Detecting Duplicate Entities for Ontology Reconciliation

Bachelor - Abschlusskolloquium

Motivation

- In ontologies can be multiple entities for the same real world object: duplicates.

- E.g.: “Barbara Adler” taken from Freebase

Name	Facts	Occurrences
Barbara Adler	Musician, Poet, …	1
Barbara Adler	Screenwriter, Producer, …	2
Barbara Adler	Mother of Lauren Adler, Woman, Person	2
Barbara Adler	Person appearing in film	1
Barbara Adler	Mother of Amy and Richie Adler, Wife of Abe Adler, Woman, Person	1
Barbara Adler	Woman, Person	202

Motivation II

- Information scattered across multiple entities
- Redundant information is stored
- For Semantic Search: Pollutes the results

Causes for Duplicity

- Merging of different ontologies
- Favoring of duplicates over information loss
- Addition of user-generated content
- Generally: Depends on definition of '*real world object*'

Challenges

- 22+ GB of data
- 5000+ different predicates
- No domain specific solution
- Entities with few and with many relations
- Preferably: traceable decision-making process

Detecting Duplicates

1. Preprocess ontological data to allow chunk-wise processing

2. Chunk-wise fill data structure with entity data

3. Pair-wise compare relations of entities with the same name

4. Depending on outcome of weighted relation comparisons, mark as duplicates

Preprocessing

- Repeat triple for reverse occurrences & sort
- Preprocessing allows chunkwise iteration

Mars (m/09cws)	Orbited by	Phobos (m/0dxqj)
…	…	…
Phobos (m/0dxqj)	is-a	Moon (m/03yxlwb)
Phobos (m/0dxqj)	is-a	Natural satellite (m/0d_23)
Phobos (m/0dxqj)	*Orbited by (reversed)*	*Mars (m/09cws)*
…	…	…
Phobos (m/0krzj)	is-a	Deity (m/02knxz7)
Phobos (m/0krzj)	Parents	Aphrodite (m/0mpp)
…	…	…

Data structure (entity)

- Entities consist of: id, name, float-relations, date-relations and set-relations

Phobos (m/0dxqj)	is-a	Moon (m/03yxlwb)
Phobos (m/0dxqj)	is-a	Natural satellite (m/0d_23)
Phobos (m/0dxqj)	Orbited by (reversed)	Mars (m/09cws)
Phobos (m/0dxqj)	Discovery Date	1977-08-18
Phobos (m/0dxqj)	Orbital Period (days)	0.3189
Phobos (m/0dxqj)	Orbit Eccentricity	0.0151

Data structure (entity)

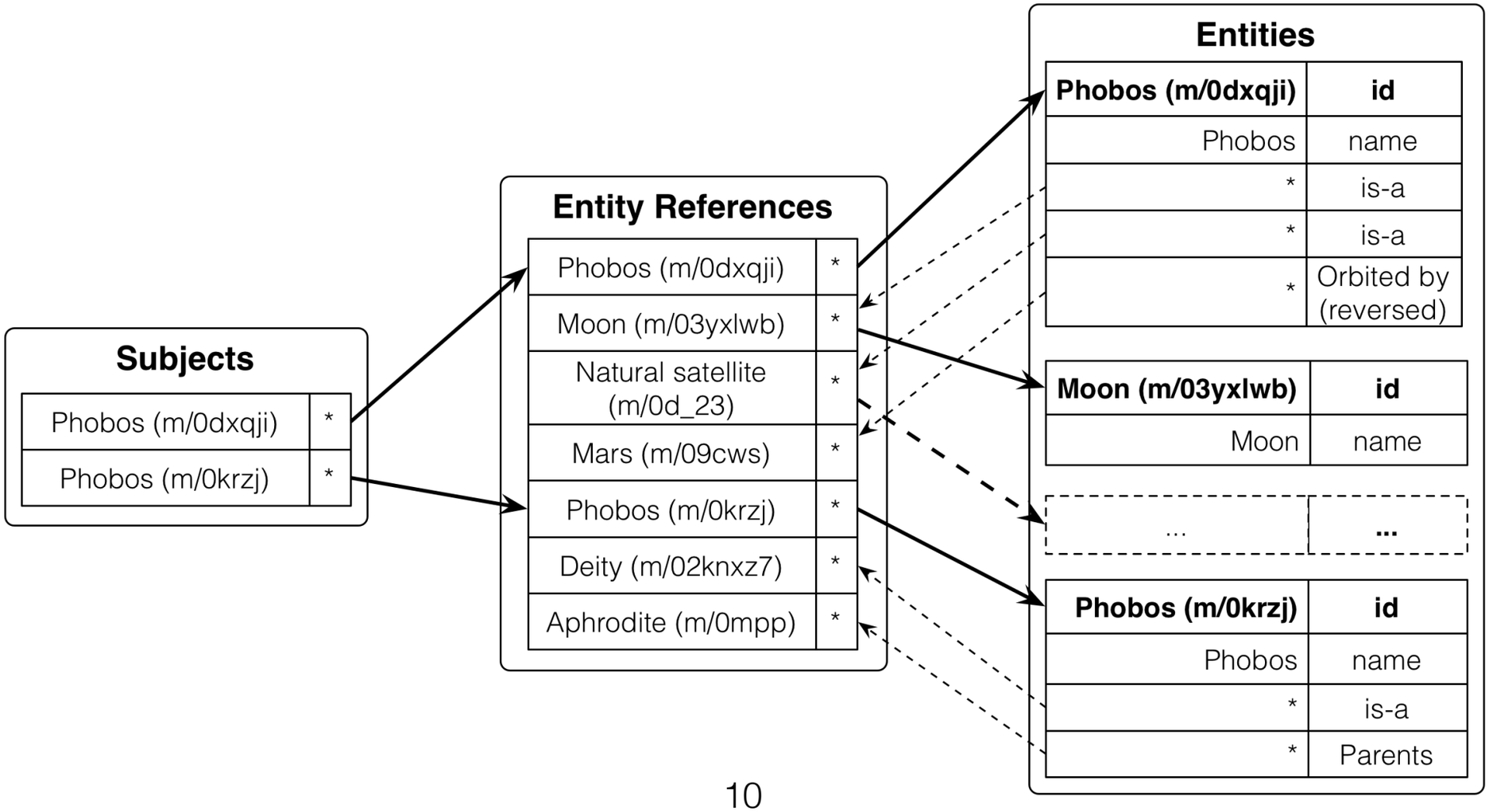
- Entities consist of: id, name, float-relations, date-relations and set-relations

- id: **Phobos (m/0dxqj)**, name: **Phobos**

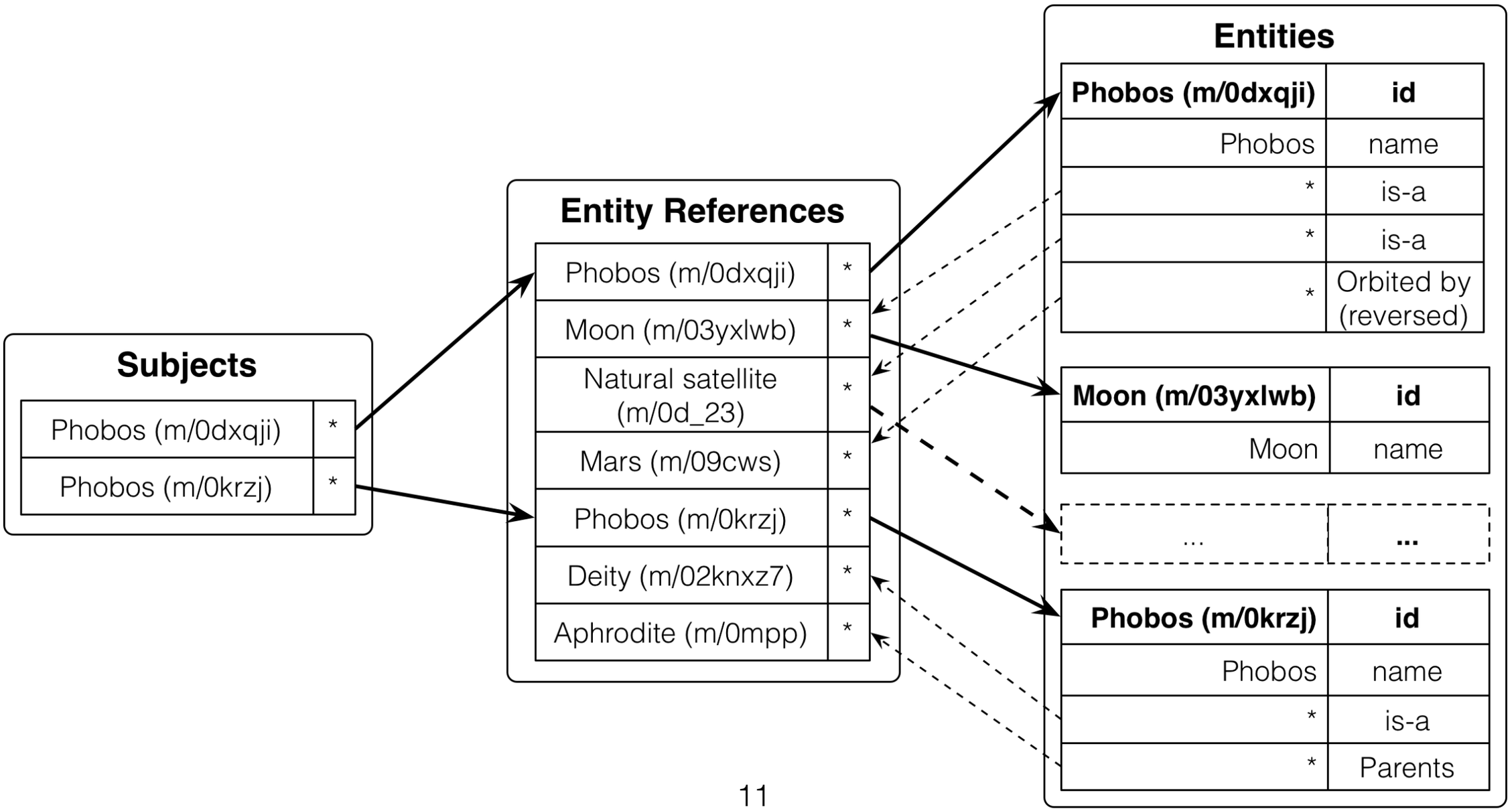
set relation	is-a	Moon (m/03yxlwb) Natural satellite (m/0d_23)
set relation	Orbited by (reversed)	Mars (m/09cws)
date relation	Discovery Date	1977-08-18
float relation	Orbital Period (days)	0.3189
float relation	Orbit Eccentricity	0.0151

Data structure

- Subjects *and* objects are treated as entities
- Additional layer of indirection for linking duplicates



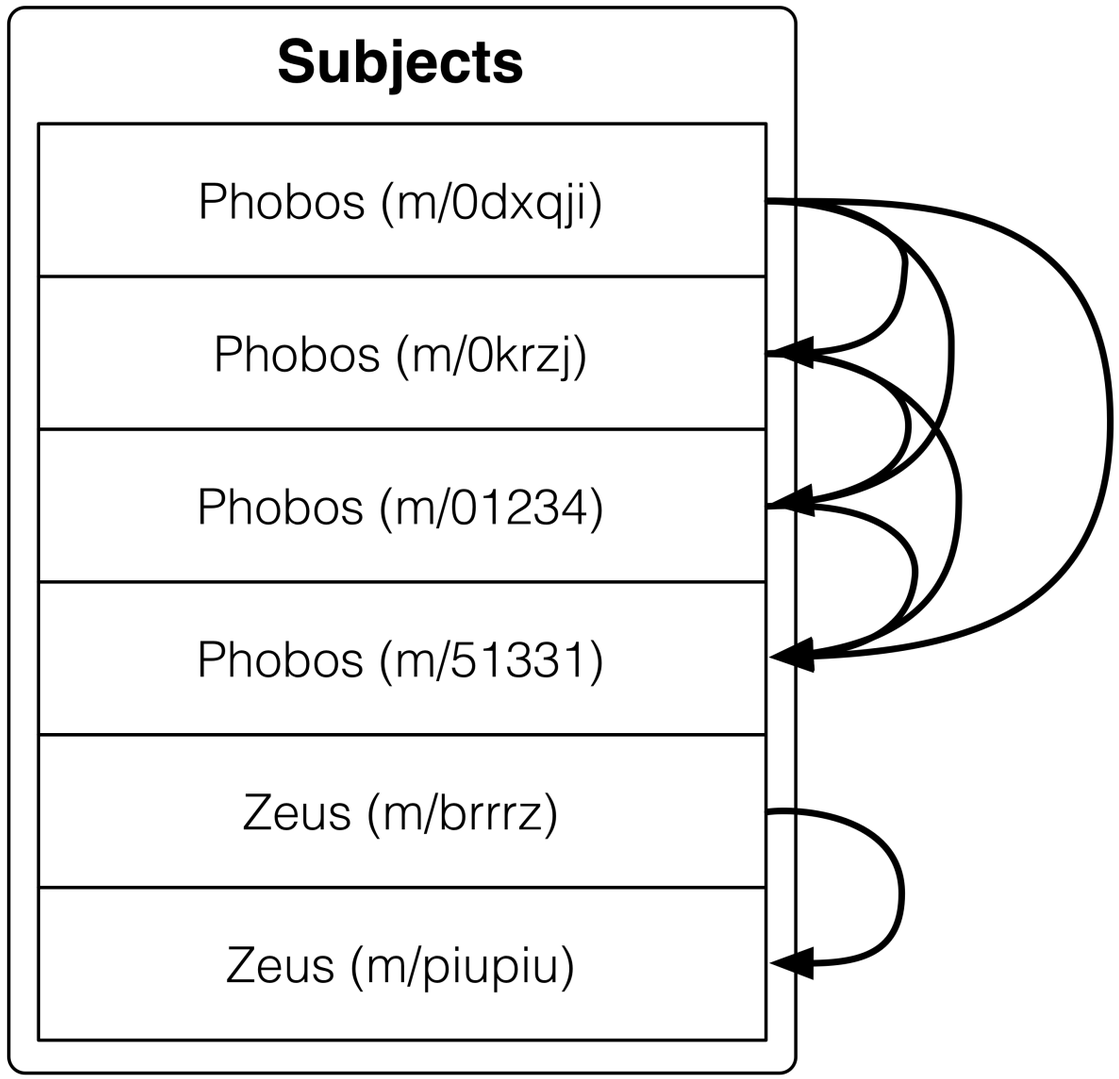
Phobos (m/0dxqj)	is-a	Moon (m/03yxlwb)
Phobos (m/0dxqj)	is-a	Natural satellite (m/0d_23)
Phobos (m/0dxqj)	Orbited by (reversed)	Mars (m/09cws)
Phobos (m/0krzj)	is-a	Deity (m/02knxz7)
Phobos (m/0krzj)	Parents	Aphrodite (m/0mpp)



Comparison

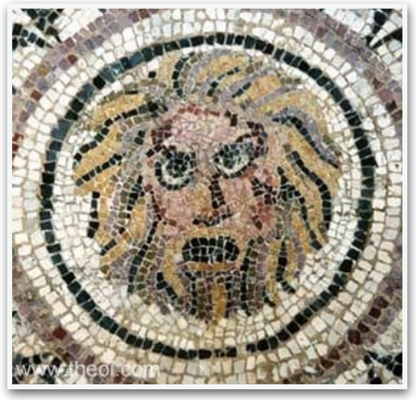
- Pair-wise comparison of entities with the same name
- Measure similarity of two entities based on their relations
- Scoring schema for each relation

Pairwise Comparison



- Pair-wise comparison of each entity with the same name
- $\frac{n(n-1)}{2}$ comparisons per name

Phobos vs Phobos



vs

id	is-a	Parents	Orbited by (reverse)
Phobos (m/0dxqji)	• Moon • Natural satellite	-	• Mars
Phobos (m/0krzj)	• Deity	• Aphrodite	-

- 1 common relation with disjoint facts
- 1 mutually disjoint relation

➡ *No way.*

Scoring schema

id	Starred in	Spouse	Date of Birth
Brad Pitt (m/zomb13)	• World War Z • Moneyball	• Angelina Jolie	1963-12-18
Brad Pitt (m/1_rul3)	• Fight Club • Mr. & Mrs. Smith	• Angelina Jolie	1963-12-18

id	Gender	Nationality	Date of Birth
Bob Miller (m/0815)	• Male	• United States	1988-04-20
Bob Miller (m/0814)	• Male	• United States	1945-12-24

➡ Some relations are more defining than others

Scoring schema

id	is-a	Parents	Siblings
Phobos (m/h0rr0r)	• Deity	• Aphrodite	-
Phobos (m/0krzj)	• Deity	• Aphrodite • Ares	• Deimos • Eros

➡ Additional facts to common or new relations

id	is-a	Children	Spouse
Bob Mayer (m/0815)	• Person	• Lady Gaga	-
Bob Mayer (m/0816)	• Person	-	• Marissa Mayer

➡ Mutually disjoint relations can be defining

Scoring schema

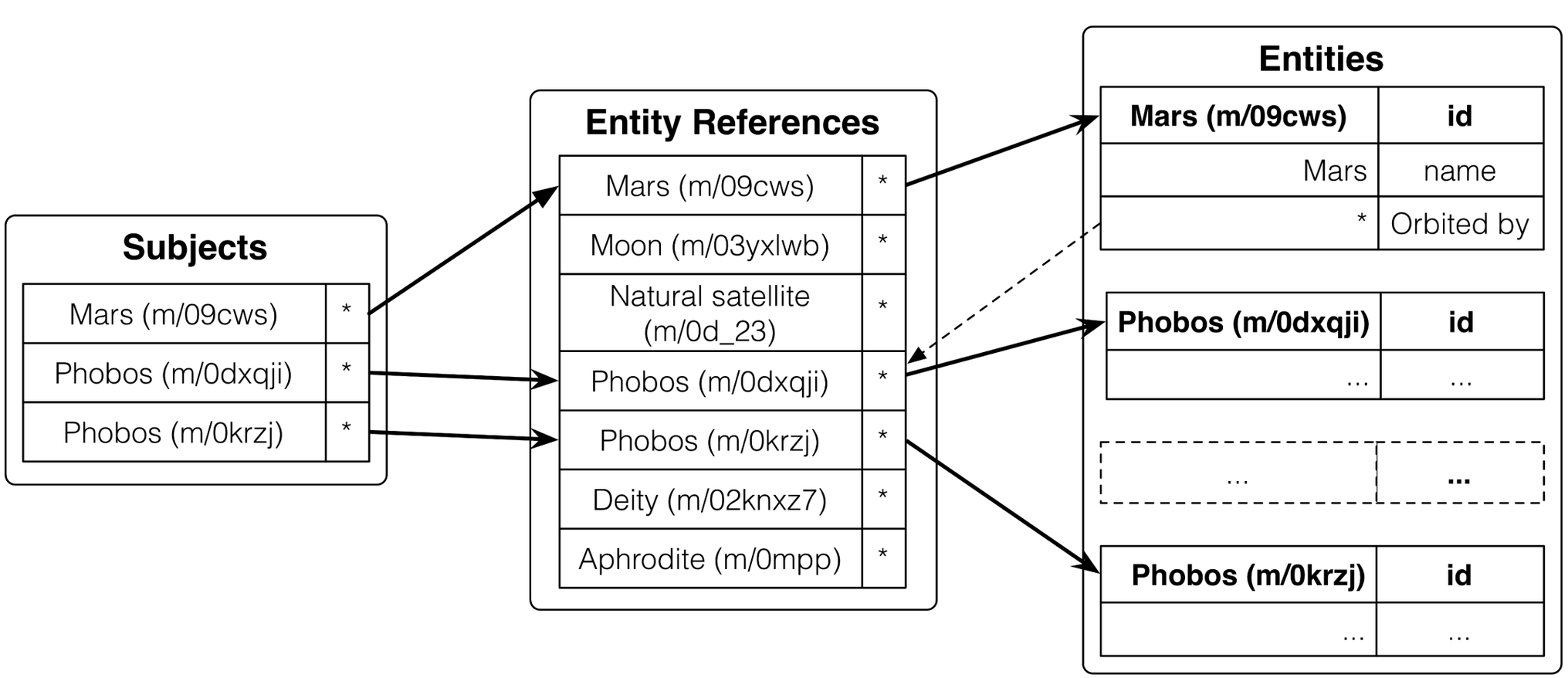
- Sum of scores over common relations divided by the smaller entities number of relations
- I.e.: average score over common relations and a penalty for every mutually disjoint relation

$$score(e_1, e_2) = \frac{\sum_{r \in R_1 \cap R_2} score_r}{min(|R_1|, |R_2|)}$$

- This score needs to overcome the global threshold

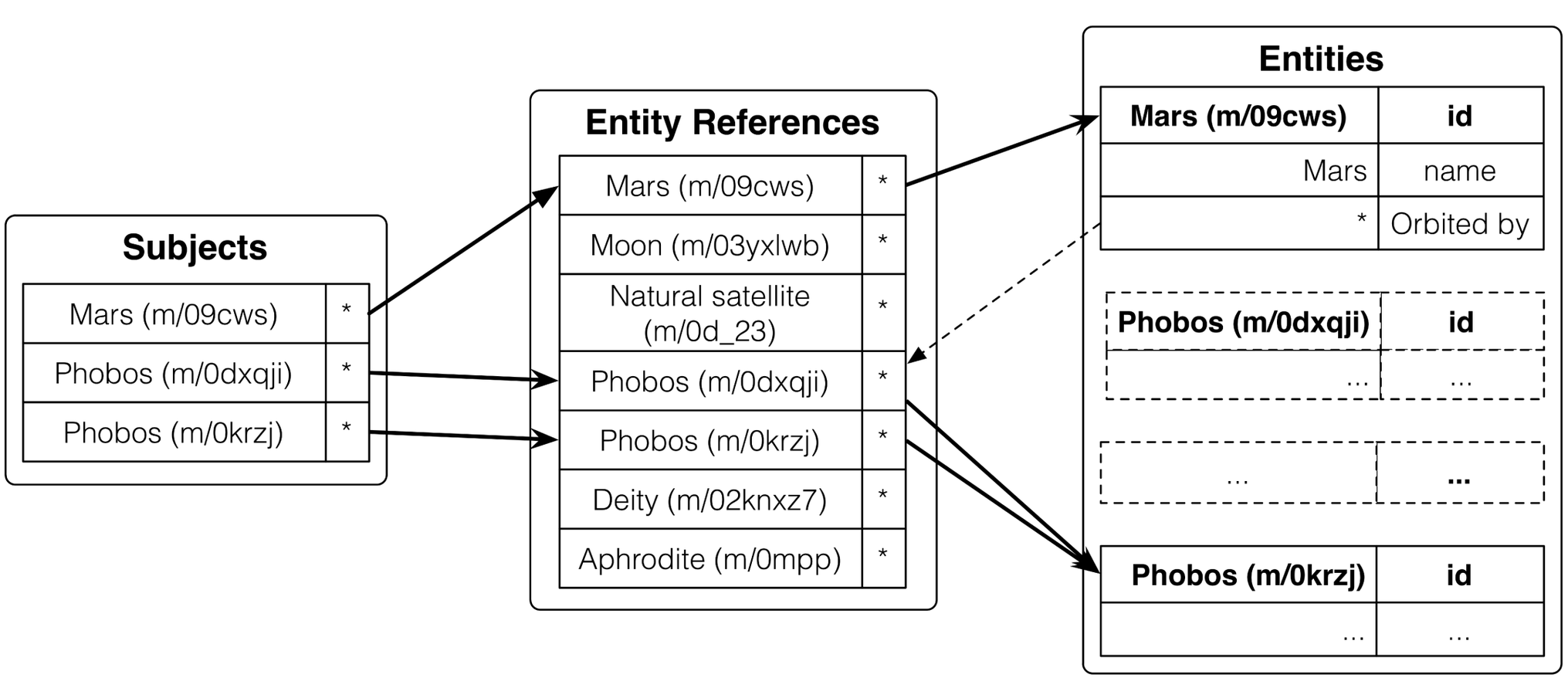
Merging Entities

- Redirection of the entity's reference & merging of their relations



Merging Entities

- Improves chances of future comparisons
- Duplicates are saved across chunks



Application

- Implementation in C++ with g++ and c++-11
- Input ontology (~22GB):
- 300m facts, 50m entities, 30m different names
- 52% of entities share a name with another entity
- only 18% of names are used by multiple entities

Results

- 1.2m, 977k and 810k duplicates with a medium, high and very high global threshold respectively

- Medium threshold: 94% correct, 2% false, 4% undecided

- Entities with few relations are more troublesome

- Quite a lot of entities are only stumps and have no uniquely identifiable information