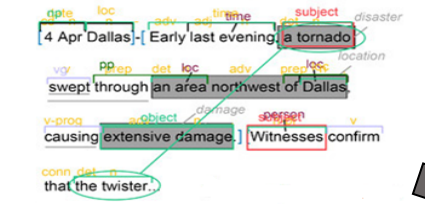
Probabilistic Declarative Information Extraction

Daisy Zhe Wang, Eirinaios Michelakis,
Michael J. Franklin, Minos Garofalakis, Joseph M. Hellerstein

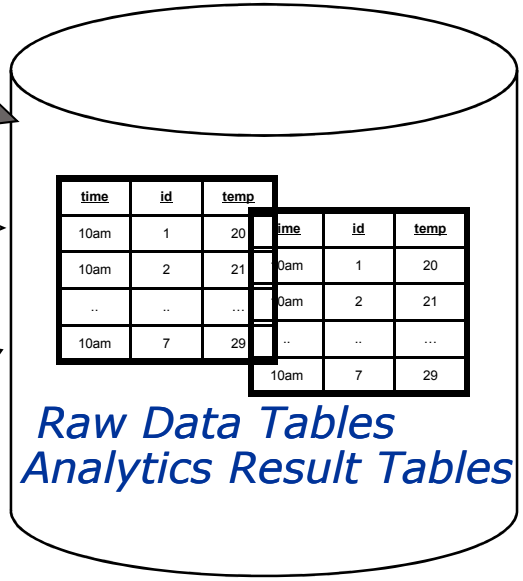
ICDE, Long Beach, 2nd March, 2010

State-of-the-art ---- DB meets ML



Probabilistic Data Sources

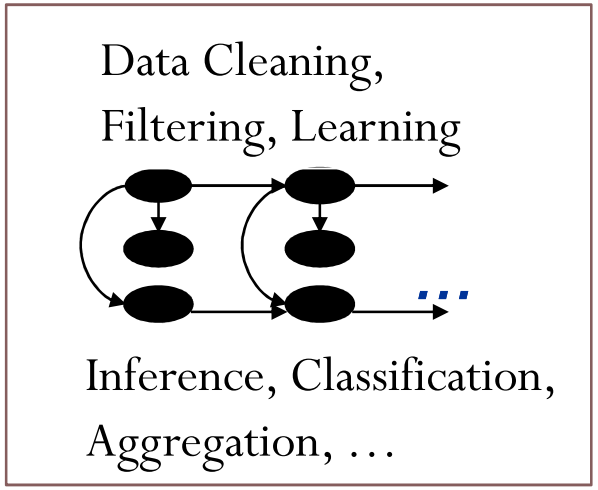
*SELECT ***
FROM ANALYTICS



Raw Data Tables
Analytics Result Tables

Relational DBMS

INPUT FILE



Statistical ML Packages

OUTPUT FILE

Problems and Solutions

- Problems
  - Scalability of SML packages
  - Expensive Data Transfer Cost
  - DB only as a Data Storage
- Solutions – Efficient and Scalable In-Database ML
  - Represent Model and Probabilistic Data as First-class Objects
  - Implement Inference as Native Operators
  - Support Probabilistic Queries
  - Optimize across Inference and Relational Operations

Probabilistic Information Extraction (IE)

- IE Tasks: unstructured text → structured data
- Rule-based IE (e.g. DBLife)
- Probabilistic IE: better accuracy, flexibility and adaptability
  - Naïve Bayes (NB)
  - Hidden Markov Model (HMM)
  - Conditional Random Fields (CRF)
- Database Research on IE
  - Declarative IE
  - Probabilistic Databases

Probabilistic Declarative IE

Conditional Random Fields (CRF)

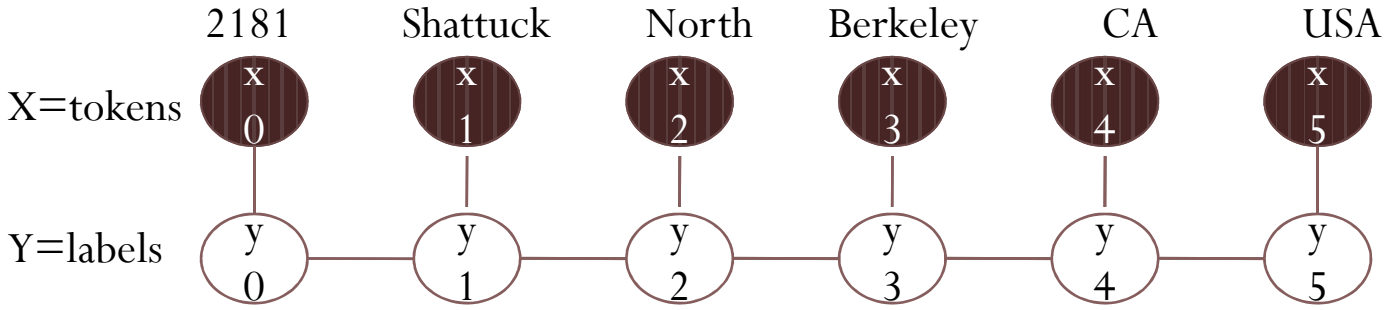
Text (address string):

E.g., *"2181 Shattuck North Berkeley CA USA"*

Possible Extraction Worlds:

x	2181	Shattuck	North	Berkeley	CA	USA	
y1	**apt. num**	**street name**	**city**	**city**	**state**	**country**	(0.6)
y2	**apt. num**	**street name**	**street name**	**city**	**state**	**country**	(0.1)

CRF Model:

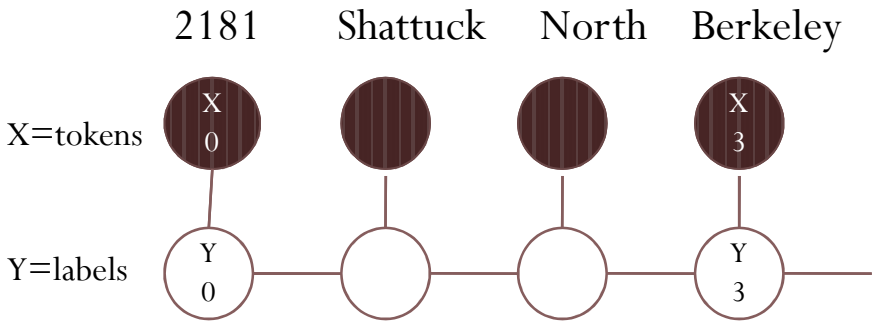


Viterbi Top-k Inference on CRF

Viterbi Dynamic Programming Algorithm:

$$V(i,y) = \begin{cases} \max_{y'}(V(i-1,y') \\ \quad + \sum_{k=1}^{K} \lambda_k f_k(y,y',x_i)), & \text{if} \quad i \geqslant 0 \\ 0, & \text{if} \quad i = -1. \end{cases} \qquad (3)$$

CRF Model:

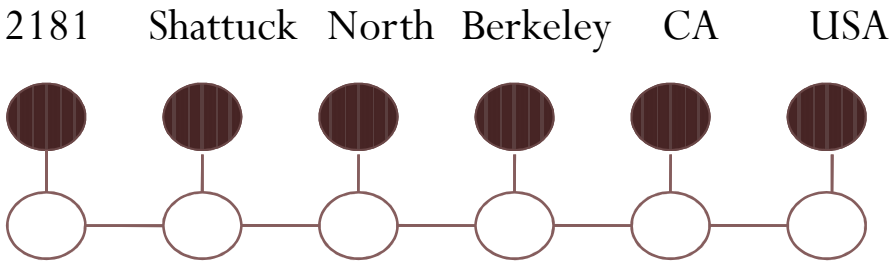


Dynamic Programming V matrix:

pos	street num	street name	city	state	country
0	5	1	0	1	1
1	2	15	7	8	7
2	12	24	21	18	17
3	21	32	32	30	26
4	29	40	38	42	35
5	39	47	46	46	50

Text Data and CRF Representations

2181 Shattuck North Berkeley CA USA
342 S Montezuma St Prescott AZ USA
P.O. Box 210732 Minneapolis MN USA
9330 Eastex Fwy Houston TX USA
225 16th West Vancouver BC CAN
1084 Salk Road Pickering ON CAN ...



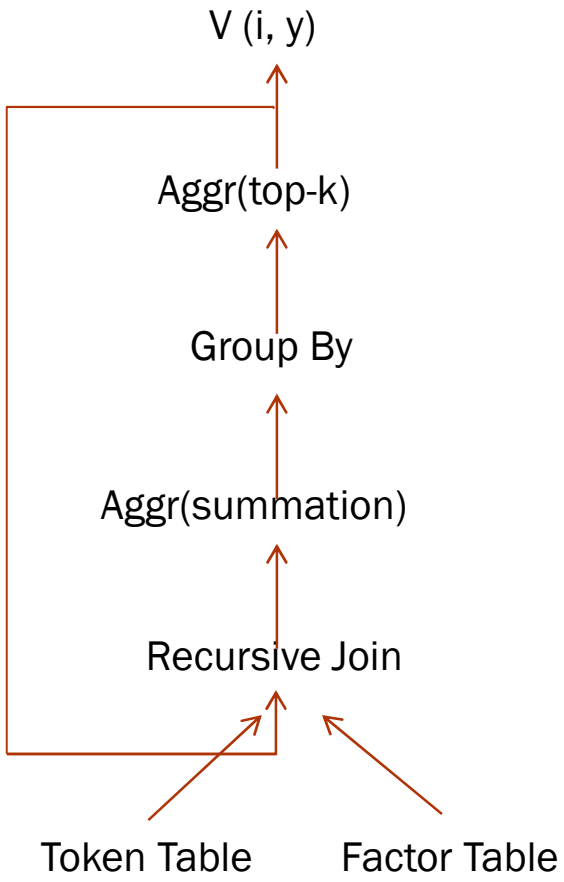
docID	pos	token	Label
1	0	2181	
1	1	Shattuck	
1	2	North	
1	3	Berkeley	
1	4	CA	
1	5	USA	

Token Table

token	prevLabel	label	score
2181(DIGIT)	null	street num	22
2181(DIGIT)	null	street name	5
…	..	..	
Berkeley	streetname	street name	10
Berkeley	streetname	city	25
..	..	..	

Factor Table

Implement Dynamic Programming Algorithm using Recursive Queries

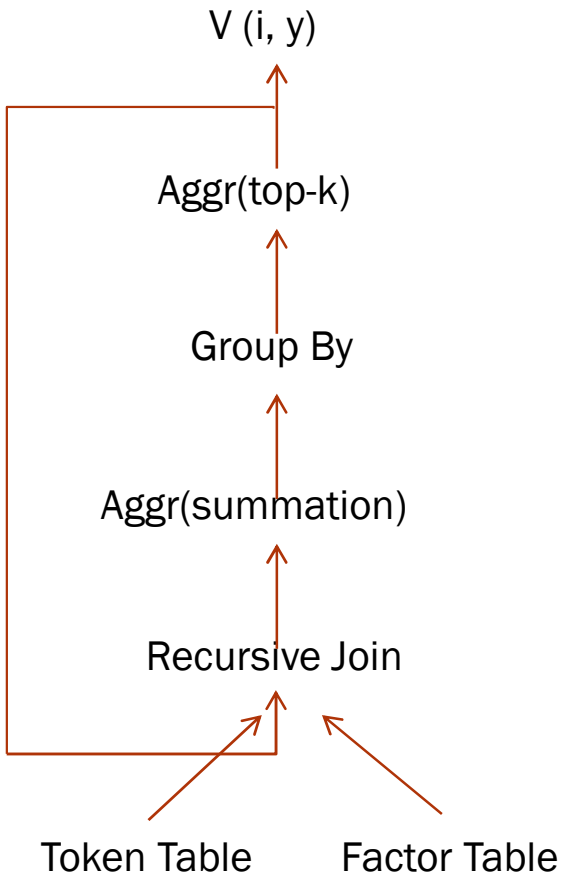


Dynamic Programming V matrix:

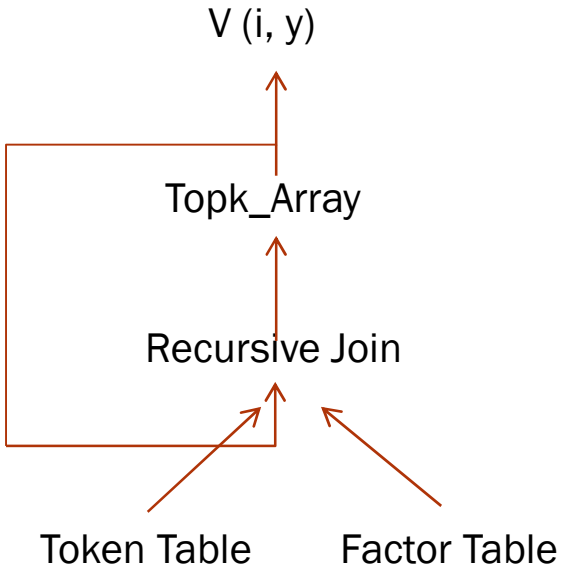
pos	street num	street name	city	state	country
0	5	1	0	1	1
1	2	15	7	8	7
2	12	24	21	18	17
3	21	32	32	30	26
4	29	40	38	42	35
5	39	47	46	46	50

ViterbiPerDoc (ViterbiAllDoc)

Viterbi Implemented in SQL

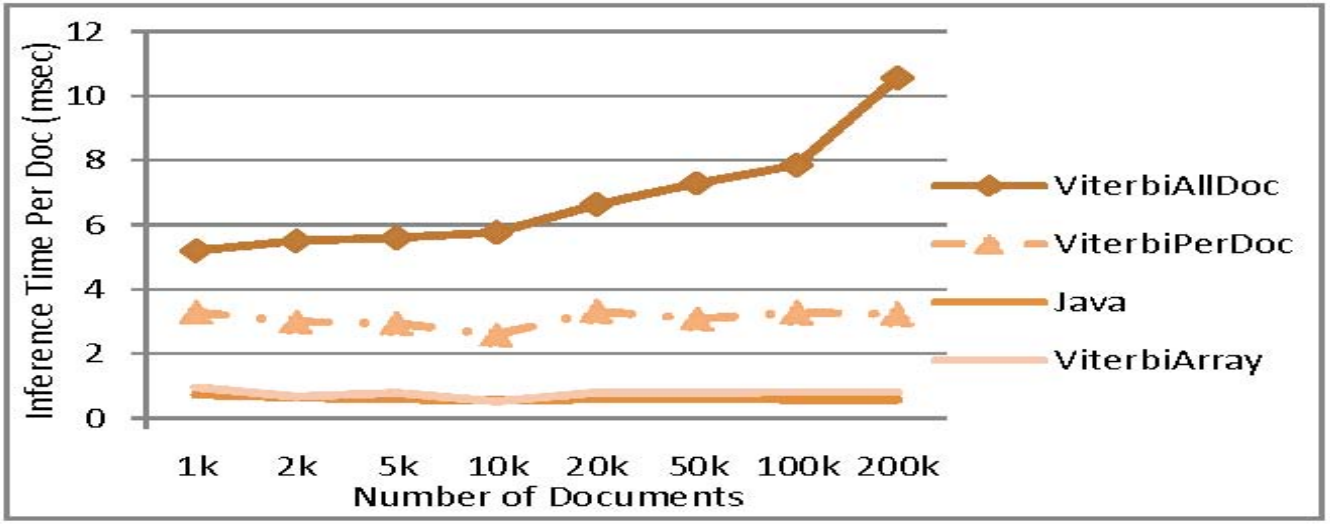


ViterbiPerDoc (ViterbiAllDoc)



ViterbiArray

Evaluation: [Runtime Efficiency] SQL Implementations of Viterbi



Conclusions

- Probabilistic Declarative IE
  - Motivation of In-Database ML – Efficiency, Scalability
  - Text Data and CRF Model Representation
  - Viterbi Inference SQL Implementation
  - Experiment Results
- BayesStore System **[VLDB'08]**
- Current & Future Work
  - Probabilistic Querying over CRF-based IE **[Submitted to PVLDB]**
  - Parallelization
  - Other IE Tasks (e.g. Coreference)

Thank you! … Questions? ☺

SQL Implementation of CRF model can be downloaded at BayesStore Project Page:
http://www.cs.berkeley.edu/~daisyw/BayesStore.html