

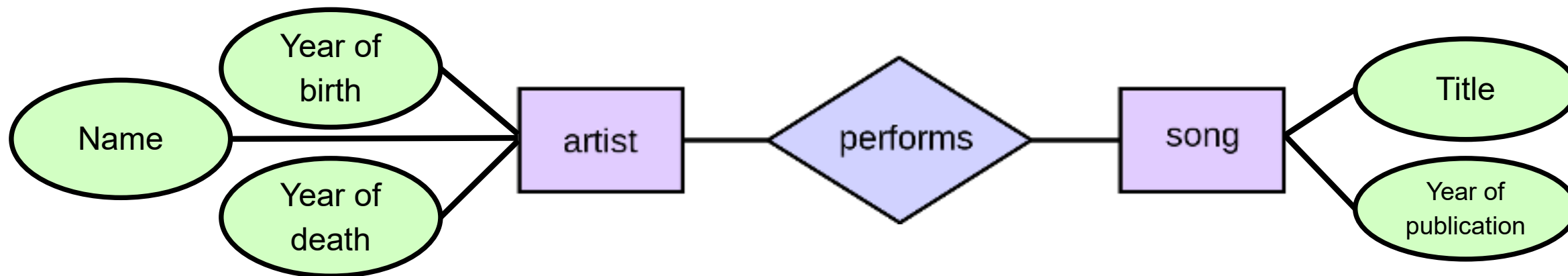
All non-prime attributes of the relation are dependent on the whole of every candidate key



Name	Given name	Year of Birth	Year of Death	<u>Title</u>	<u>Album</u>
Hendrix	Jimi	1942	1970	All Along the Watchtower	Electric Ladyland
Winehouse	Amy	1983	2011	Rehab	Back to Black
Winehouse	Amy	1983	2011	Back to Black	Back to Black
Cobain	Kurt	1967	1994	Come as you are	Nevermind

A violation of the 2. NF because non-prime attributes containing artist information depend on album only but not on the title of the track!

2. Normal Form



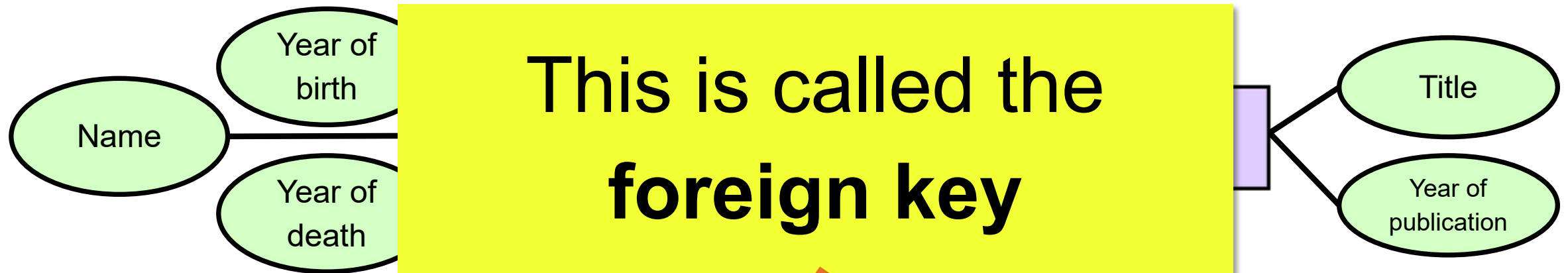
artist

AID	Name	Given name	Year of Birth	Year of Death
1	Hendrix	Jimi	1942	1970
2	Cobain	Kurt	1967	1994
3	Winehouse	Amy	1983	2011

song

SID	Title	Year of publication
1	All Along the Watchtower	1968
2	Rehab	2006
3	Back to Black	2006
4	Come as you are	1991

2. NF a) Foreign Keys



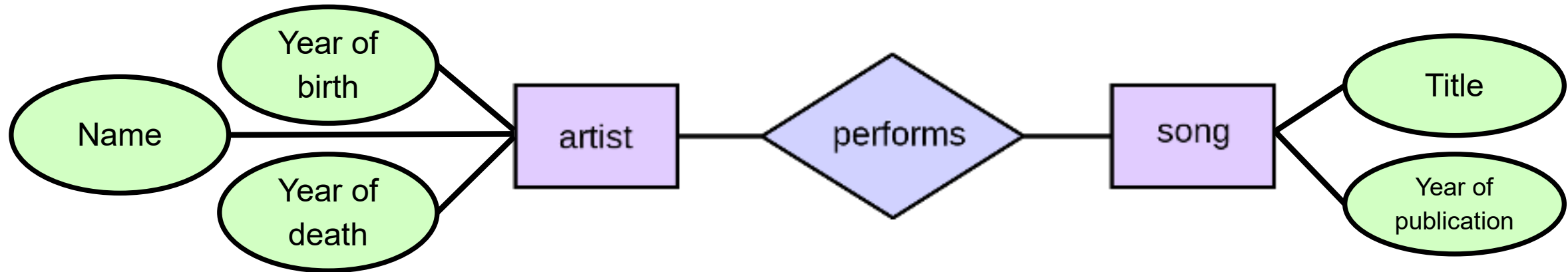
artist

AID	Name	Given name	Year of Birth	Year of Death
1	Hendrix	Jimi	1942	1970
2	Cobain	Kurt	1967	1994
3	Winehouse	Amy	1983	2011

song

SID	AID	Title	Year of publication
1	1	All Along the Watchtower	1968
2	3	Rehab	2006
3	3	Back to Black	2006
4	2	Come as you are	1991

2. NF b) Association table



artist

AID	Name	Given name	Year of Birth	Year of Death
1	Hendrix	Jimi	1942	1970
2	Cobain	Kurt	1967	1994
3	Winehouse	Amy	1983	2011

performs

AID	SID
1	1
2	4
3	2
3	3

song

SID	Title	Year of publication
1	All Along the Watchtower	1968
2	Rehab	2006
3	Back to Black	2006
4	Come as you are	1991

As you can see: both ER entities and relations can be relations in the relational model!

3. Normal Form

A table is in 2. NF

All attributes are functionally dependent solely on the primary key

<u>CD_ID</u>	Album title	Interpret	Year of foundation	Year of publication
1	Electric Ladyland	Jimi Hendrix	1963	1968
3	Back to Black	Amy Winehouse	2002	2006
4	Nevermind	Nirvana	1987	1991
5	Lioness: Hidden Treasures	Amy Winehouse	2002	2011

Album title and interpret depend on CD_ID but year of foundation has a transitive dependency on interpret.
Violation of 3.NF creates redundancy in the database.



Relational Databases

Database systems based on the relational model

Most relational databases use SQL as their query language

Still the de-facto industry standard

ACID transactions intend to guarantee data validity

- Atomicity: indivisible and irreducible series of operations
- Consistency: data is only changed in allowed ways. Ambiguous meaning in different systems
- Isolation: defining how and when database operations become visible to other users
- Durability: transactions that have been committed will survive permanently



NoSQL!

Non-relational databases provide a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

Provide 3 operations:

- Insert(k,v)
- Lookup(k)
- Delete(k)

Extremely simple and efficient from a provider-perspective

These databases scale!

Relaxed database consistency model.

Most Join/Selection/... operations are offloaded to the application

Semantic Web: RDF

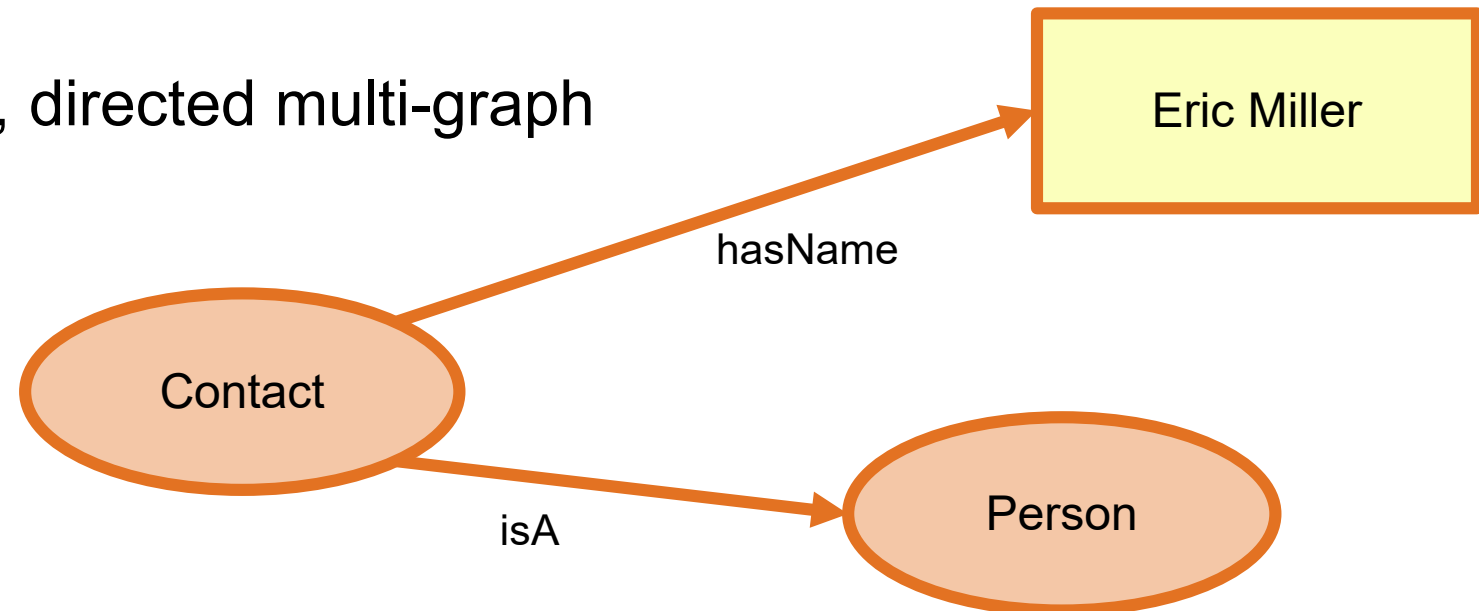
Classical modeling approach based on the idea of making statements about resources

Data model consisting of SPO-triples: Subject-Predicate-Object

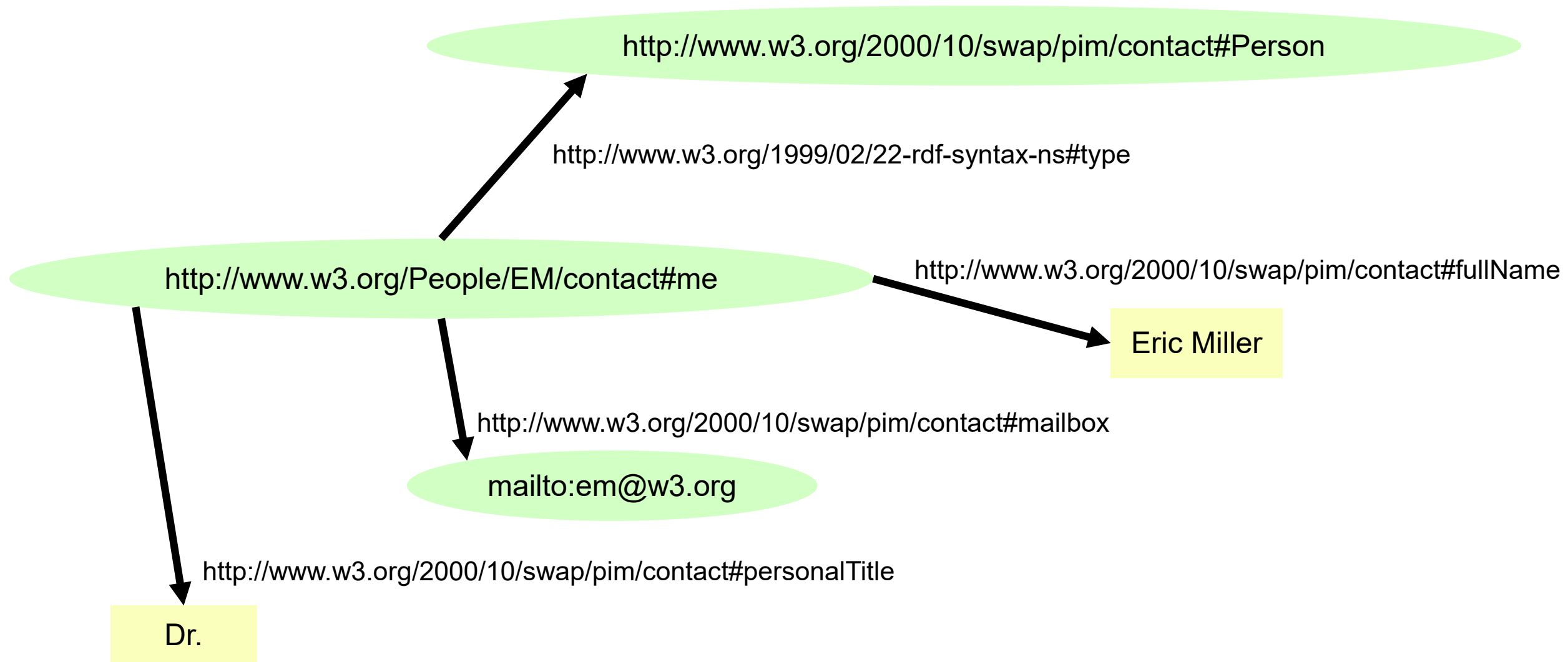
Subject represents the resource, the predicate denotes aspects of the resource and expresses relation between subject and object

Allows modeling disparate, abstract concepts very efficiently

Often visualized as a labeled, directed multi-graph



RDF Predicates are globally unique



Graph databases

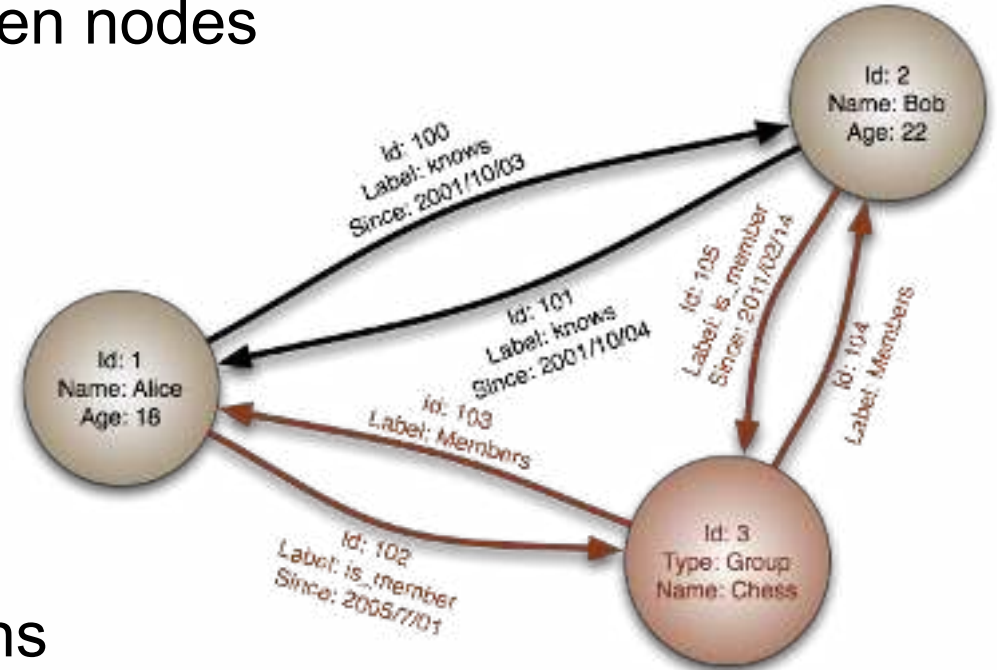
Graph databases explicitly model relations between nodes

Examples:

Social Graph, Consumption Graph,
Mobile Graph, Interest Graph, ...

Graph databases store pointers to records
of adjacent nodes eliminating costly join operations

True power of graph databases shows for queries that are more than one level deep,
i.e. querying friends of friends of friends...



Wrap-up

Data Modeling



Abstraction is Key

General meaning: identify basic rules and features from specific examples

In Computer Science:

- **Avoid repetition!**
Generalize program code so that it can be run in different contexts
- **Separate concepts and implementation!**
Separate abstract view on data from actual implementation of databases, for example

With all forms: there is no natural abstraction!

User requirements, world view, prejudices, and blind spots influence the process of abstraction

Steps of Data Modeling

1. Conceptual data modeling:
identification and description of that part of the world a modeler is modeling
notation of the findings
2. Logical data modeling:
defining tables of a database – translating the conceptual model
3. Physical data modeling:
optimizing the database design
actual implementation
usually not done by a data modeler

Best case:

Conceptual and Logical model are *independent* of an actual implementation

Distinction Between Steps

Conceptual Model

Organizes information

Makes logical model easy to derive

Captures semantics of information

Logical Model

Provides structure to the data that defines a set of suitable algorithms

Achieves computability (often through mathematical models)

Powerful formal abstraction

Thanks.

mirco.schoenfeld@uni-bayreuth.de