ADDING SEMANTIC TO IR

Adding Semantics to IR (or Adding Ranking to DB)

	Structured data (records)	Unstructured data (documents)
Unstructured search (keywords)	Keyword Search on Relational Graphs (BANKS, Discover, DBexplorer, …) + Web 2.0	IR Systems Search Engines + Digital Libraries + Enterprise Search
Structured search (SQL,XQuery)	DB Systems + Text + Relax. & Approx. + Ranking	Querying entities & relations from IE (Libra, ExDB, NAGA, …)

Why semantic search?

- "We are at the beginning of search." (Marissa Mayer)
  - Solved large classes of queries, e.g. navigational
  - Heavy investment in computational power
  - Remaining queries are hard, not solvable by brute force, and require a deep understanding of the world and human cognition
- Background knowledge and metadata can help to address poorly solved queries

Poorly solved information needs

- Ambiguous searches
  - paris hilton
- Long tail queries
  - george bush (and I mean the beer brewe
- Multimedia search
  - paris hilton sexy
- Imprecise or overly precise searches
  - jim hendler
  - pictures of strong adventures people
- Precise searches for descriptions
  - countries in africa
  - 32 year old computer scientist living in barcelona
  - reliable digital camera under 300 dollars

Many of these queries would not be asked by users, who learned over time what search technology can and can not do.

Document retrieval and data retrieval

- Information Retrieval (IR) support the retrieval of documents (document retrieval)
  - Representation based on lightweight syntax-centric models
  - Work well for topical search
  - Not so well for more complex information needs
  - Web scale
- Database (DB) and Knowledge-based Systems (KB) deliver more precise answers (data retrieval)
  - More expressive models
  - Allow for complex queries
  - Retrieve concrete answers that precisely match queries
  - Not just matching and filtering, but also joins
  - Limitations in scalability

Semantic search

- Target (combination of) document and data retrieval
- Semantic search is a retrieval paradigm that
  - Exploits the structure/semantics of the data or explicit background knowledge to understand user intent and the meaning of content
  - Incorporates the intent of the query and the meaning of content into the search process (**semantic models**)
- Wide range of semantic search systems
  - Employ different semantic models, possibly at **different steps** of the search process and in order to support **different tasks**

Combination of data and document

- Documents with metadata
  - Metadata may be embedded inside the document
  - *I'm looking for **documents** that mention countries in Africa.*
- Data retrieval
  - Structured data, but searchable text fields
  - *I'm looking for **directors**, who have directed movies where the synopsis mentions dinosaurs.*

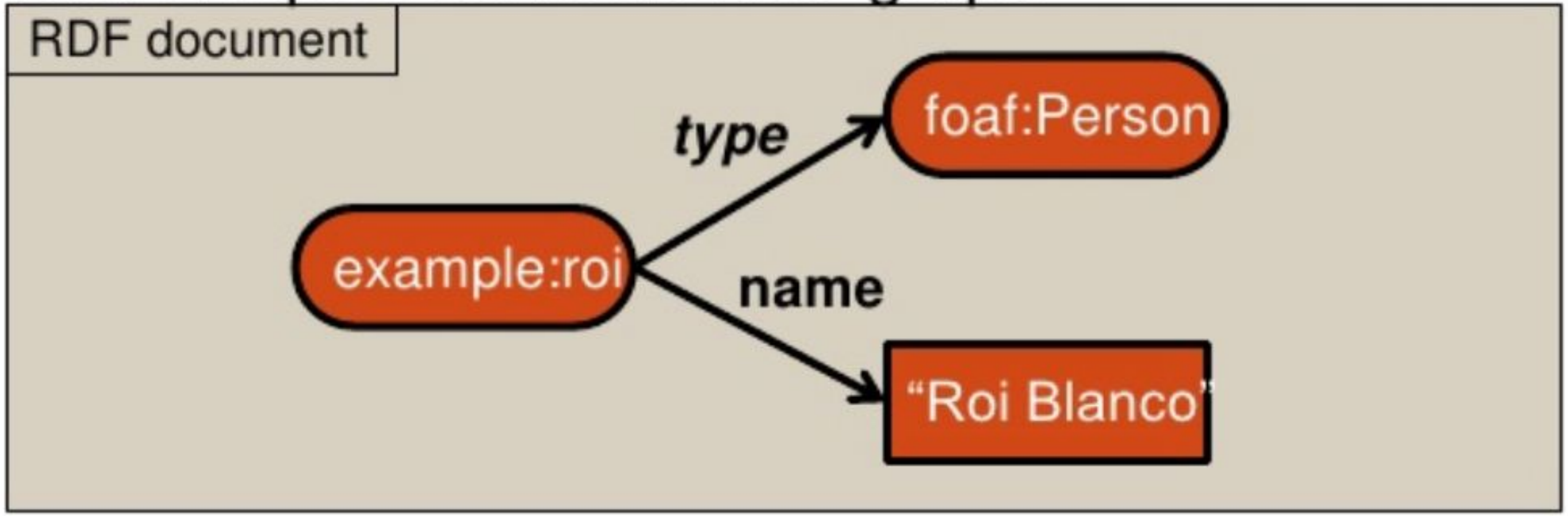
Semantic web data

- Data on the Web is not directly accessible
  - Most web pages are generated from databases, but formatted for human consumption
  - APIs offer limited views over data
- Two solutions
  - Extraction using **Information Extraction** (IE) techniqu
    - Out of scope for this tutorial
  - Relying on publishers to expose structured data using standard **Semantic Web** formats

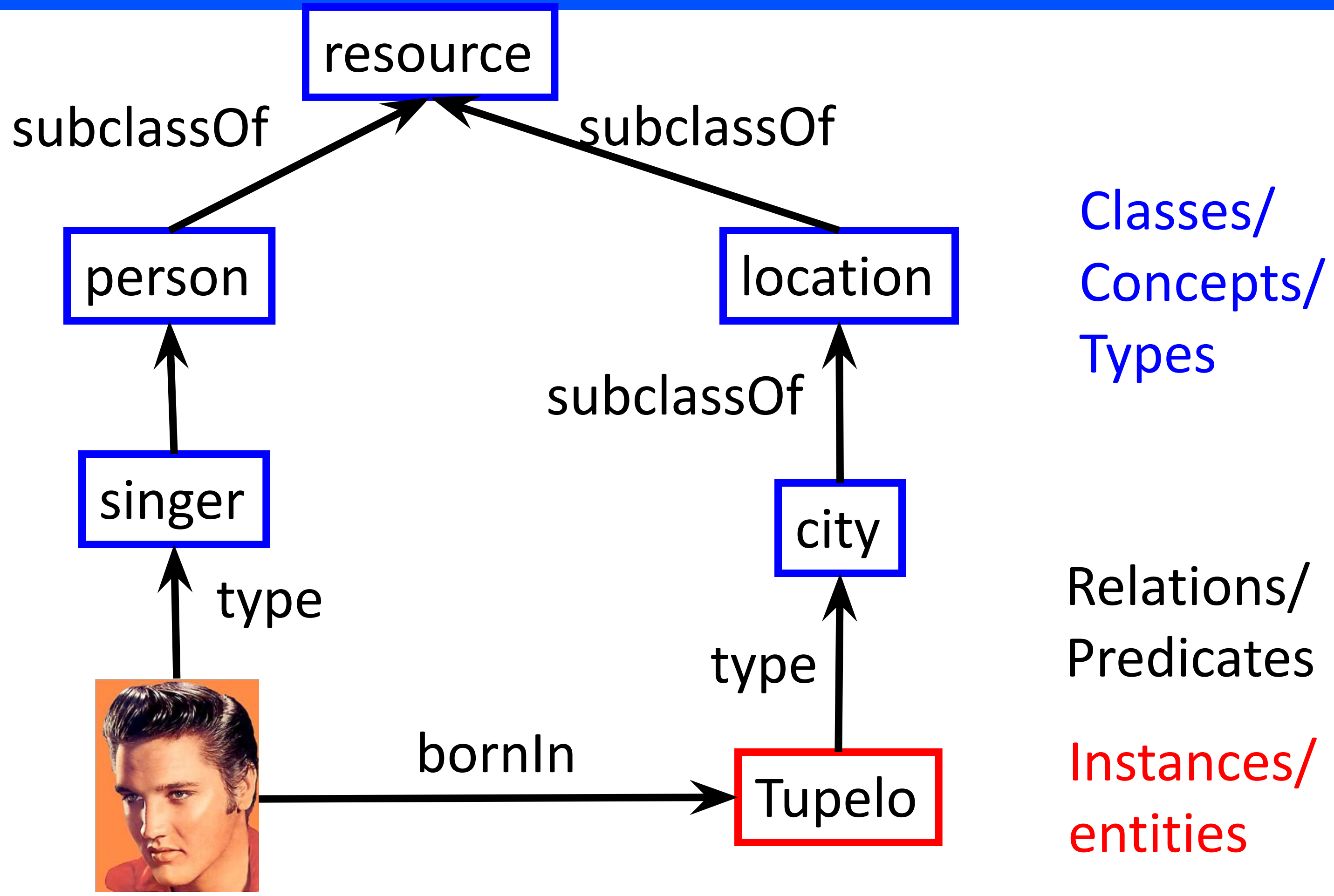
- Natural Language Processing
  - Named entity recognition and disambiguation, sentiment analysis etc.
- Extraction of information about entities
  - *Suchanek et al. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, WWW, 2007.*
  - *Wu and Weld. Autonomously Semantifying Wikipedia, CIKM 2007.*
- Extraction from HTML tables
  - *Cafarella et al. WebTables: Exploring the Power of Tables on the Web. VLDB 2008*
- Wrapper induction
  - *Kushmerick et al. Wrapper Induction for Information ExtractionText extraction. IJCAI 2007*
- Filling web forms automatically (form-filling)
  - *Madhavan et al. Google's Deep-Web Crawl. VLDB 2008*

- Sharing data across the Web
  - Standard data model
    - RDF
  - A number of syntaxes (file formats)
    - RDF/XML, RDFa
  - Powerful, logic-based languages for schemas
    - OWL, RIF
  - Query languages and protocols
    - HTTP, SPARQL

- Each resource (thing, entity) is identified by a URI
  - Globally unique identifiers
  - URLs a subset of URIs
  - Often abbreviated using namespaces
    - e.g. example:roi = http://example.org/roi
- RDF represents knowledge as a set of triples
  - Each triple is a single fact about the entity (an attribute or a relationship)
- A set of triples forms an RDF graph

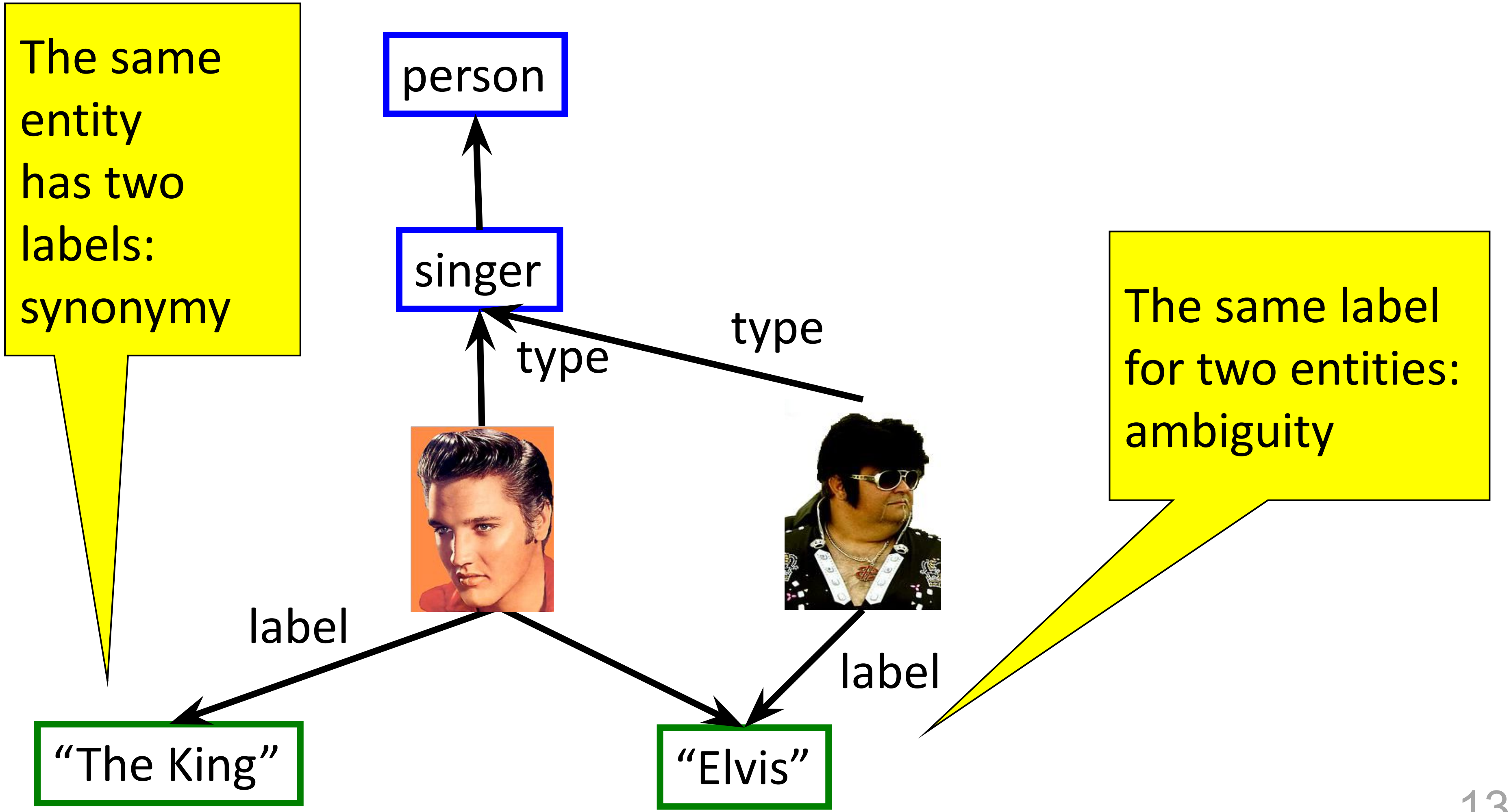


Knowledge Bases are labeled graphs



A knowledge base can be seen as a directed labeled multi-graph, where the nodes are entities and the edges relations.

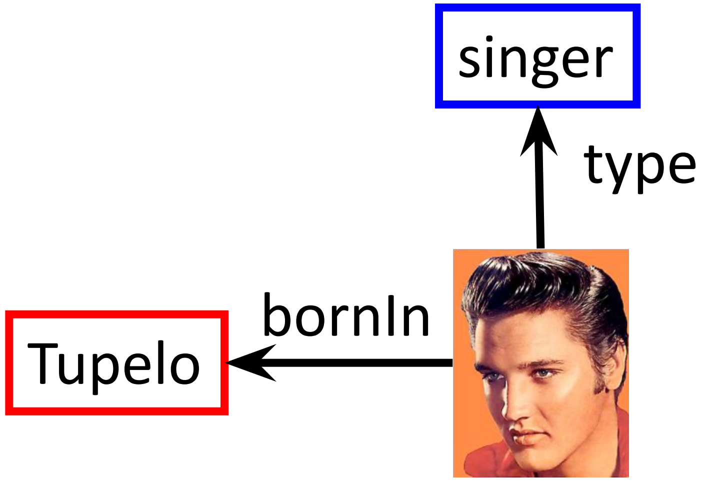
An entity can have different labels



Different views of a knowledge base

We use "RDFS Ontology" and "Knowledge Base (KB)" synonymously.

Graph notation:



Triple notation:

Subject	Predicate	Object
Elvis	type	singer
Elvis	bornIn	Tupelo
...	...	...

Logical notation:

type(Elvis, singer)
bornIn(Elvis,Tupelo)
...

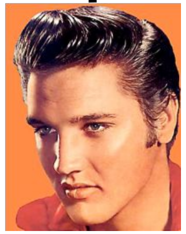
The Goal is finding classes and instances: entities search

person

subclassOf

singer

type



Which classes exist?
(aka entity types, unary predicates, concepts)

Which subsumptions hold?

Which entities belong to which classes?

Which entities exist?

Don‘t Let Me Be Misunderstood

Keyword query: Max Planck

or

Keyword query: Greek art Paris

or

Semantic Search

Concept query:
Person = „Max Planck“

Concept query:
„Greek art“
& Location = „Paris“

Entity search

Instead of „interpreting“ text with background knowledge, extract facts and search entities, attributes, and relations

<u>Motivation and Applications:</u>

- Web search for vertical domains
  (products, traveling, entertainment, scholarly publications, intelligence agencies, etc.)
- preparation for natural-language QA
- step towards better Deep-Web search, digital libraries, e-science

<u>Example systems:</u>

- Libra (MSR), EntityRank (UIUC), ExDB (UW Seattle), NAGA (MPII), …
- probably all commercial search engines have some support for entities

<u>Typical system architecture:</u>

Information Extraction (IE): Text to records

Information Extraction (IE): Text to Records

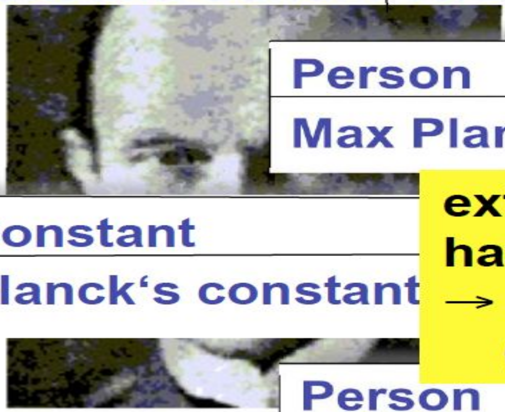
Max Planck

Max Karl Ernst Ludwig Planck (April 23, 1858 – October 4, 1947) was a German physicist wh… is considered to be the inventor of quantum theory.

Born in Kiel, Planck started his physics studies at Munich University in 1874, graduating in 1879 in Berlin. He returned to München in 1880 to teach at the university, and moved to Kiel in 1885. There he married Marie Merck in 1886. In 1889, he moved to Berlin, where from 1892 on he held the chair of theoretical physics.

In 1899, he discovered a new fundamental constant, which is named Planck's constant, and is, for example, used to calculate the energy of a photon. Also that year, he de… own set of units of measurement based on fundamental physical constan… year later, he discovered the law of heat radiation, which is named Planck's … Radiation. This law became the basis of quantum theory, which emer… later in cooperation with Albert Einstein and Niels Bohr.

Person	BirthDate	BirthPlace	…
Max Planck	4/23, 1858	Kiel	
Albert Einstein	3/14, 1879	Ulm	
Mahatma Gandhi	10/2, 1869	Porbandar	



Person	ScientificResult
Max Planck	Quantum Theory

Constant
Planck's constant

extracted facts often have confidence < 1
→ DB with uncertainty (probabilistic DB)

Person	Collaborator
Max Planck	Albert Einstein
Max Planck	Niels Bohr

Person	Organization
Max Planck	KWG / MPG

combine NLP, pattern matching, lexicons, statistical learning

IE Technology: Rules, Patterns, Learning

For heterogeneous sources and for natural-language text:

- **NLP techniques** (parser, PoS tagging) for tokenization
- **identify patterns** (regular expressions) as features
- **train statistical learners** for segmentation and labeling (HMM, CRF, SVM, etc.), augmented with lexicons
- use learned model to **automatically tag** newly seen input