

Probabilistic Databases: Introduction

Antoine Amarilli



Uncertain data: Practical motivations




Numerous sources of **uncertain data**:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemata
- Imprecise automated processes (information extraction, NLP, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

Use case: Web information extraction

Recently-Learned Facts

Refresh

instance	iteration	date learned	confidence
oliguric_phase is a non-disease physiological condition	1111	06-jul-2018	97.5  
alaska airlines is an organization	1114	25-aug-2018	100.0  
heating_insurance_policies is a physical action	1111	06-jul-2018	90.4  
n98_12 is a term used by physicists	1111	06-jul-2018	94.2  
dragonball_z_super_butoden_2 is software	1111	06-jul-2018	100.0  
general_motors_corp is a company headquartered in the city detroit	1116	12-sep-2018	100.0  
the companies herald and la compete with eachother	1111	06-jul-2018	99.6  
stanford hired montgomery	1111	06-jul-2018	98.4  
kimn is a radio station in the city denver	1116	12-sep-2018	100.0  
radisson_sas_portman_hotel is a park in the city central_london	1116	12-sep-2018	100.0  

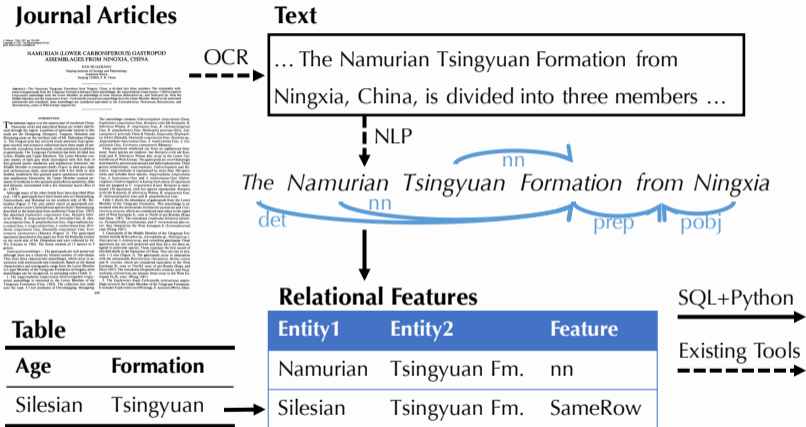
Never-ending Language Learning (NELL, CMU), <http://rtw.ml.cmu.edu/rtw/kbbrowser/>

Use case: Web information extraction

Subject	Predicate	Object	Confidence
Elvis Presley	diedOnDate	1977-08-16	97.91%
Elvis Presley	isMarriedTo	Priscilla Presley	97.29%
Elvis Presley	influences	Carlo Wolff	96.25%

YAGO, <https://www.yago-knowledge.org/>

Other use case: Information extraction from scientific articles



Other use case: Crowdsourcing

All HITs

1-10 of 2751 Results

Sort by:



[Show all details](#)

[Hide all details](#)

1

[2](#)

[3](#)

[4](#)

[5](#)

>

[Next](#)

>>

[Last](#)

Transcribe data

[View a HIT in this group](#)

Requester: p9r

HIT Expiration Date: Nov 18, 2015 (23 hours 59 minutes)

Reward: \$0.03

Time Allotted: 45 minutes

Description: Please transcribe the data from the following images

Keywords: [transcribe](#), [handwriting](#), [data entry](#)

Qualifications Required:

HIT approval rate (%) is greater than 90

Classify Receipt

[View a HIT in this group](#)

Requester: Jon Breliq

HIT Expiration Date: Nov 24, 2015 (6 days 23 hours)

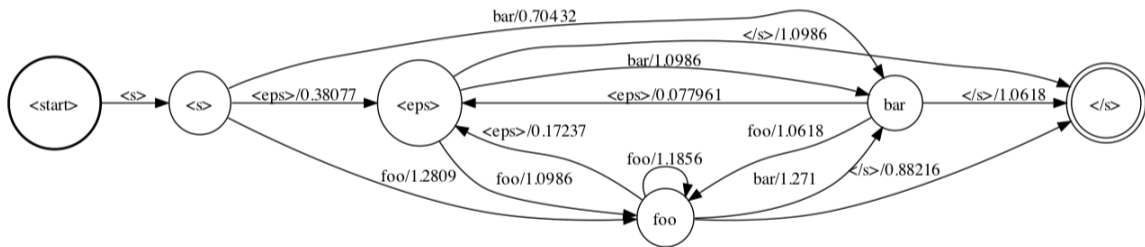
Reward: \$0.02

Time Allotted: 20 minutes

Description: Looking at a receipt image, identify the business of the receipt

Keywords: [image](#), [receipt](#), [categorize](#), [transcribe](#), [extract](#), [data](#), [entry](#), [transcription](#), [text](#), [easy](#), [qualification](#), [jon](#), [breliq](#), [prod](#)

Other use case: Speech recognition and OCR



Different types of uncertainty

- The uncertainty can be **qualitative** (e.g., NULL)...
- ... or **quantitative** (e.g., 95%)

Further, there are different types:

- **Unknown** value: NULL in an RDBMS
- **Alternative** between several possibilities: either A or B or C
- **Imprecision on a numeric value**: a sensor gives a value that is an approximation of the actual value
- **Confidence in a fact as a whole**: cf. information extraction
- **Structural uncertainty**: the schema of the data itself is uncertain
- **Missing data**: we know that some data is missing (open-world semantics)

What happens to this uncertainty?

Naive solution

Forget about uncertainty, or apply a threshold after each computation step

What happens to this uncertainty?

Naive solution

Forget about uncertainty, or apply a threshold after each computation step

Ideal solution

Instead of neglecting uncertainty, let's manage it rigorously throughout the whole process of answering a query

What happens to this uncertainty?

Naive solution

Forget about uncertainty, or apply a threshold after each computation step

Ideal solution

Instead of neglecting uncertainty, let's manage it rigorously throughout the whole process of answering a query

Also: it leads to interesting theoretical questions! :)

Possible worlds semantics

Idea: use a representation system

Possible world: A **regular** (deterministic) relational database

Possible worlds semantics

Idea: use a representation system

Possible world: A **regular** (deterministic) relational database

Uncertain database: (Compact) representation of a **set of possible worlds**

Possible worlds semantics

Idea: use a representation system

Possible world: A **regular** (deterministic) relational database

Uncertain database: (Compact) representation of a **set of possible worlds**

Probabilistic database: (Compact) representation of a **probability distribution over possible worlds,**

Possible worlds semantics

Idea: use a representation system

Possible world: A **regular** (deterministic) relational database

Uncertain database: (Compact) representation of a **set of possible worlds**

Probabilistic database: (Compact) representation of a **probability distribution over possible worlds**, either:

finite: a set of possible worlds, each with their probability

continuous: more complicated

Possible worlds semantics

Idea: use a representation system

Possible world: A **regular** (deterministic) relational database

Uncertain database: (Compact) representation of a **set of possible worlds**

Probabilistic database: (Compact) representation of a **probability distribution over possible worlds**, either:

finite: a set of possible worlds, each with their probability

continuous: more complicated

date	teacher	
08	Diego	0.9
09	Paolo	0.8
09	Floris	0.7

Contents of this class

- Present the most common **models** of probabilistic data
 - Focus on the **simplest one**, tuple-independent databases (TID)

Contents of this class

- Present the most common **models** of probabilistic data
 - Focus on the **simplest one**, tuple-independent databases (TID)
- Introduce the **probabilistic query evaluation** problem (PQE):
 - Central task: evaluating queries over probabilistic databases

Contents of this class

- Present the most common **models** of probabilistic data
 - Focus on the **simplest one**, tuple-independent databases (TID)
- Introduce the **probabilistic query evaluation** problem (PQE):
 - Central task: evaluating queries over probabilistic databases
- Present the **dichotomy** by Dalvi and Suciu on the complexity of PQE for UCQs

Contents of this class

- Present the most common **models** of probabilistic data
 - Focus on the **simplest one**, tuple-independent databases (TID)
- Introduce the **probabilistic query evaluation** problem (PQE):
 - Central task: evaluating queries over probabilistic databases
- Present the **dichotomy** by Dalvi and Suciu on the complexity of PQE for UCQs
- Present **treewidth-based approaches** to efficient PQE

Contents of this class

- Present the most common **models** of probabilistic data
 - Focus on the **simplest one**, tuple-independent databases (TID)
- Introduce the **probabilistic query evaluation** problem (PQE):
 - Central task: evaluating queries over probabilistic databases
- Present the **dichotomy** by Dalvi and Suciu on the complexity of PQE for UCQs
- Present **treewidth-based approaches** to efficient PQE
- Give an overview of **other topics** on probabilistic databases

Contents of this class

- Present the most common **models** of probabilistic data
 - Focus on the **simplest one**, tuple-independent databases (TID)
- Introduce the **probabilistic query evaluation** problem (PQE):
 - Central task: evaluating queries over probabilistic databases
- Present the **dichotomy** by Dalvi and Suciu on the complexity of PQE for UCQs
- Present **treewidth-based approaches** to efficient PQE
- Give an overview of **other topics** on probabilistic databases
- **Next class (Jan 9)**: Liat will talk about **incomplete information**