A COURSE ON PROBABILISTIC DATABASES

Probabilistic Databases

- Data: standard relational data, plus probabilities that measure the degree of uncertainty

- Queries: standard SQL queries, whose answers are annotated with output probabilities

A Little History of Probabilistic DBs

Early days

- Wong'82
- Shoshani'82
- Cavallo&Pittarelli'87
- Barbara'92
- Lakshmanan'97,'01
- Fuhr&Roellke'97
- Zimanyi'97

Main challenge:
Query Evaluation
(=Probabilistic Inference)

Recent work

- Stanford (Trio)
- UW (MystiQ)
- Cornell (MayBMS)
- Oxford (MayBMS)
- U.of Maryland
- IBM Almaden (MCDB)
- Rice (MCDB)
- U. of Waterloo
- UBC
- U. of Florida
- Purdue University
- U. of Wisconsin

Why?

Many applications need to manage uncertain data

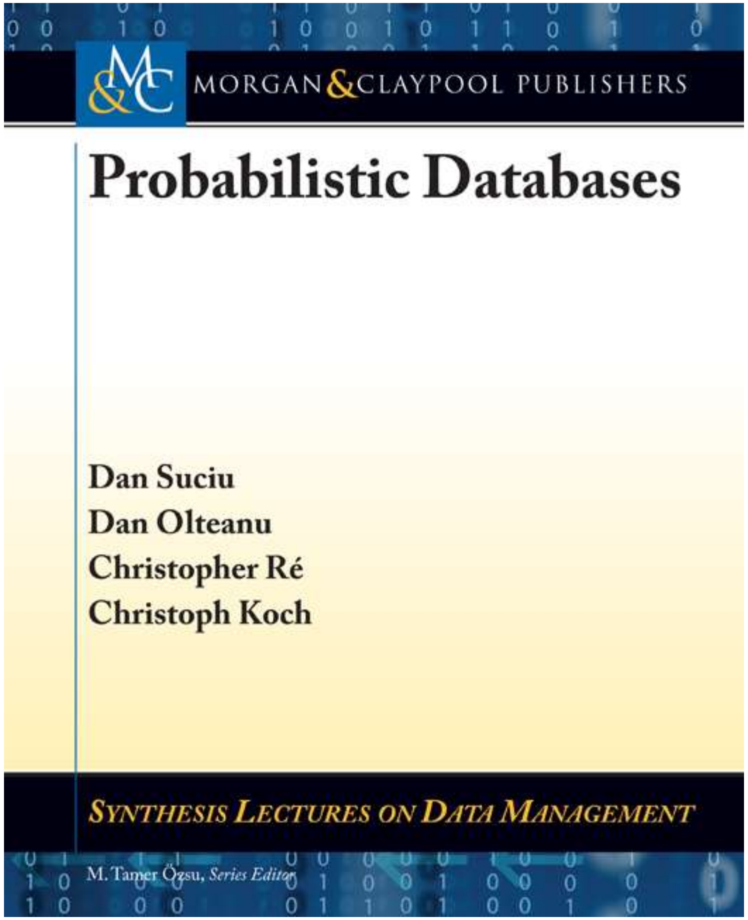
- Information extraction
- Knowledge representation
- Fuzzy matching
- Business intelligence
- Data integration
- Scientific data management
- Data anonymization

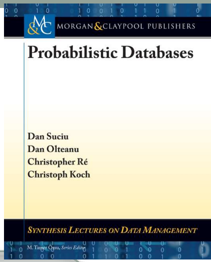
What?

- Probabilistic Databases extend Relational Databases with probabilities
- Combine Formal Logic with Probabilistic Inference
- Requires a new thinking for both databases and probabilistic inference

This Course: Query Evaluation

Based on the book:

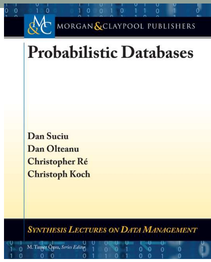




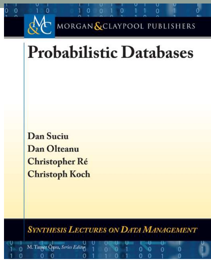
Outline of the Tutorial

Part 1	1.	Motivating Applications	
	2.	The Probabilistic Data Model	Chapter 2
Part 2	3.	Extensional Query Plans	Chapter 4.2
	4.	The Complexity of Query Evaluation	Chapter 3
Part 3	5.	Extensional Evaluation	Chapter 4.1
Part 4	6.	Intensional Evaluation	Chapter 5
	7.	Conclusions	

What You Will Learn



- Background:
  - Relational data model: tables, queries, relational algebra
  - PTIME, NP, #P
  - Model counting: DPLL, OBDD, FBDD, d-DNNF
- In detail:
  - Extensional plans, extensional evaluation, running them in postgres
  - The landscape of query complexity: from PTIME to #P-complete,
  - Query compilation: Read-Once Formulas, OBDD, FBDD, d-DNNF
- Less detail:
  - The #P-hardness proof, complexity of BDDs
- Omitted:
  - Richer data models: BID, GM, XML, continuous random values)
  - Approximate query evaluation,
  - Ranking query answers



Related Work. See book, plus:

These references are not in the book

- Wegener: Branching programs and binary decision diagrams: theory and applications, 2000
- Dalvi, S.: The dichotomy of probabilistic inference for unions of conjunctive queries, JACM'2012
- Huang, Darwiche: DPLL with a Trace: From SAT to Knowledge Compilation, IJCAI 2005
- Beame, Li, Roy, S.: Lower Bounds for Exact Model Counting and Applications in Probabilistic Databases, UAI'13
- Gatterbauer, S.: Oblivious Bounds on the Probability of Boolean Functions, under review

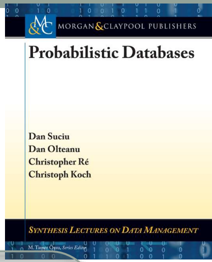
The applications are from:

- Ré, Letchner, Balazinska, S: Event queries on correlated probabilistic streams. SIGMOD Conference 2008
- Gupta, Sarawagi: Creating Probabilistic Databases from Information Extraction Models. VLDB 2006
- Stoyanovich, Davidson, Milo, Tannen: Deriving probabilistic databases with inference ensembles. ICDE 2011
- Beskales, Soliman, Ilyas, Ben-David: Modeling and Querying Possible Repairs in Duplicate Detection. PVLDB 2009
- Kumar, Ré: Probabilistic Management of OCR Data using an RDBMS. PVLDB 2011

A COURSE ON PROBABILISTIC DATABASES

Lecture 1: Motivating Applications

Outline



Part 1

1. Motivating Applications
2. The Probabilistic Data Model — Chapter 2

Part 2

3. Extensional Query Plans — Chapter 4.2
4. The Complexity of Query Evaluation — Chapter 3

Part 3

5. Extensional Evaluation — Chapter 4.1

Part 4

6. Intensional Evaluation — Chapter 5
7. Conclusions

[Gupta'2006]

Example 1: Information Extraction

```
52-A Goregaon West Mumbai 400 076
```

CRF

Standard DB: keep the most likely extraction

Id	House_no	Area	City	Pincode	Prob
1	52	Goregaon West	Mumbai	400 062	0.1
1	52-A	Goregaon	West Mumbai	400 062	0.2
1	52-A	Goregaon West	Mumbai	400 062	0.5
1	52	Goregaon	West Mumbai	400 062	0.2

Probabilistic DB: keep most/all extractions to increase recall

Key finding: the probabilities given by CRFs correlate well with the precision of the extraction.

[Stoyanovich’2011]

Example 2: Modeling Missing Data

id	age	edu	inc	nw
t_1	20	HS	?	?
t_2	20	BS	50K	100K
t_3	20	?	50K	?
t_4	20	HS	100K	500K
t_5	20	?	?	?
t_6	20	HS	50K	100K
t_7	20	HS	50K	500K
t_8	?	HS	?	?
t_9	30	BS	100K	100K
t_{10}	30	?	100K	?
t_{11}	30	HS	?	?
t_{12}	30	MS	?	?
t_{13}	40	BS	100K	100K
t_{14}	40	HS	?	?
t_{15}	40	BS	50K	500K
t_{16}	40	HS	?	500K
t_{17}	40	HS	100K	500K

Standard DB: NULL

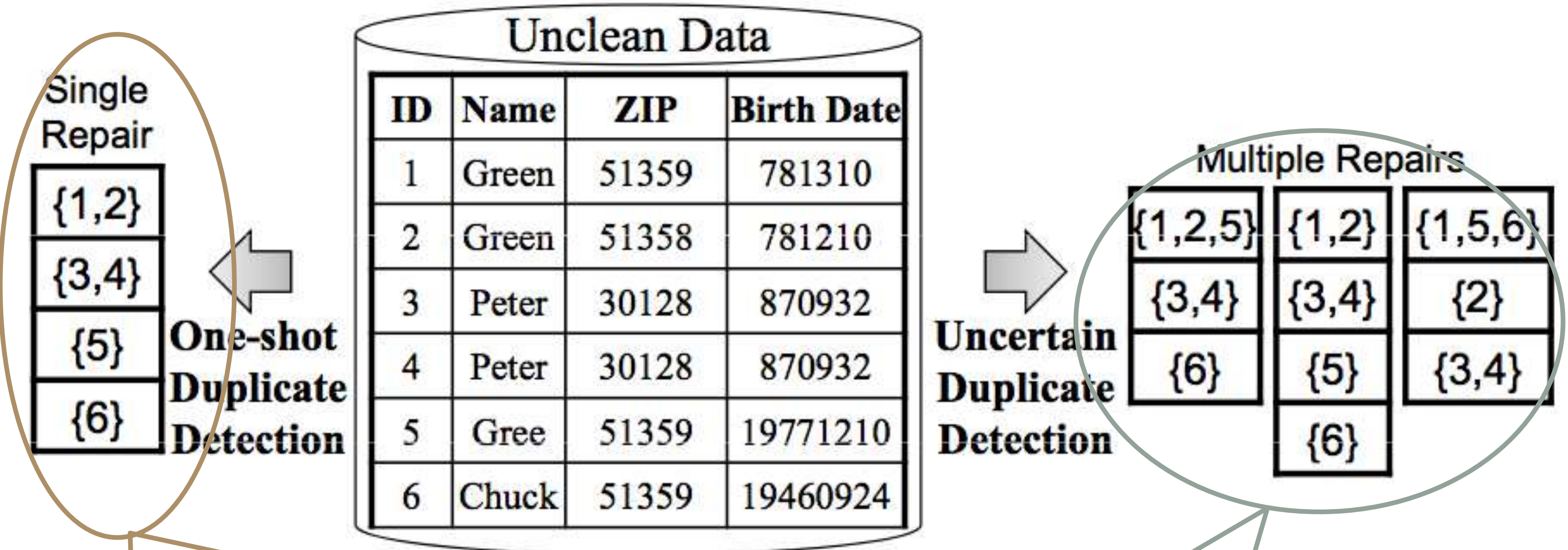
Probabilistic DB: distribution on possible values

id	age	edu	inc	nw	prob
$t_{12}.1$	30	MS	50K	100K	0.30
$t_{12}.2$	30	MS	50K	500K	0.45
$t_{12}.3$	30	MS	100K	100K	0.10
$t_{12}.4$	30	MS	100K	500K	0.15

Key technique: Meta Rule Semi-Lattice for inferring missing attributes.

[Beskales'2009]

Example 3: Data Cleaning



Unclean Data

ID	Name	ZIP	Birth Date
1	Green	51359	781310
2	Green	51358	781210
3	Peter	30128	870932
4	Peter	30128	870932
5	Gree	51359	19771210
6	Chuck	51359	19460924

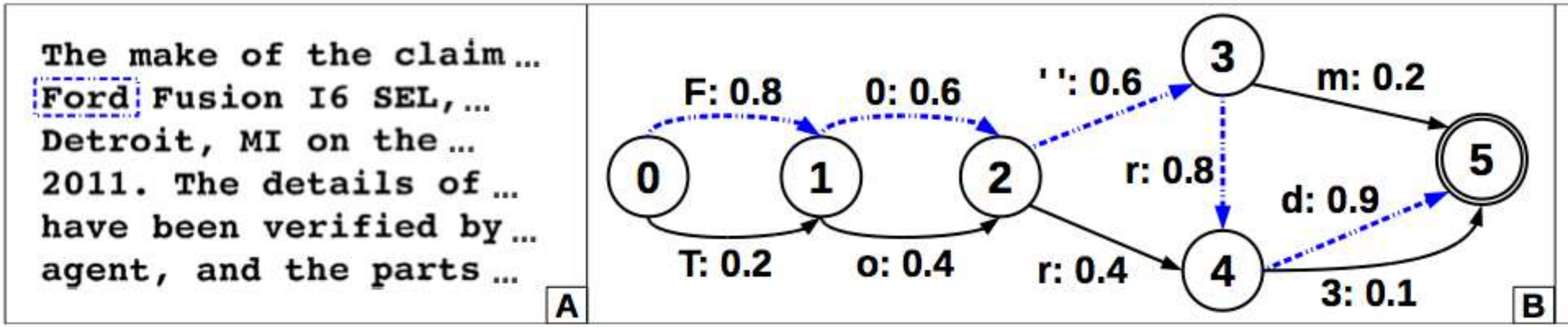
Standard DB
cleaning data means choosing one possible repair

Probabilistic DB
keep many/all possible repairs

Challenge: Representing multiple repairs.
[Beskaes'2009] restrict to hierarchical repairs.

[Kumar'2011]

Example 4: OCR



They use OCRopus from Google Books: output is a stohastic automaton
Traditionally: retain only the Maximum Apriori Estimate (MAP)
With a probabilistic database: may retain several alternative recognitions: increase recal

```
SELECT DocId, Loss
FROM Claims
WHERE Year = 2010
    AND DocData LIKE '%Ford%';
```

Summary of Applications

- Structured, but uncertain data
- Modeled as probabilistic data
- Answers to SQL queries annotated with probabilities

Probabilistic database:

- Combine data management with probabilistic inference