Representation Formalisms for Uncertain Data

Jennifer Widom

with

Anish Das Sarma
Omar Benjelloun
Alon Halevy

and other participants in the Trio Project

!! Warning !!

This work is preliminary and in flux

Ditto for these slides

Luckily I'm among friends…

Some Context

Trio Project

We're building a new kind of DBMS in which:

1. Data
2. Accuracy
3. Lineage

are all first-class interrelated concepts

Potential applications

– Scientific and sensor databases
– Data cleaning and integration
– Approximate query processing
– And others…

Context (cont'd)

We began by investigating the accuracy component

= uncertainty (more on terminology coming)

Recently we've made progress tying together uncertainty + lineage

Approaching a New Problem

1) Work in a void for a while
2) Then see what others have done
3) Adjust and proceed

Void Part 1

Defined initial Trio Data Model (TDM) [CIDR ’05]

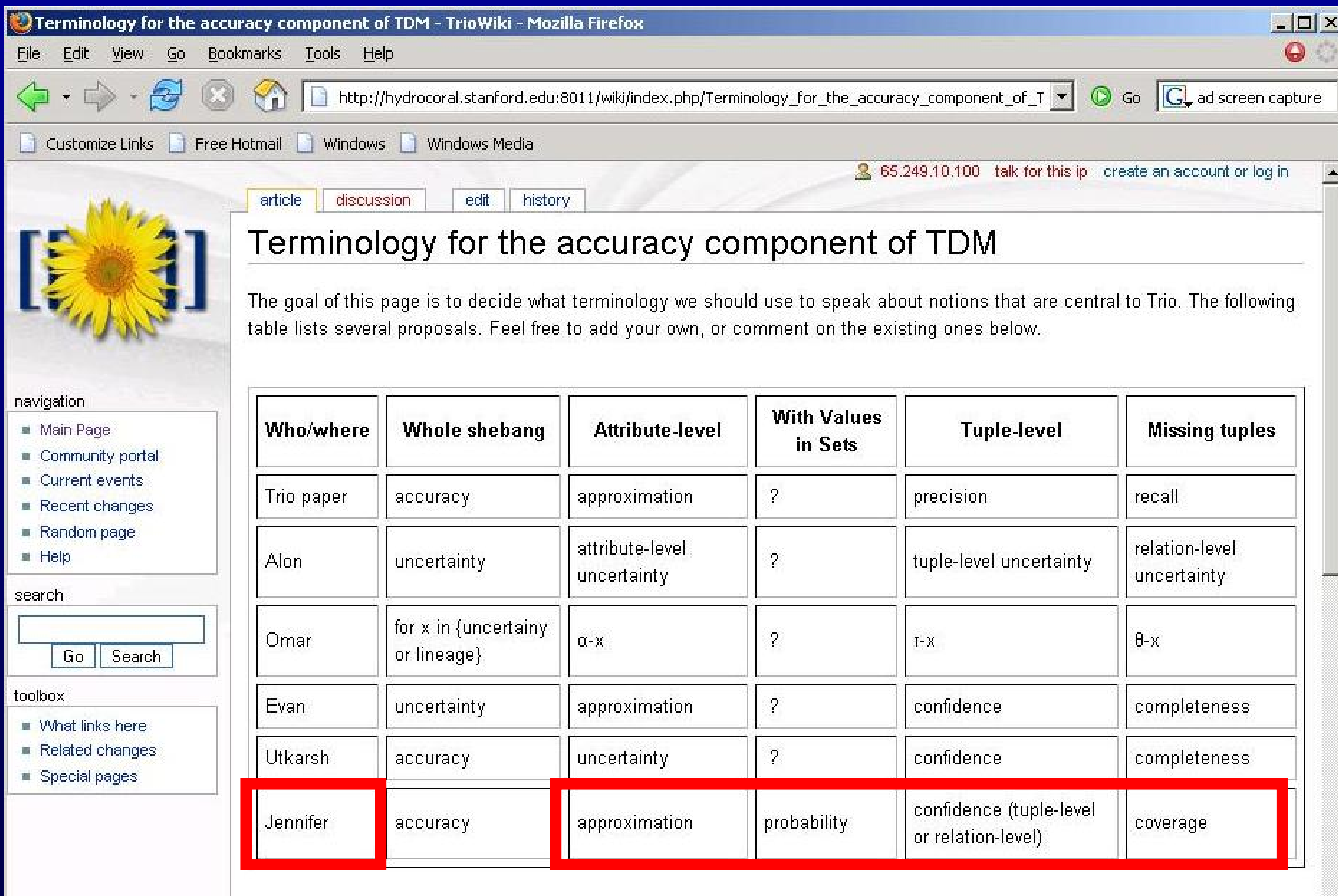
- Based primarily on applications and intuition

Accuracy component of initial TDM

A sub-trio:

1. Attribute-level approximation
2. Tuple-level (or relation-level) confidence
3. Relation-level coverage

Terminology Wars



Terminology for the accuracy component of TDM - TrioWiki - Mozilla Firefox

http://hydrocoral.stanford.edu:8011/wiki/index.php/Terminology_for_the_accuracy_component_of_T

article | discussion | edit | history

Terminology for the accuracy component of TDM

The goal of this page is to decide what terminology we should use to speak about notions that are central to Trio. The following table lists several proposals. Feel free to add your own, or comment on the existing ones below.

Who/where	Whole shebang	Attribute-level	With Values in Sets	Tuple-level	Missing tuples
Trio paper	accuracy	approximation	?	precision	recall
Alon	uncertainty	attribute-level uncertainty	?	tuple-level uncertainty	relation-level uncertainty
Omar	for x in {uncertainy or lineage}	α-x	?	τ-x	θ-x
Evan	uncertainty	approximation	?	confidence	completeness
Utkarsh	accuracy	uncertainty	?	confidence	completeness
Jennifer	accuracy	approximation	probability	confidence (tuple-level or relation-level)	coverage

TDM: Approximation (Attributes)

Broadly, an approximate value is a set of possible values along with a probability distribution over them

TDM: Approximation (Attributes)

Specifically, each Trio attribute value is either:

1) **Exact value (default)**
2) **Set of values, each with *prob* $\in$ [0,1] (Σ=1)**
3) **Min + Max for a range (uniform distribution)**
4) **Mean + Deviation for Gaussian distribution**

Type 2 sets may include "unknown" ($\perp$)

Independence of approximate values within a tuple

TDM: Confidence (Tuples)

Each tuple **t** has confidence $\in [0,1]$

- Informally: chance of **t** correctly belonging in relation
- Default: **confidence=1**
- Can also define at relation level

TDM: Coverage (Relations)

- Each relation **R** has coverage $\in [0,1]$
  - Informally: percentage of correct **R** that is present
  - Default: **coverage=1**

Void Part 2

- Started fiddling around with TDM accuracy
  - Suitability for applications
  - Expressiveness in general
  - Operations on data
- Immediately encountered interesting issues
  - Modeling is nontrivial
  - Operation behavior is nonobvious
  - Completeness and closure

End Void

Time to…

- Read up on other work
- Study a simplified accuracy model
- Get formal
- Change terminology ☺

Uncertain Database: Semantics

Definition: An uncertain database represents a set of possible (certain) databases

a.k.a. “possible worlds” “possible instances”

Example: Jennifer attends workshop on Monday; Mike attends on Monday, Tuesday, or not at all

person	day
Jennifer	Monday
Mike	Monday

Instance1

person	day
Jennifer	Monday
Mike	Tuesday

Instance2

person	day
Jennifer	Monday

Instance3

Restricted TDM Accuracy

1. Approximation: or-sets
2. Confidence: maybe-tuples (denoted "?")
3. Coverage: omit

Straightforward mapping to possible-instances

person	day	
Jennifer	Monday	
Mike	{Monday,Tuesday}	?

maps to the three possible-instances on previous slide

Properties of Repsentations

- Restricted-TDM is one possible representation for uncertain databases
- A representation is well-defined if we know how to map any database in the representation to its set of possible instances
- A representation is complete if every set of possible instances can be represented

➢ Unfortunately, TDM (restricted or not) is incomplete

Incompleteness

person	day
Jennifer	Monday
Mike	Tuesday

Instance1

person	day
Jennifer	Monday

Instance2

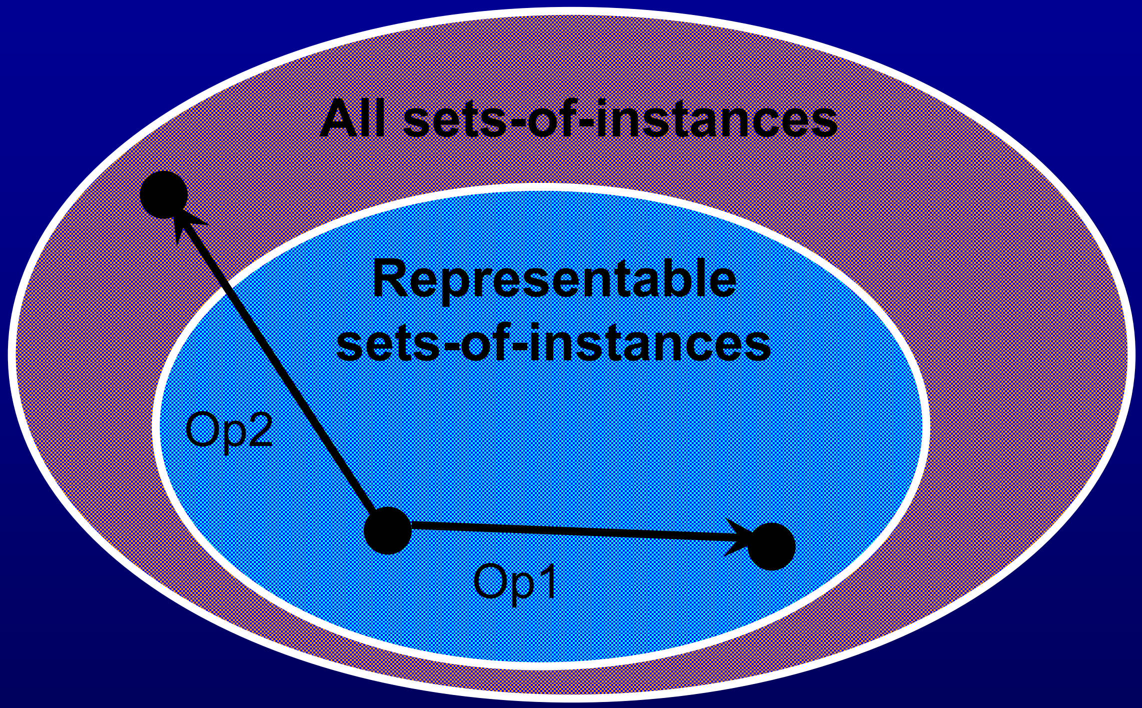
person	day
Mike	Tuesday

Instance3

person	day	
Jennifer	Monday	?
Mike	Tuesday	?

generates 4th instance: empty relation

Completeness vs. Closure



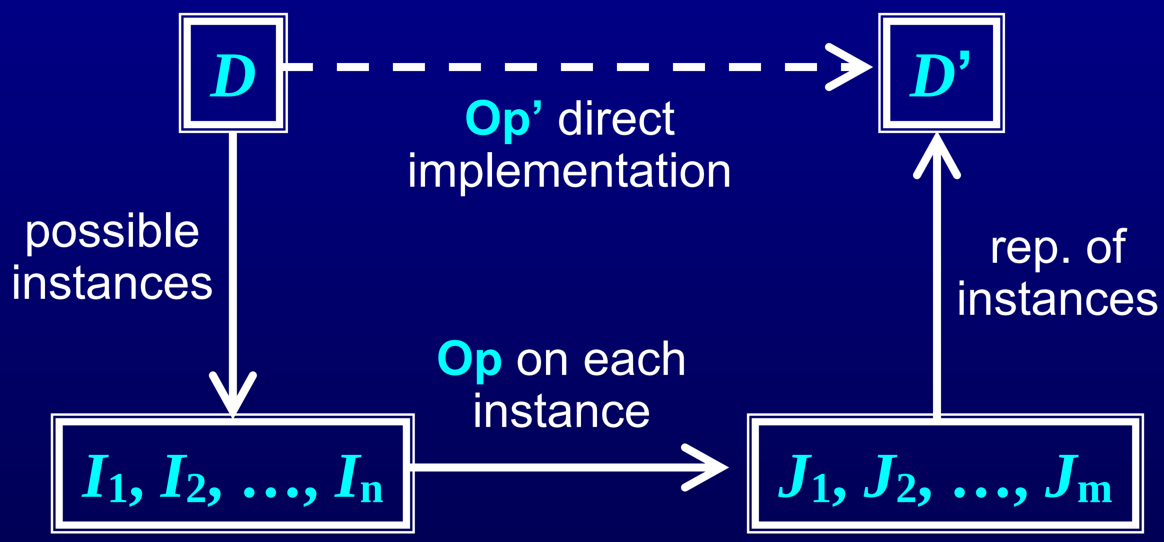
Completeness:
blue=pink

Closure:
arrow stays in blue

Proposition: An incomplete representation is still interesting if it's expressive enough and closed under all required operations

Operations: Semantics

Easy and natural (re)definition for any standard database operation (call it **Op**)



Closure:
up-arrow always exists

Note: Completeness $\Rightarrow$ Closure

Closure in TDM

Unfortunately, TDM (restricted or not) is not closed under many standard operations

Next:

1. Examples of non-closure in TDM
2. Suggest possible extensions to the representation (hereafter "model")
3. Hierarchy of models based on expressiveness

Non-Closure Under Join (Ex. 1)

person	day
Mike	{Monday,Tuesday}

$\bowtie$

day	food
Monday	chicken
Tuesday	fish

Result has two possible instances:

person	day	food
Mike	Monday	chicken

Instance1

person	day	food
Mike	Tuesday	fish

Instance2

Not representable with or-sets and ?

Need for Xor

Result has two possible instances:

person	day	food
Mike	Monday	chicken

Instance1

person	day	food
Mike	Tuesday	fish

Instance2

Representable with Xor constraint

	person	day	food
t1	Mike	Monday	chicken
t2	Mike	Tuesday	fish

Constraint: **t1 XOR t2**

Non-Closure Under Join (Ex. 2)

person	day
Mike	{Monday,Tuesday}

$\bowtie$

day	food
Monday	chicken
Monday	pie

Result has two possible instances:

person	day	food
Mike	Monday	chicken
Mike	Monday	pie

Instance1

person	day	food

Instance2

Not representable with or-sets and ?

Need for Iff

Result has two possible instances:

person	day	food
Mike	Monday	chicken
Mike	Monday	pie

Instance1

person	day	food

Instance2

Representable with ≡ (Iff) constraint

person	day	food	
Mike	Monday	chicken	**t1**
Mike	Monday	pie	**t2**

Constraint: **t1 ≡ t2**

Constraints ⇒ Completeness?

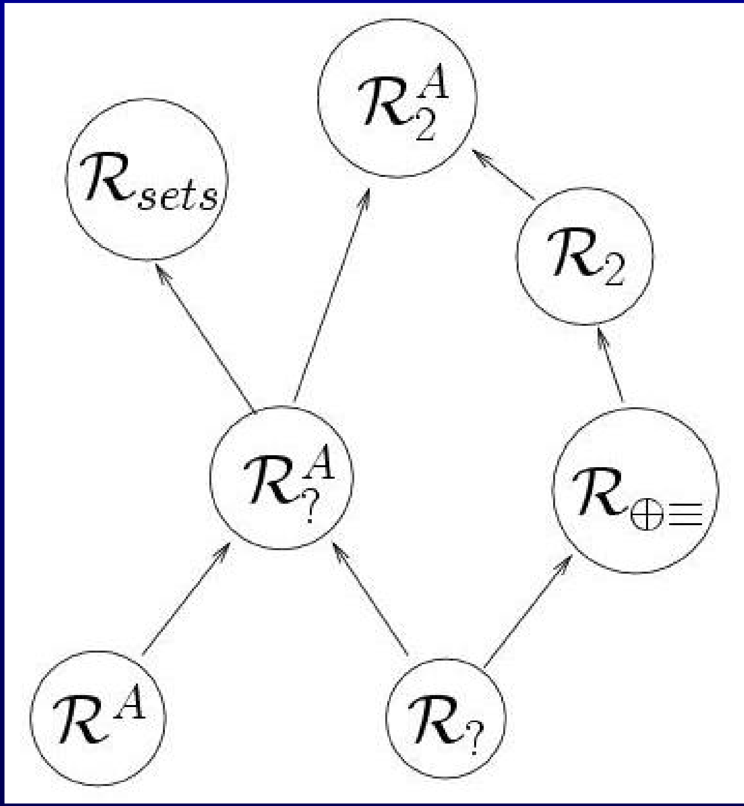
Full propositional logic: YES

Xor and **Iff**: NO

General 2-clauses: NO

How about "vertical or" (tuple-sets)? NOPE

Hierarchy of Incomplete Models



R	relations
A	or-sets
?	maybe-tuples
2	2-clauses
sets	tuple-sets

Closure

But remember:

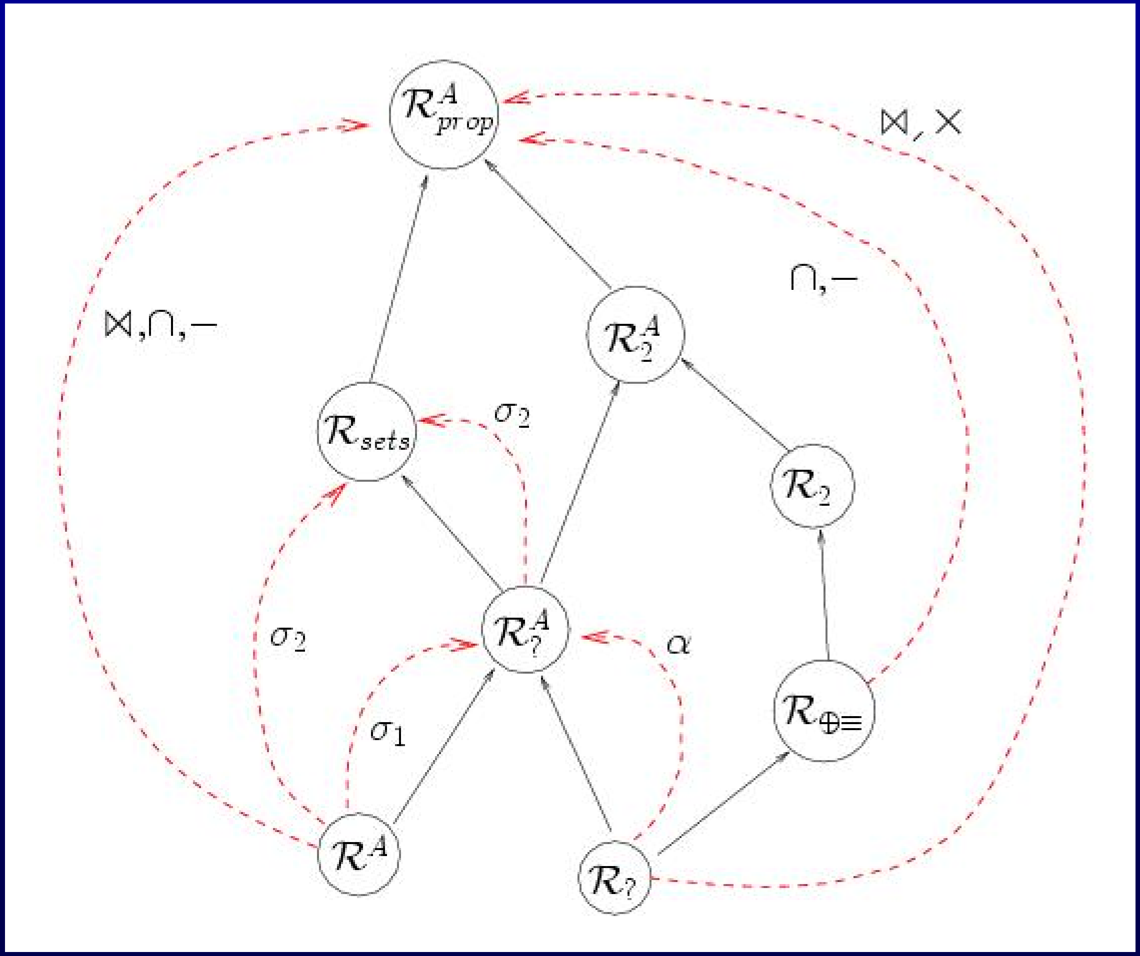
- Completeness may not be necessary
- Closure may be good enough

Are any of these models closed under standard relational operations?

Closure Table

Closure-Model	$\mathcal{R}^A$	$\mathcal{R}_?$	$\mathcal{R}^A_?$	$\mathcal{R}_{\oplus\equiv}$,$\mathcal{R}_2$,$\mathcal{R}^A_2$,$\mathcal{R}_{sets}$
Union	Y	Y	Y	Y
$Select_{ee}$	Y	Y	Y	Y
$Select_{es}$	N	Y	Y	Y
$Select_{ss}$	N	Y	N	Y
Intersection	N	Y	N	N
Cross Product	Y	N	N	N
Join	N	N	N	N
Difference	N	Y	N	N
Projection	Y	Y	Y	Y
Duplicate Elimination	N	Y	N	N
Aggregation	N	N	N	N

Closure Diagram



Omitted:

- Self-loops
- Subsumed arrows to root

Some Final Theory (for today)

Instance membership: Given instance I and uncertain relation R, is I an instance of R?

Instance certainty: Given instance I and uncertain relation R, is I R's only instance?

Tuple membership: Given tuple *t* and uncertain relation R, is *t* in any of R's instances?

Tuple certainty: Given tuple *t* and uncertain relation R, is *t* in all of R's instances?

Many of these problems are NP-Hard in complete models but polynomial in our incomplete models

Back to Reality

What does all of this mean for the Trio project?

Fundamental dilemma:

- Restricted-TDM: intuitive, understandable, incomplete
- Unrestricted-TDM: more complex, still incomplete
- Complete models: even more complex, nonintuitive

Uses of Incomplete Models

Sufficient for some applications

- Incomplete model can represent data
- Closed under required operations

Two-layer approach

- Underlying complete model
- Incomplete "working" model for users (recall Mike's chicken and fish)
- Challenge: approximate approximation

Lineage and Uncertainty

Trio: Data + Accuracy + Lineage

Surprise: Restricted-TDM (v2) + Lineage is complete and (therefore) closed

person	day
Mike	{Monday,Tuesday}

A1 A2

⋈

day	food
Monday	chicken
Tuesday	fish

person	day	food
Mike	Monday	chicken
Mike	Tuesday	fish

Lineage: A1, …

Lineage: A2, …

Current Challenges

Pursue uncertainty+lineage

Remainder of accuracy model

- Probability distributions
- Intervals, Gaussians
- Confidence values
- Coverage

Querying uncerrtainty

Ex: Find all people with ≥ 3 alternate days

➢ Can we generalize the possible-instances semantics?

Search term: stanford trio