InfoPlanet

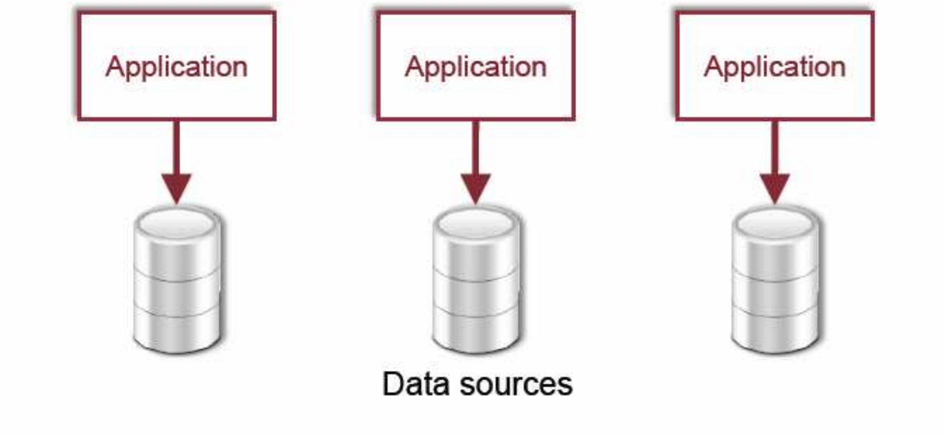
Università e Ricerca aprono la strada a nuovi utilizzi aziendali per dati e informazioni

Roberto Sicconi, Director DeepQA Opportunities, IBM USA

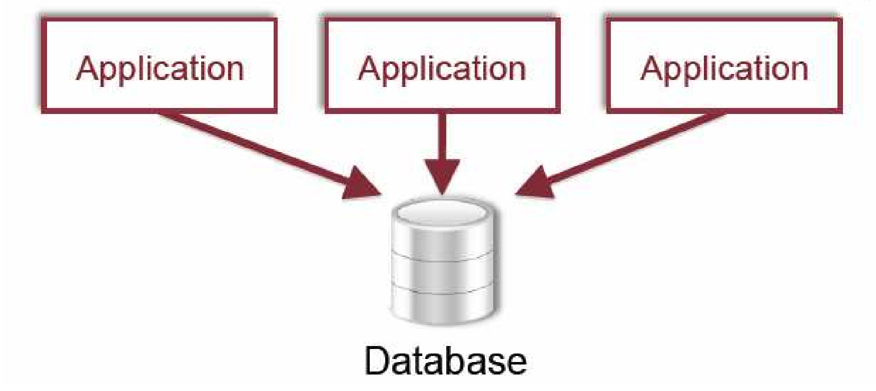
Maurizio Lenzerini, Professore Ordinario di Base dei Dati, Università La Sapienza di Roma

Information system architecture enabled by DBMS

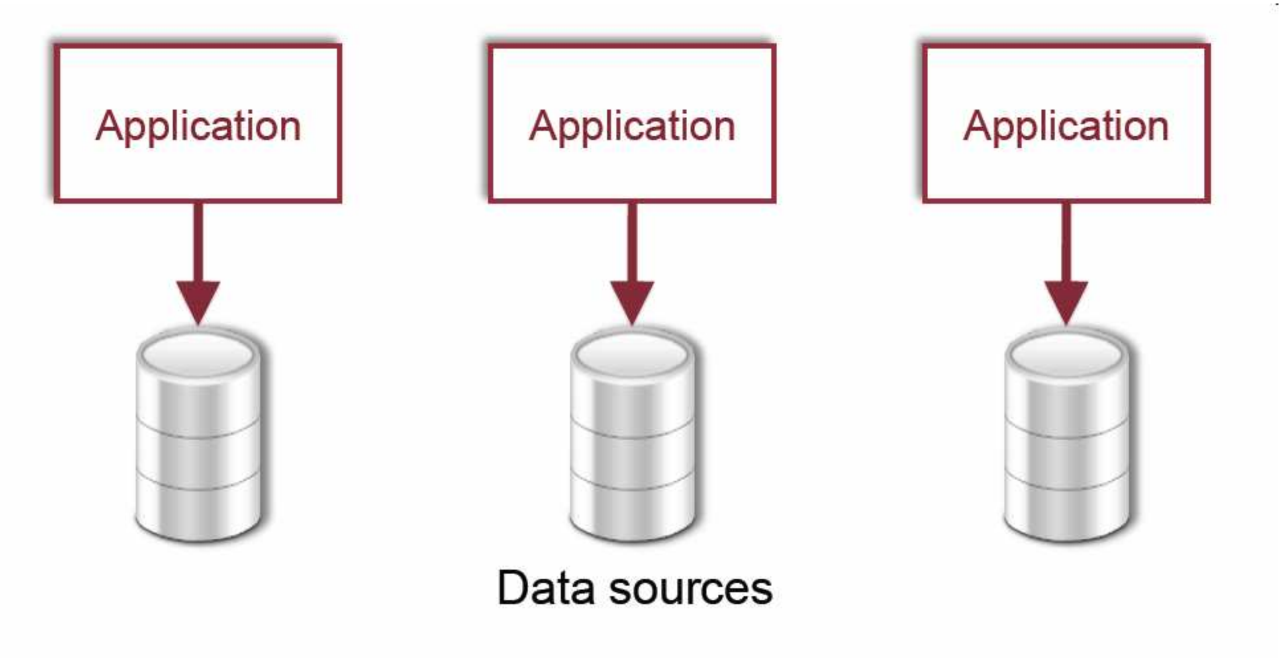
Pre-DBMS architecture (need of a unified data storage):



“Ideal information system architecture” with DBMS (’80s):



Actual information system structure



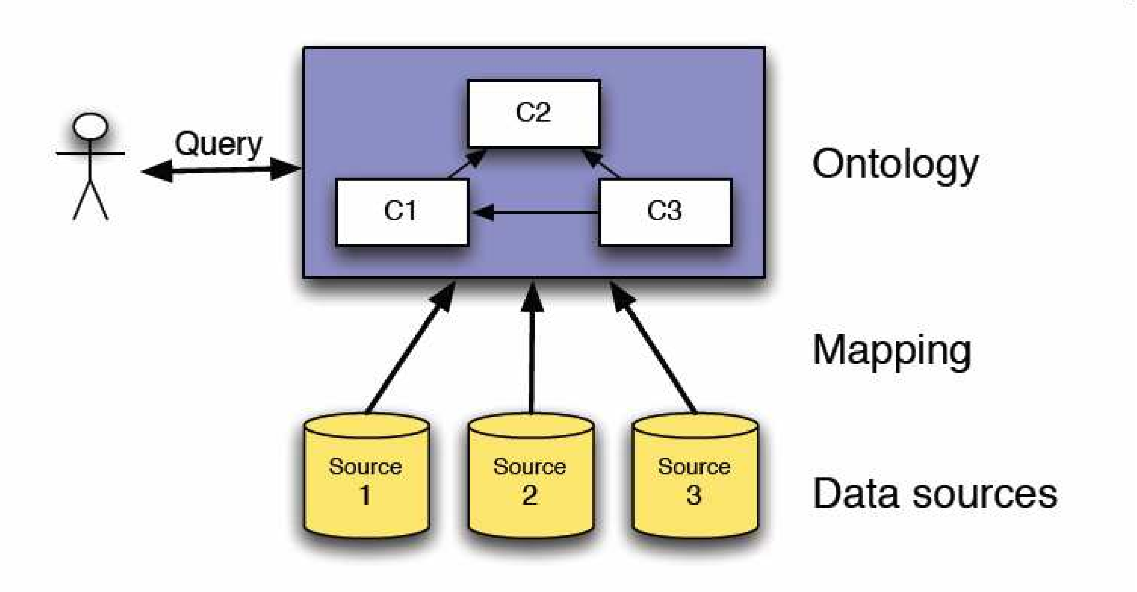
- Distributed, redundant, application-dependent, and mutually incoherent data
- Desperate need of a coherent, conceptual, unified view of data

Ontology-based data management: basic idea

Use Knowledge Representation and Reasoning principles and techniques for a new way of managing data.

- Leave the data where they are
- Build a conceptual specification of the domain of interest, in terms of knowledge structures (**semantic transparency**)
- Map such knowledge structures to concrete resources (e.g., data sources)
- Express all services over the abstract representation
- Automatically translate knowledge services to data services

Ontology-based data management: architecture



Based on three main components:

- **Ontology**, used as the conceptual layer to give clients a unified conceptual specification of the domain.
- **Data sources**, representing external, independent, heterogeneous, storage (or, more generally, computational) structures.
- **Mappings**, used to semantically link data at the sources to the ontology.

Which languages?

- Which **language** for the ontology?
- Which **language** for the mappings?
- Which **language** for expressing services (i.e., queries) over the ontology?

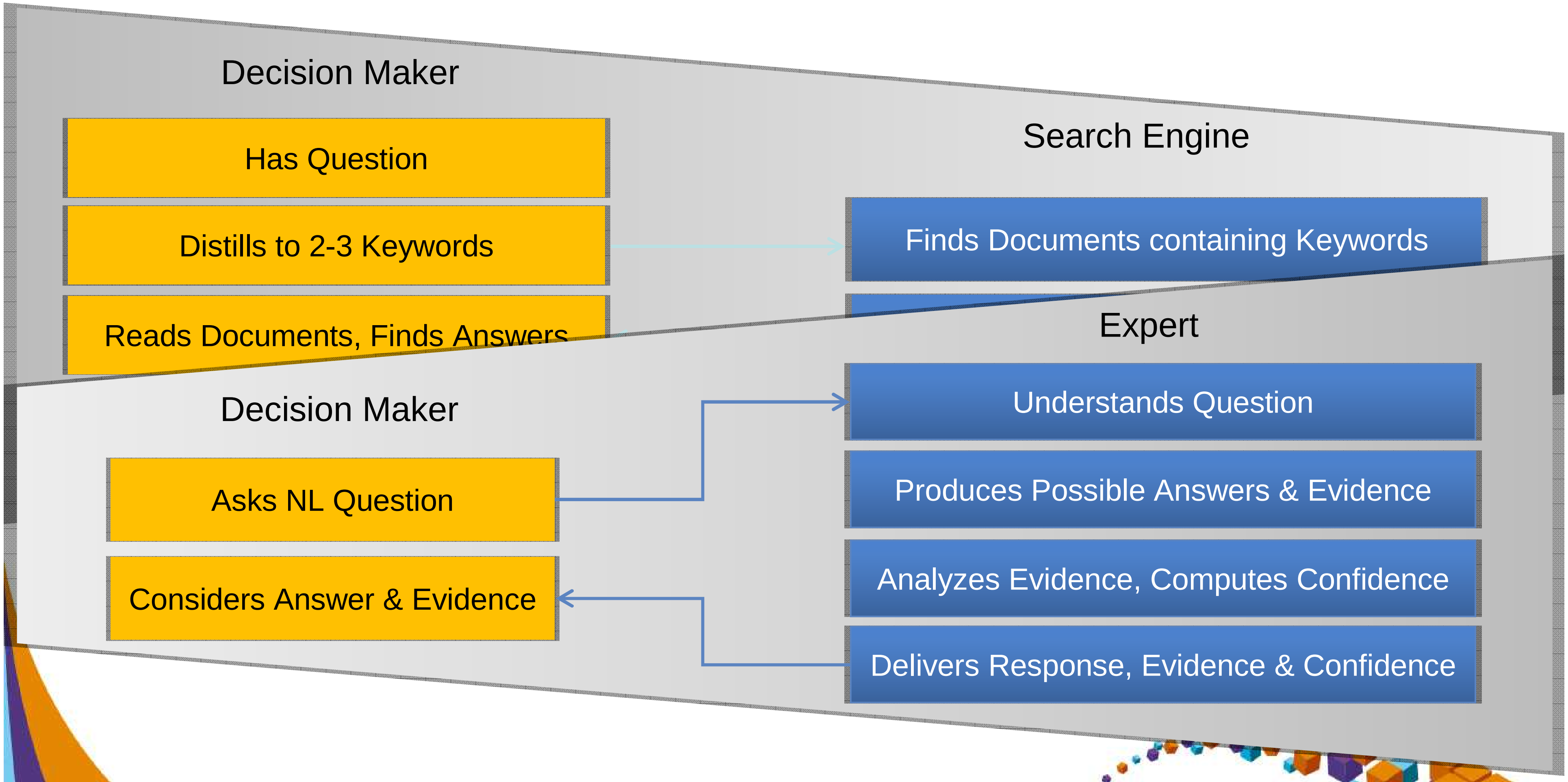
Challenge: optimal compromise between expressive power and data complexity.

InfoPlanet

InfoPlanet

Governance per i tuoi dati,
opportunità per il tuo business.
Con IBM.

Informed Decision Making: Search vs. Expert Q&A



Decision Maker

Has Question

Distills to 2-3 Keywords

Reads Documents, Finds Answers

Search Engine

Finds Documents containing Keywords

Decision Maker

Asks NL Question

Considers Answer & Evidence

Expert

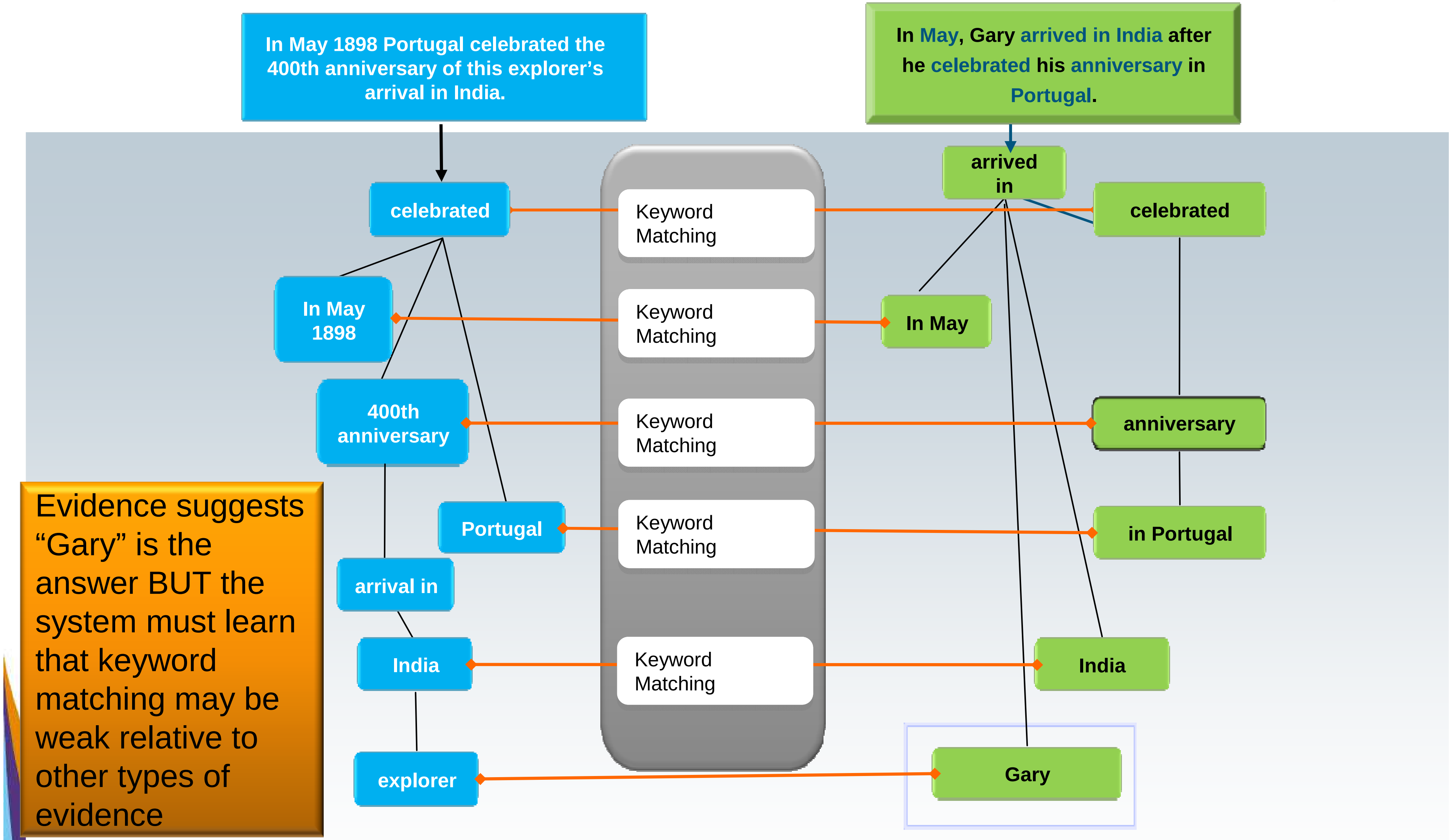
Understands Question

Produces Possible Answers & Evidence

Analyzes Evidence, Computes Confidence

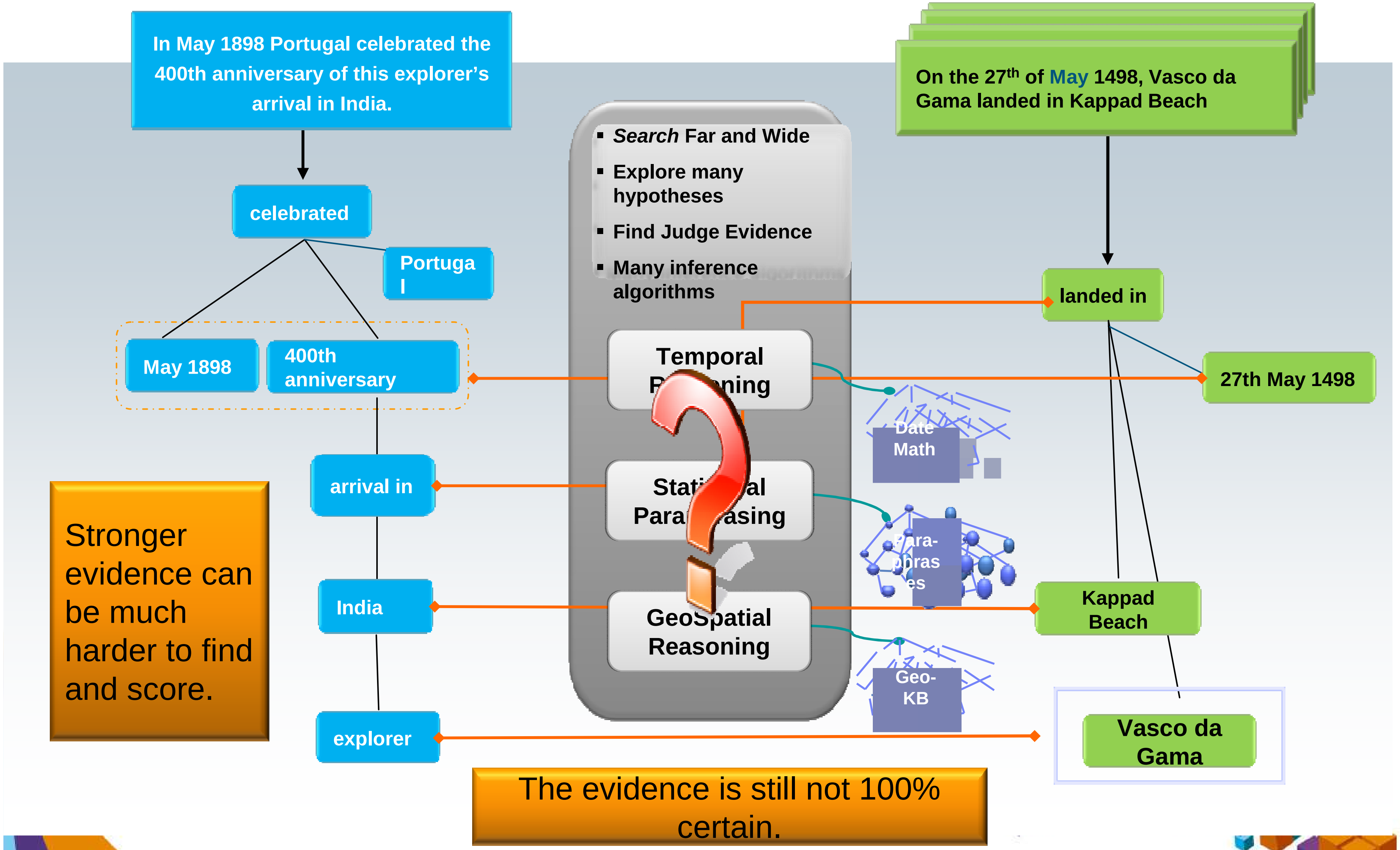
Delivers Response, Evidence & Confidence

Different Types of Evidence: Keyword Evidence



Evidence suggests "Gary" is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence

Different Types of Evidence: Deeper Evidence



In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

celebrated

Portugal

May 1898

400th anniversary

arrival in

India

explorer

- *Search* Far and Wide
- Explore many hypotheses
- Find Judge Evidence
- Many inference algorithms

Temporal R ning

Stati al Para asing

GeoSpatial Reasoning

Date Math

Para-phrases

Geo-KB

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

landed in

27th May 1498

Kappad Beach

Vasco da Gama

Stronger evidence can be much harder to find and score.

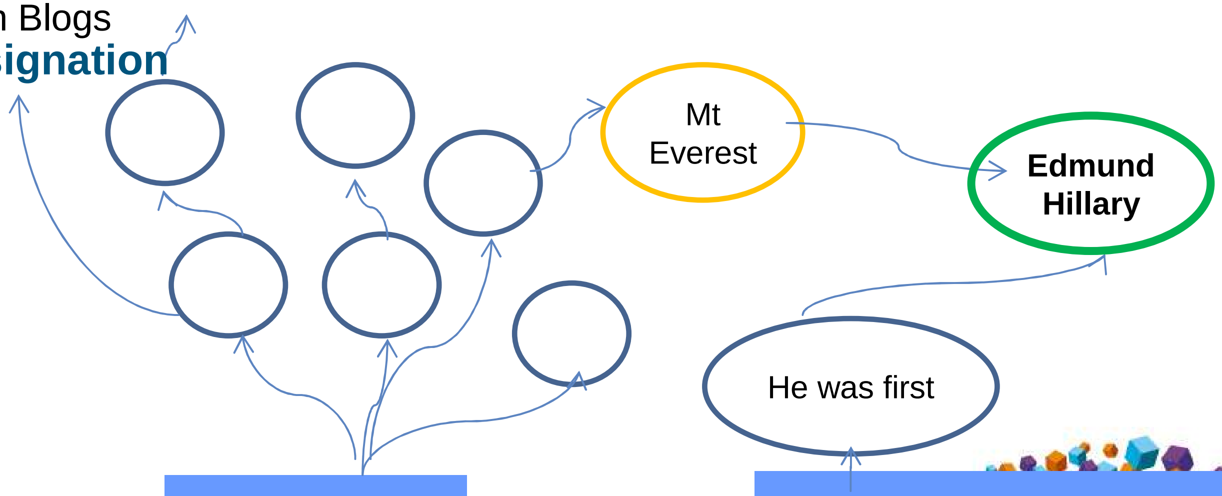
The evidence is still not 100% certain.

Examples from Jeopardy! clues and missing links

- This **fish** was thought to be extinct millions of years ago until one was found off South Africa in 1938
- Category: ENDS IN "TH"
- Answer: **coelacanth**

- When hit by electrons, a phosphor gives off electromagnetic energy in this **form**
- Category: General Science
- Answer: **light (or photons)**

- Secy. Chase just submitted **this** to me for the third time--guess what, pal. This time I'm accepting **it**
- Category: Lincoln Blogs
- Answer: **his resignation**



On hearing of the discovery of George Mallory's body, he told reporters he still thinks he was first.

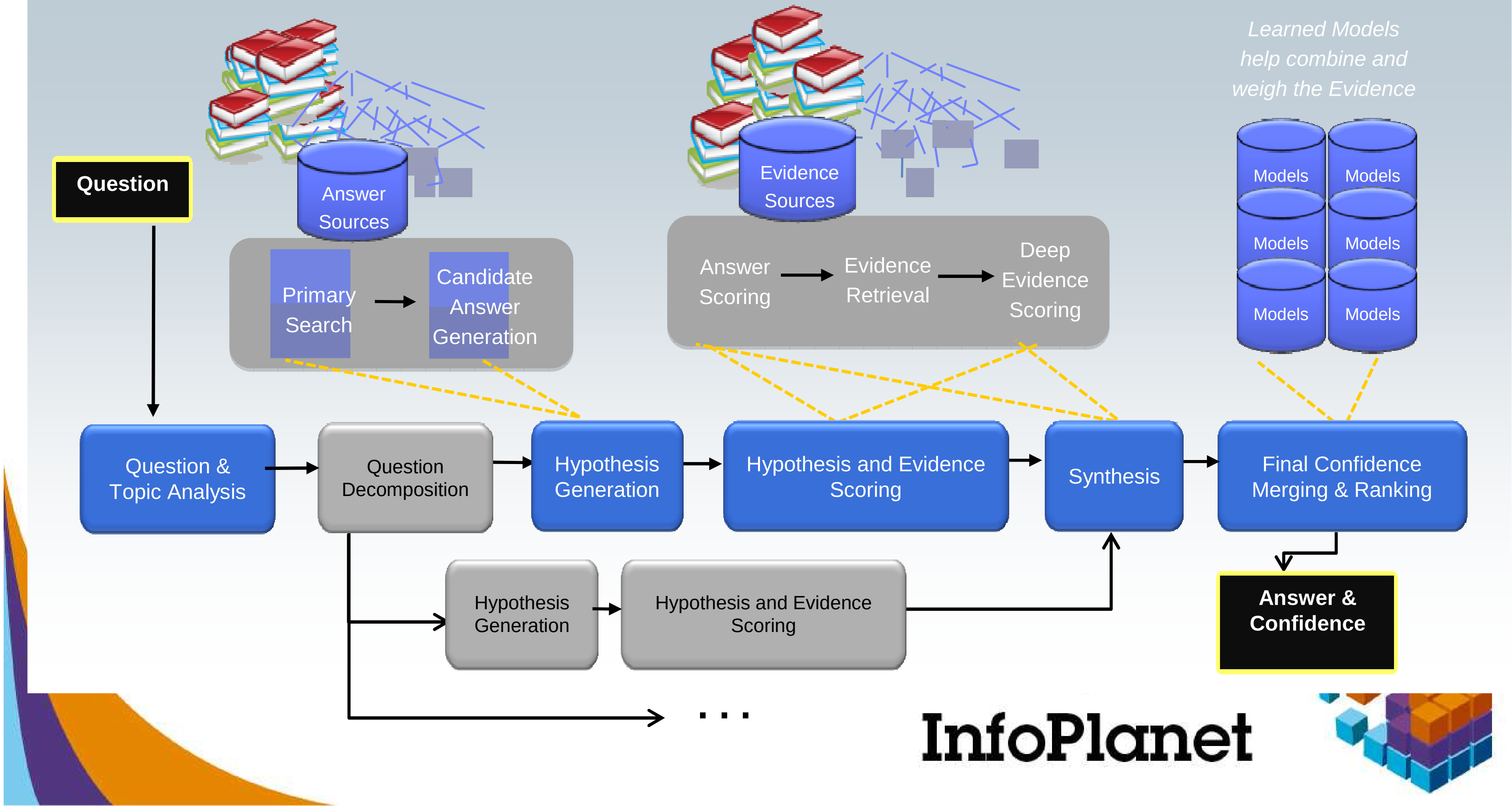
InfoPlanet

Governance per i tuoi dati,
opportunità per il tuo business.
Con IBM.

DeepQA: The Technology Behind Watson

Massively Parallel Probabilistic Evidence-Based Architecture

DeepQA *generates and scores many hypotheses using an extensible collection of* ***Natural Language Processing**, **Machine Learning** and **Reasoning Algorithms.*** *These gather and weigh evidence over both unstructured and structured content to determine the answer with the best confidence.*



InfoPlanet

InfoPlanet

Governance per i tuoi dati,
opportunità per il tuo business.
Con IBM.

Semantic technologies for Data Management

Based on the idea that the ontology is the heart of the information system.

- *Ontology-based data access and integration*
- *Ontology-based privacy-aware data access*
- *Ontology-based data quality*
- *Ontology-based data restructuring*
- *Ontology-based data update*
- *Ontology-based service management*

General requirements:

- large data collections
- efficiency at least with respect to size of data (data complexity)

Ontology-based data access and integration

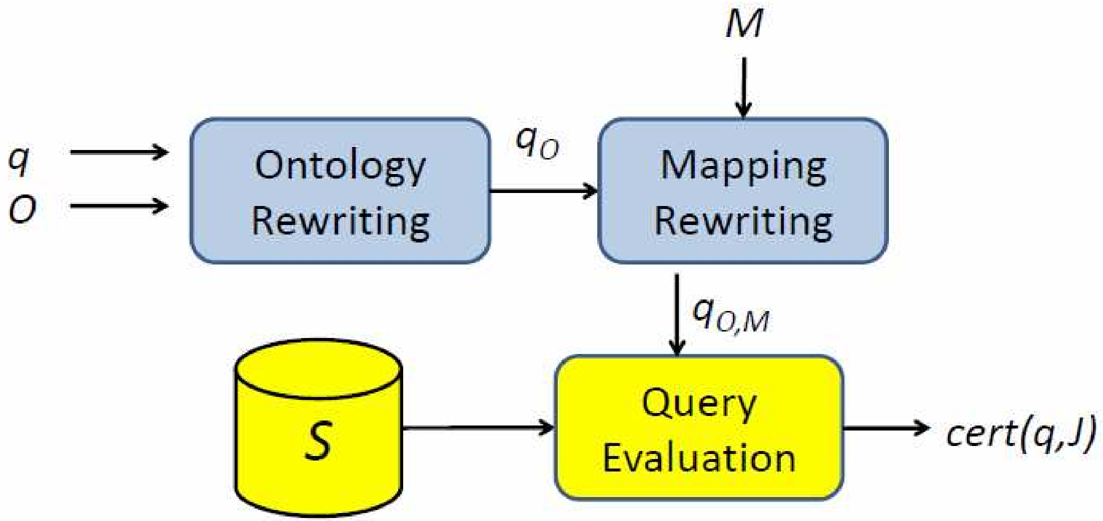
- Which language for the ontology?
  - *$DL\text{-}Lite_{A,id}$*

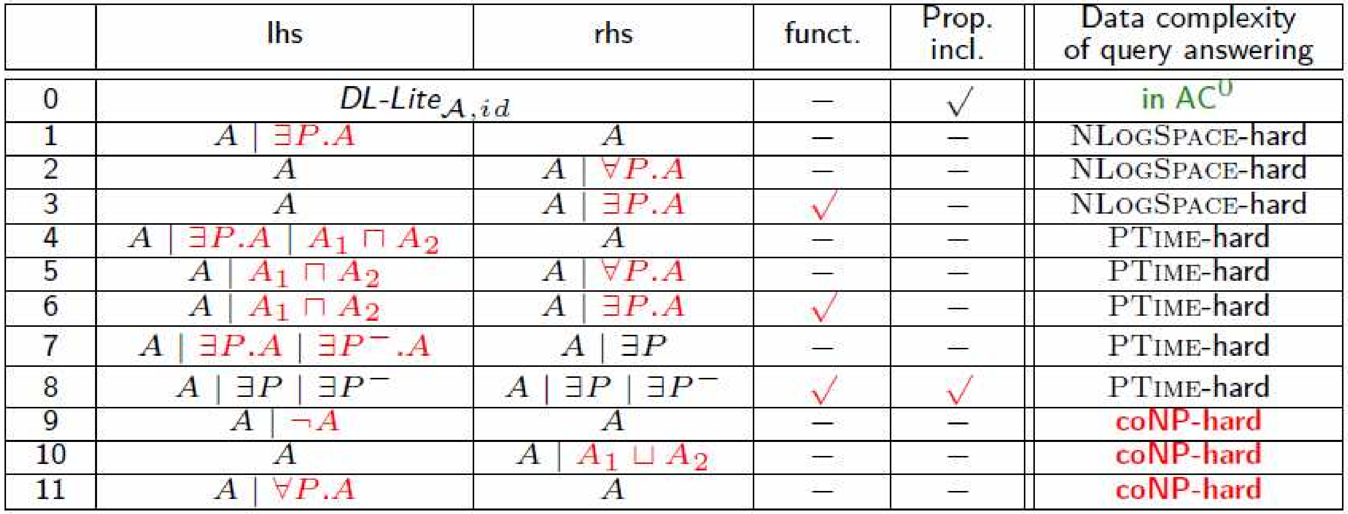
- Which language for the mappings?
  - *FOL-to-CQ, with object constructors*

- Which language for expressing queries over the ontology?
  - *Essentially UCQs*

Challenge: optimal compromise between expressive power and data complexity.

Ontology-based data access and integration





	lhs	rhs	funct.	Prop. incl.	Data complexity of query answering
0	*DL-Lite*$_{\mathcal{A},id}$		–	√	in AC^0
1	$A \mid \exists P.A$	A	–	–	NLOGSPACE-hard
2	A	$A \mid \forall P.A$	–	–	NLOGSPACE-hard
3	A	$A \mid \exists P.A$	√	–	NLOGSPACE-hard
4	$A \mid \exists P.A \mid A_1 \sqcap A_2$	A	–	–	PTIME-hard
5	$A \mid A_1 \sqcap A_2$	$A \mid \forall P.A$	–	–	PTIME-hard
6	$A \mid A_1 \sqcap A_2$	$A \mid \exists P.A$	√	–	PTIME-hard
7	$A \mid \exists P.A \mid \exists P^-.A$	$A \mid \exists P$	–	–	PTIME-hard
8	$A \mid \exists P \mid \exists P^-$	$A \mid \exists P \mid \exists P^-$	√	√	PTIME-hard
9	$A \mid \neg A$	A	–	–	coNP-hard
10	A	$A \mid A_1 \sqcup A_2$	–	–	coNP-hard
11	$A \mid \forall P.A$	A	–	–	coNP-hard

Ontology-based privacy-aware data access

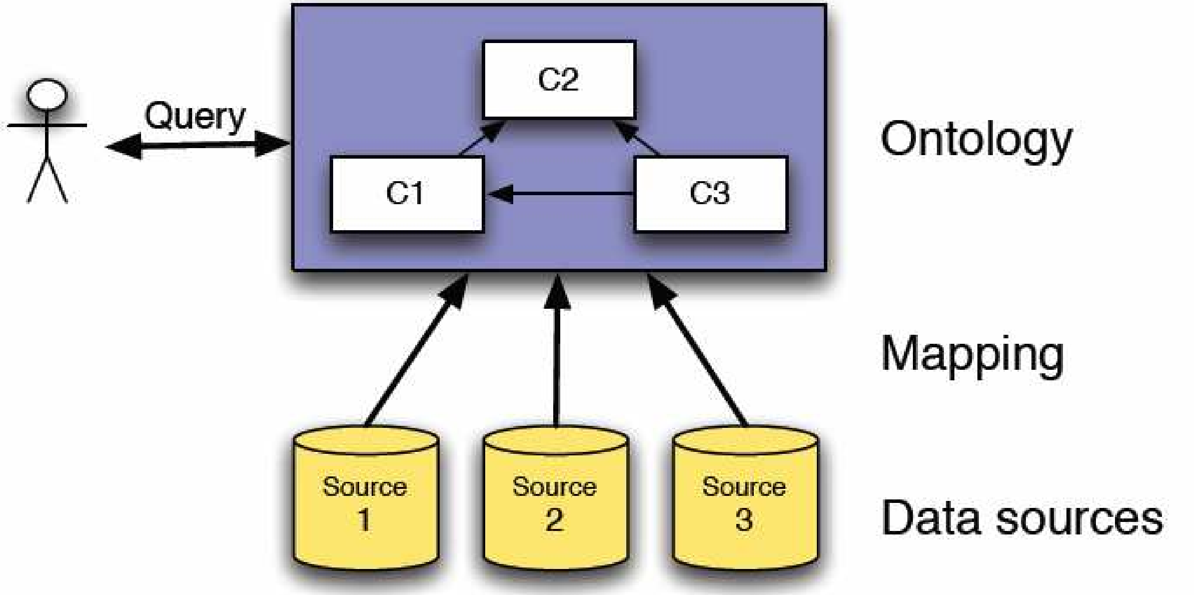
- What can be seen by a user can be formalized by means of a set of views (called authorization views) over the ontology
- The query answering algorithm can ensure that the answer returned to the user can be derived only by the knowledge represented by the authorization views

Ontology-based data quality

- Checking the quality of the data sources can be done by comparing the information content of the sources with the ontology
- The quality of query answering can be improved by using logic-based techniques for “repairing” inconsistencies

Ontology-based data restructuring

We can restructure our data by materializing the data according to the ontology



Ontology-based data update

- The idea is that users can express, besides queries, updates over the ontology

- Challenges:
  - What is the semantics of an update expressed over the ontology?
  - How to push the updates from the ontology to the data sources?

InfoPlanet

Ontology-based service management

- The idea is that one can express, besides queries and updates, services over the ontology

- Challenges:
  - What is the right language to express services?
  - How to compare services?
  - How to automatically compose services to dynamically devise new services the updates from the ontology to the data sources?

InfoPlanet

Governance per i tuoi dati,
opportunità per il tuo business.
Con IBM.

Potential Business Applications

Healthcare / Life Sciences: Diagnostic Assistance, Evidence-Based, Collaborative Medicine

Tech Support: Help-desk, Contact Centers

Enterprise Knowledge Management and Business Intelligence

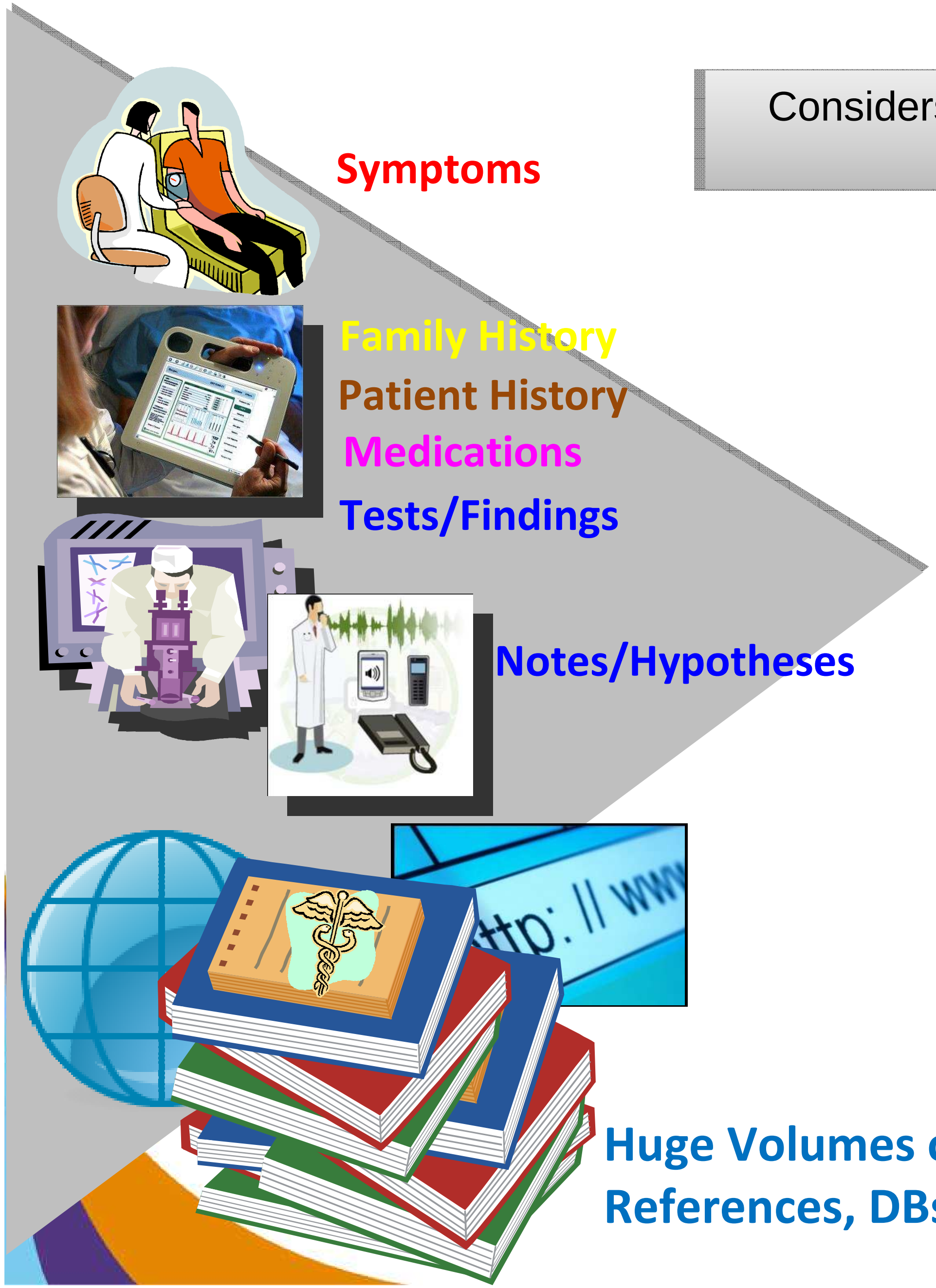
Government: Improved Information Sharing and Education

InfoPlanet

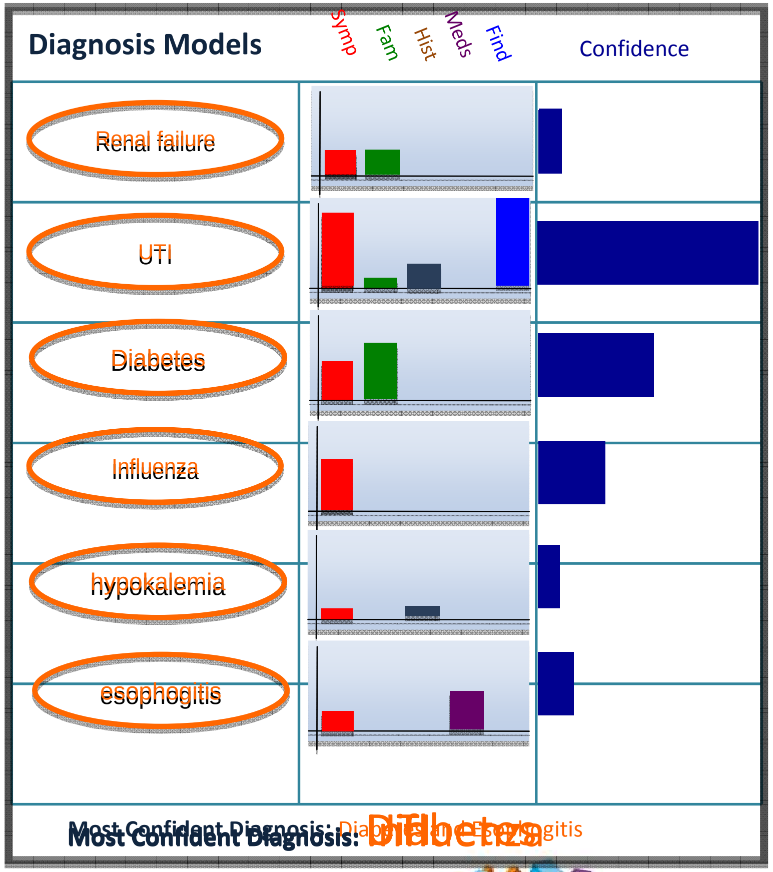
Differential Diagnosis with DeepQA

- **Capabilities**
  - **Support physicians in the differential diagnosis process**
  - **Address best known sources of diagnostic errors**
  - **Deal with ambiguous, incomplete, conflicting information** (both in declared symptoms, observations, findings, …and in the knowledge sources)
  - **Leverage both structured** (e.g. lab tests, EMR, ontologies) **and unstructured** (e.g. reports, papers, knowledge bases) **data**
  - Perform **statistical analysis of multiple** partially overlapping **unstructured evidences**
  - Help **identify "red herrings"** (anomalies in patient history data (e.g. incorrect lab tests results) that may lead to incorrect conclusions)
  - **Point to missing** information that would help reduce ambiguity and improve the quality of the diagnosis
  - **Real-time response**, except for periodic pre-processing of data sources when updates are made available

DeepQA in Continuous Evidence-Based Diagnostic Analysis



Considers and synthesizes a broad range of evidence improving quality, reducing cost



Huge Volumes of Texts, Journals, References, DBs etc.

InfoPlanet

InfoPlanet

Governance per i tuoi dati,
opportunità per il tuo business.
Con IBM.