Probabilistic Fact-Finding at Scale

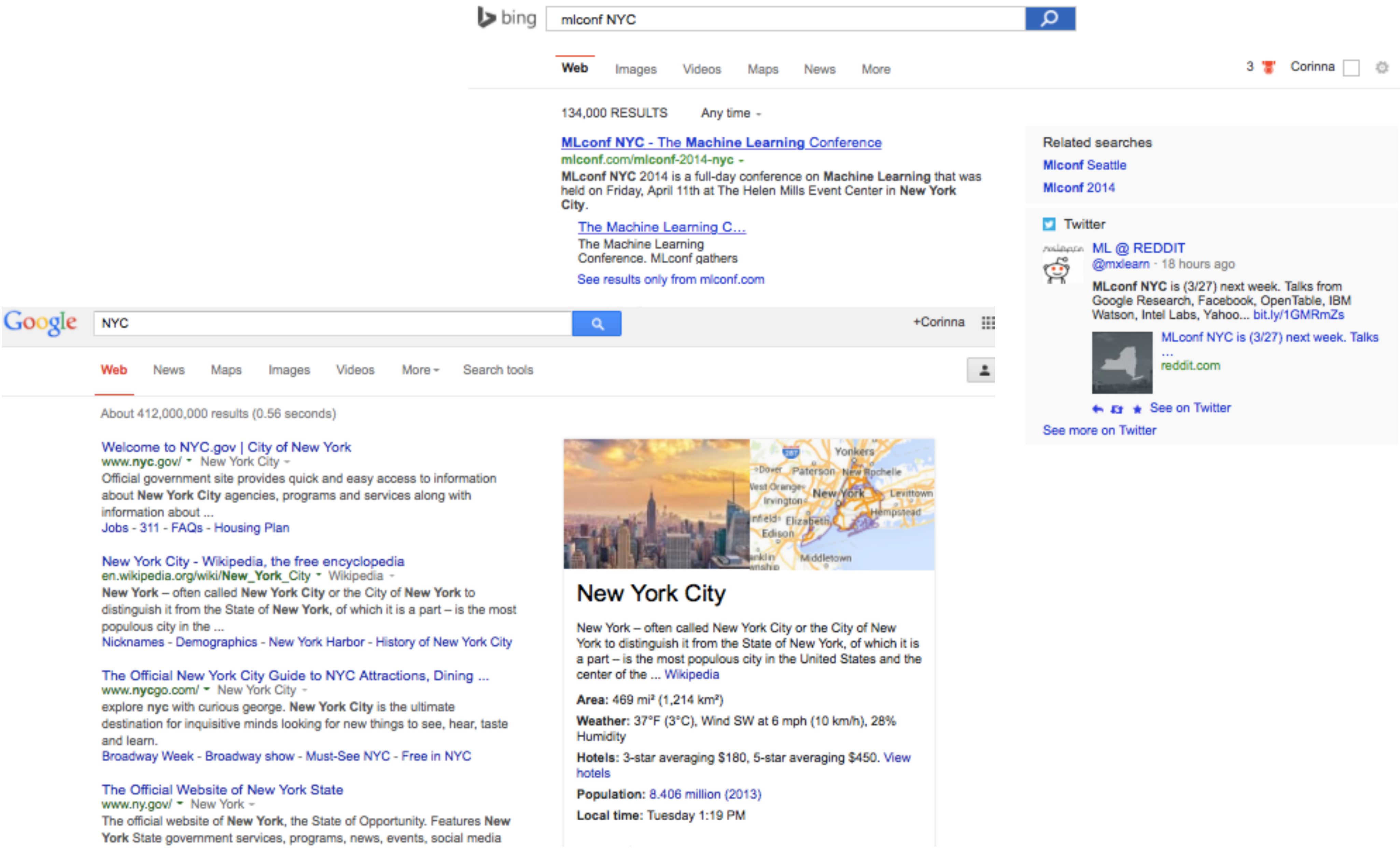
Large-scale machine learning at Google

Corinna Cortes

Google Research

corinna@google.com

Knowledge Panels a Commodity



134,000 RESULTS Any time

MLconf NYC - The Machine Learning Conference
mlconf.com/mlconf-2014-nyc
MLconf NYC 2014 is a full-day conference on **Machine Learning** that was held on Friday, April 11th at The Helen Mills Event Center in **New York City**.

The Machine Learning C...
The Machine Learning Conference. MLconf gathers

See results only from mlconf.com

Related searches
Mlconf Seattle
Mlconf 2014

Twitter
ML @ REDDIT
@mxlearn · 18 hours ago
MLconf NYC is (3/27) next week. Talks from Google Research, Facebook, OpenTable, IBM Watson, Intel Labs, Yahoo... bit.ly/1GMRmZs

MLconf NYC is (3/27) next week. Talks ...
reddit.com

See on Twitter
See more on Twitter

About 412,000,000 results (0.56 seconds)

Welcome to NYC.gov | City of New York
www.nyc.gov/ · New York City
Official government site provides quick and easy access to information about **New York City** agencies, programs and services along with information about ...
Jobs - 311 - FAQs - Housing Plan

New York City - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/New_York_City · Wikipedia
New York – often called **New York City** or the City of **New York** to distinguish it from the State of **New York**, of which it is a part – is the most populous city in the ...
Nicknames - Demographics - New York Harbor - History of New York City

The Official New York City Guide to NYC Attractions, Dining ...
www.nycgo.com/ · New York City
explore **nyc** with curious george. **New York City** is the ultimate destination for inquisitive minds looking for new things to see, hear, taste and learn.
Broadway Week - Broadway show - Must-See NYC - Free in NYC

The Official Website of New York State
www.ny.gov/ · New York
The official website of **New York**, the State of Opportunity. Features **New York** State government services, programs, news, events, social media

New York City

New York – often called New York City or the City of New York to distinguish it from the State of New York, of which it is a part – is the most populous city in the United States and the center of the ... Wikipedia

Area: 469 mi² (1,214 km²)

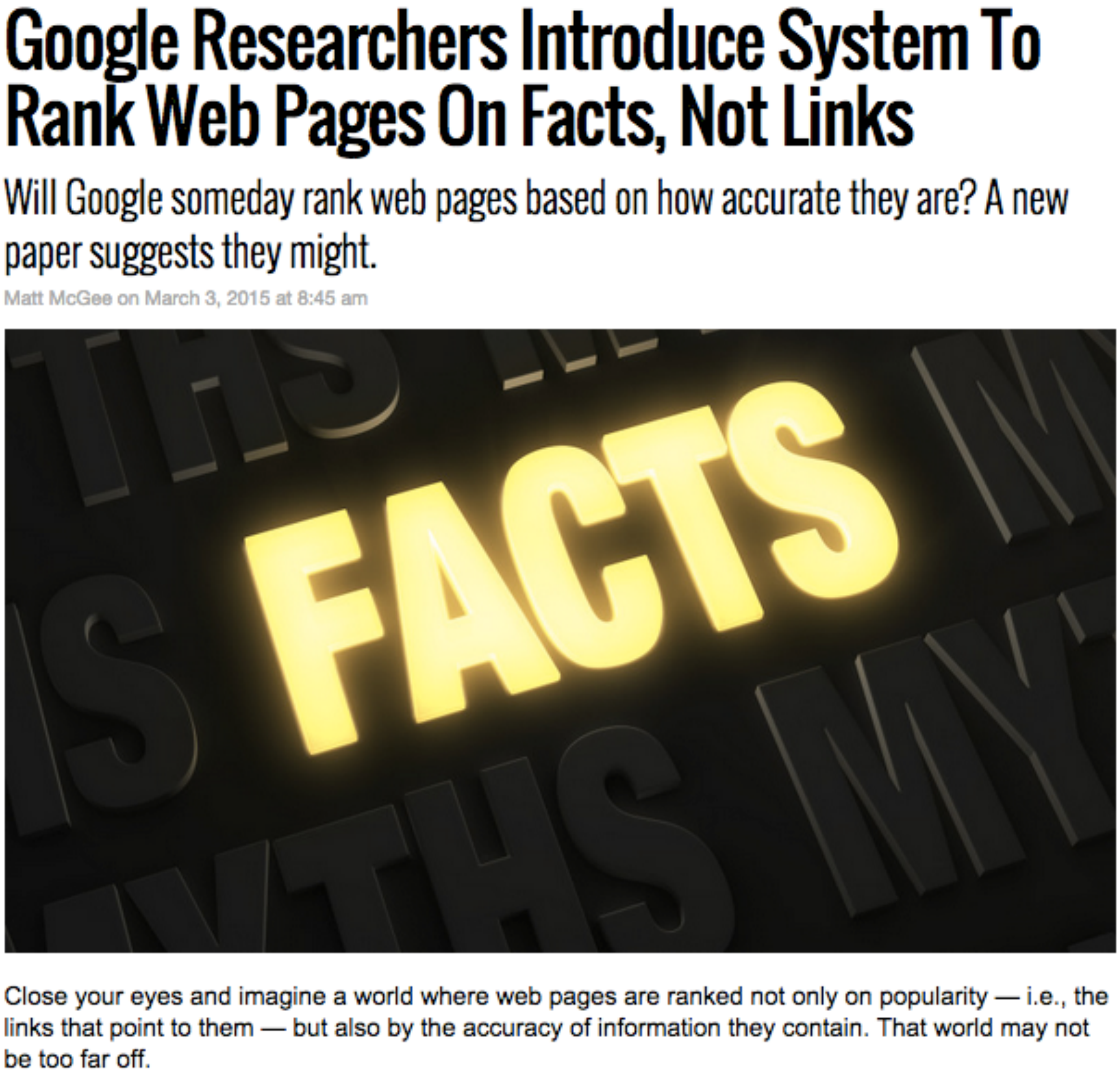
Weather: 37°F (3°C), Wind SW at 6 mph (10 km/h), 28% Humidity

Hotels: 3-star averaging $180, 5-star averaging $450. View hotels

Population: 8.406 million (2013)

Local time: Tuesday 1:19 PM

3/3/2015: http://searchengineland.com/google-researchers-introduce-system-rank-web-pages-facts-not-links-215835



Google Researchers Introduce System To Rank Web Pages On Facts, Not Links

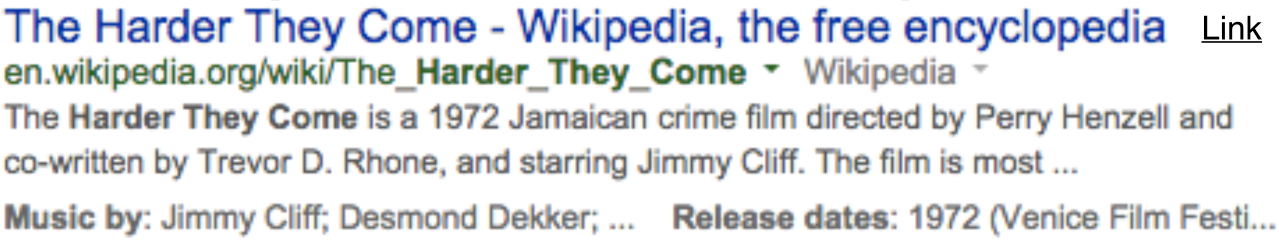
Will Google someday rank web pages based on how accurate they are? A new paper suggests they might.

Matt McGee on March 3, 2015 at 8:45 am

Close your eyes and imagine a world where web pages are ranked not only on popularity — i.e., the links that point to them — but also by the accuracy of information they contain. That world may not be too far off.

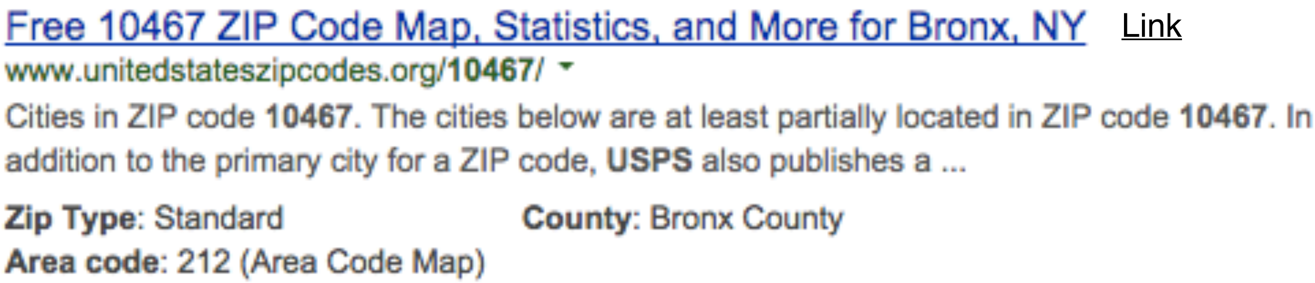
Structured Snippets

- [harder they come, harder they fall]



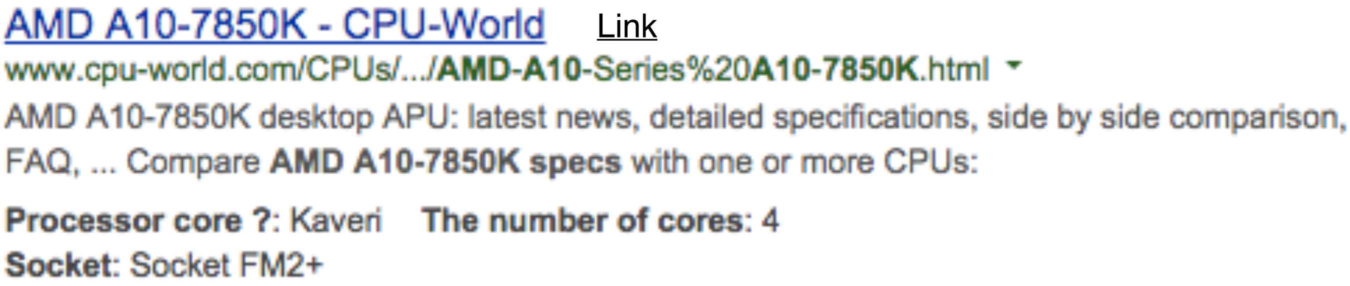
The Harder They Come - Wikipedia, the free encyclopedia Link
en.wikipedia.org/wiki/The_**Harder_They_Come** ▾ Wikipedia ▾
The **Harder They Come** is a 1972 Jamaican crime film directed by Perry Henzell and co-written by Trevor D. Rhone, and starring Jimmy Cliff. The film is most ...
Music by: Jimmy Cliff; Desmond Dekker; ... **Release dates**: 1972 (Venice Film Festi...

- [10467 usps]



Free 10467 ZIP Code Map, Statistics, and More for Bronx, NY Link
www.unitedstateszipcodes.org/**10467**/ ▾
Cities in ZIP code **10467**. The cities below are at least partially located in ZIP code **10467**. In addition to the primary city for a ZIP code, **USPS** also publishes a ...
Zip Type: Standard **County**: Bronx County
Area code: 212 (Area Code Map)

- [amd a10 7850k specs]



AMD A10-7850K - CPU-World Link
www.cpu-world.com/CPUs/.../**AMD-A10**-Series%20**A10-7850K**.html ▾
AMD A10-7850K desktop APU: latest news, detailed specifications, side by side comparison, FAQ, ... Compare **AMD A10-7850K specs** with one or more CPUs:
Processor core ?: Kaveri **The number of cores**: 4
Socket: Socket FM2+

Structured Snippets

- Launched Q3, 2014
  - http://googleresearch.blogspot.com/2014/09/introducing-structured-snippets-now.html
  - 40+ online articles (US as well as international), including articles from Engadget, PC Magazine, Search Engine Land, …

Structured Snippets

- Highlights
  - “[...] makes Google smarter and means you'll have to do less clicking in some cases, so we're all for it.” -- The Verge, 2014-09-23
  - “[...] the results themselves have usually been skimpy; you've seen preview text, and that's about it. Thankfully, Google has made that sneak peek considerably more useful.” -- Engadget, 2014-09-23
  - “[...] Now, Google is making search even more convenient.” -- PC Magazine, 2014-09-23
  - “[...] show Google's ongoing emphasis on extracting structure and entities from unstructured content.” -- Search Engine Watch, 2014-09-24
  - "[...] The big pro for users, of course, is that structured snippets make searching for specific facts easier and faster. Google’s search results are smarter and more detailed, so users will have to do less clicking to find the information they’re searching for." -- Mareeg Media, 2014-09-26
  - …

Probabilistic Fact-Finding at Scale

- **Knowledge databases** of curated facts:
  - Freebase, https://www.freebase.com
  - Google Knowledge Graph, http://www.google.com/insidesearch/features/search/knowledge.html
- **Web-scale probabilistic knowledge base** that combines extractions from Web content

Outline

- WebTables
  - How do we find the good tables on the web?
- Knowledge Vault
  - How do we find other quality facts on the web?
- Lattice Regression
  - How do we form interpretable non-linear models?

Web Tables

- "Applying WebTables in Practice", CIDR 2015
  Sreeram Balakrishnan, Alon Halevy, Boulos Harb, Hongrae Lee, Jayant Madhavan, Afshin Rostamizadeh, Warren Shen, Kenneth Wilder, Fei Wu, Cong Yu.
- March 5, 2014: 11 billion HTML tables reduced to 147 million quasi-relational Web tables,
  http://webdatacommons.org/webtables/

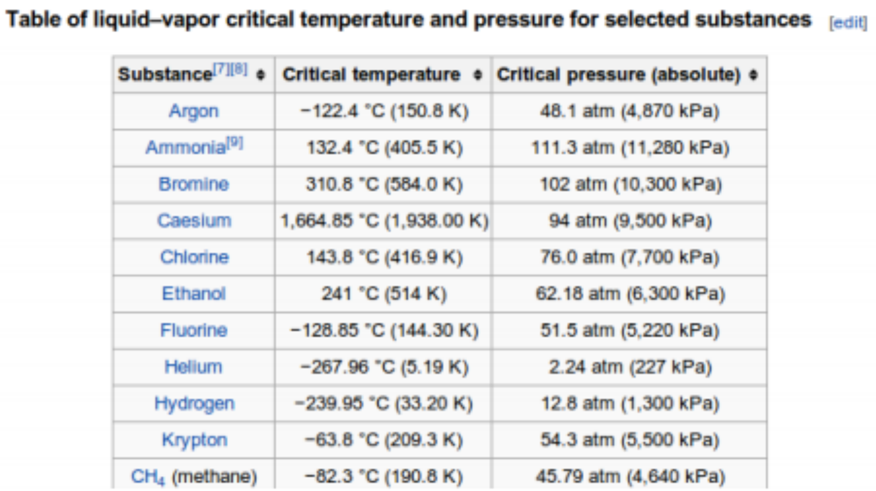
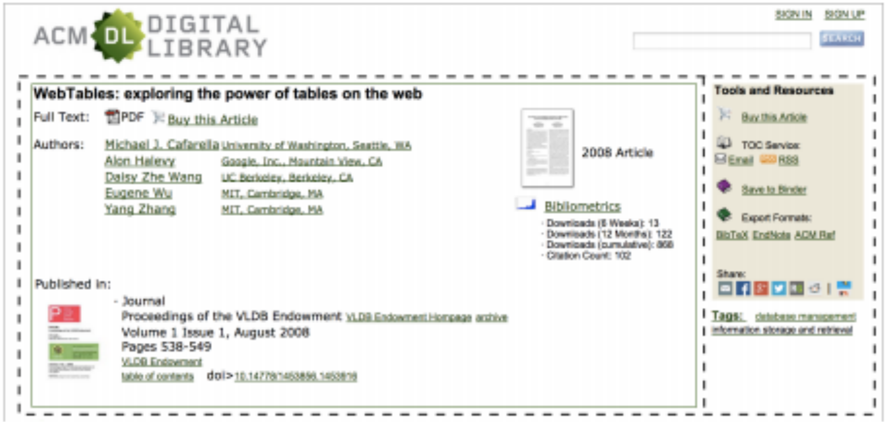


Table of liquid–vapor critical temperature and pressure for selected substances [edit]

Substance[7][8] ⬍	Critical temperature ⬍	Critical pressure (absolute) ⬍
Argon	−122.4 °C (150.8 K)	48.1 atm (4,870 kPa)
Ammonia[9]	132.4 °C (405.5 K)	111.3 atm (11,280 kPa)
Bromine	310.8 °C (584.0 K)	102 atm (10,300 kPa)
Caesium	1,664.85 °C (1,938.00 K)	94 atm (9,500 kPa)
Chlorine	143.8 °C (416.9 K)	76.0 atm (7,700 kPa)
Ethanol	241 °C (514 K)	62.18 atm (6,300 kPa)
Fluorine	−128.85 °C (144.30 K)	51.5 atm (5,220 kPa)
Helium	−267.96 °C (5.19 K)	2.24 atm (227 kPa)
Hydrogen	−239.95 °C (33.20 K)	12.8 atm (1,300 kPa)
Krypton	−63.8 °C (209.3 K)	54.3 atm (5,500 kPa)
CH_4 (methane)	−82.3 °C (190.8 K)	45.79 atm (4,640 kPa)

(a) The table shows critical temperature and pressure for various substances as a tabular data.



WebTables: exploring the power of tables on the web

Full Text: PDF Buy this Article

Authors: Michael J. Cafarella University of Washington, Seattle, WA
Alon Halevy Google, Inc., Mountain View, CA
Daisy Zhe Wang UC Berkeley, Berkeley, CA
Eugene Wu MIT, Cambridge, MA
Yang Zhang MIT, Cambridge, MA

2008 Article

Bibliometrics
· Downloads (6 Weeks): 13
· Downloads (12 Months): 122
· Downloads (cumulative): 868
· Citation Count: 102

Published in:
- Journal
Proceedings of the VLDB Endowment VLDB Endowment Hompage archive
Volume 1 Issue 1, August 2008
Pages 538-549
VLDB Endowment
table of contents doi>10.14778/1453856.1453916

Tools and Resources
Buy this Article
TOC Service:
Email RSS
Save to Binder
Export Formats:
BibTeX EndNote ACM Ref

Share:

Tags: database management information storage and retrieval

(b) The table marked by dotted line includes contents, but its primary goal is to control the layout.

Finding the ‘Good’ Tables

- **Machine learning classifier**
  - feature design
  - training example generation
  - model selection

Feature Design

- **Complicating fact:** the semantics of the table is often determined by the surrounding text;
- **Detecting subject columns**: many tables contain a subject column, the other columns contains properties of the subject:
  - Binary classifier, 94% accuracy;
  - Header row of other columns used as a schema.
- **Determining possible column classes:**
  - Annotate entries using the Google Knowledge Graph, (subject, predicate, value), pick the largest class(es).

Feature Design

- **Determining properties of the subject column**:
  - Captions may mention the properties
  - Query stream and text used to determine what properties appear with what classes. "Biperpedia: An Ontology for Search Applications", VLDB 2014, Rahul Gupta, Alon Halevy, Xuezhi Wang, Steven Whang, Fei Wu.
- **Additional structural features**
  - number of rows and columns, mean/variance of the number of characters per cell, the fraction of non-empty cells, the fraction of cells <th> tags, and the number of distinct tokens in a table.

Machine Learning

- **Training data:** original 98% negative, 2% positive
- **Heuristics**: eliminate tables based on rules:
    - tiny tables (less than 3 rows and 2 columns), calendars, password tables, table-of-content tables;
- **Simple classifier:** filter the tables based on cheap features;
- **A stratified sample** of good and bad tables:
    - 35% positive, 65% negative.
- **SVM classifier**, multi-kernel learning, high accuracy.

Outline

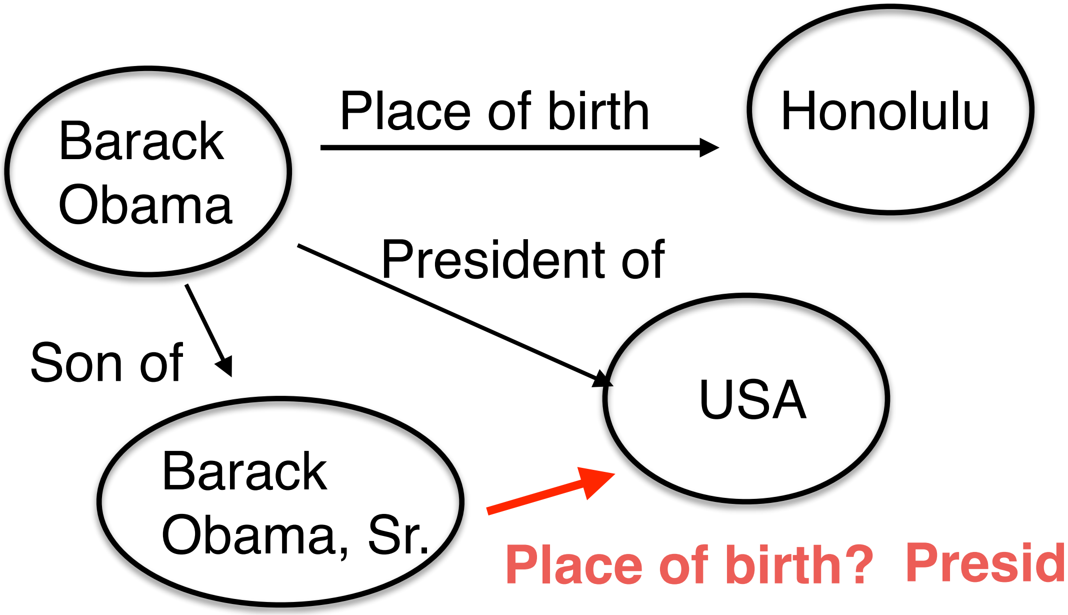
- WebTables
  - How do we find the good tables on the web?
- Knowledge Vault
  - How do we find other quality facts on the web?
- Lattice Regression
  - How do we form interpretable non-linear models?

Knowledge Vault

- "Google Researchers Introduce System To Rank Web Pages On Facts, Not Links", http://searchengineland.com/google-researchers-introduce-system-rank-web-pages-facts-not-links-215835
  - "Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources", http://arxiv.org/pdf/1502.03519v1.pdf
  - "Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion", KDD 2014, http://dl.acm.org/citation.cfm?id=2623623

Knowledge Vault

- Combines noisy extractions + prior knowledge
  - Table facts
  - Text extractors with scores derived from KG
  - Graph-based paths with scores derived from KG
  - Applies learning: P(entity, predicate, entity)



Place of birth? President of? Lives in?

Knowledge Vault

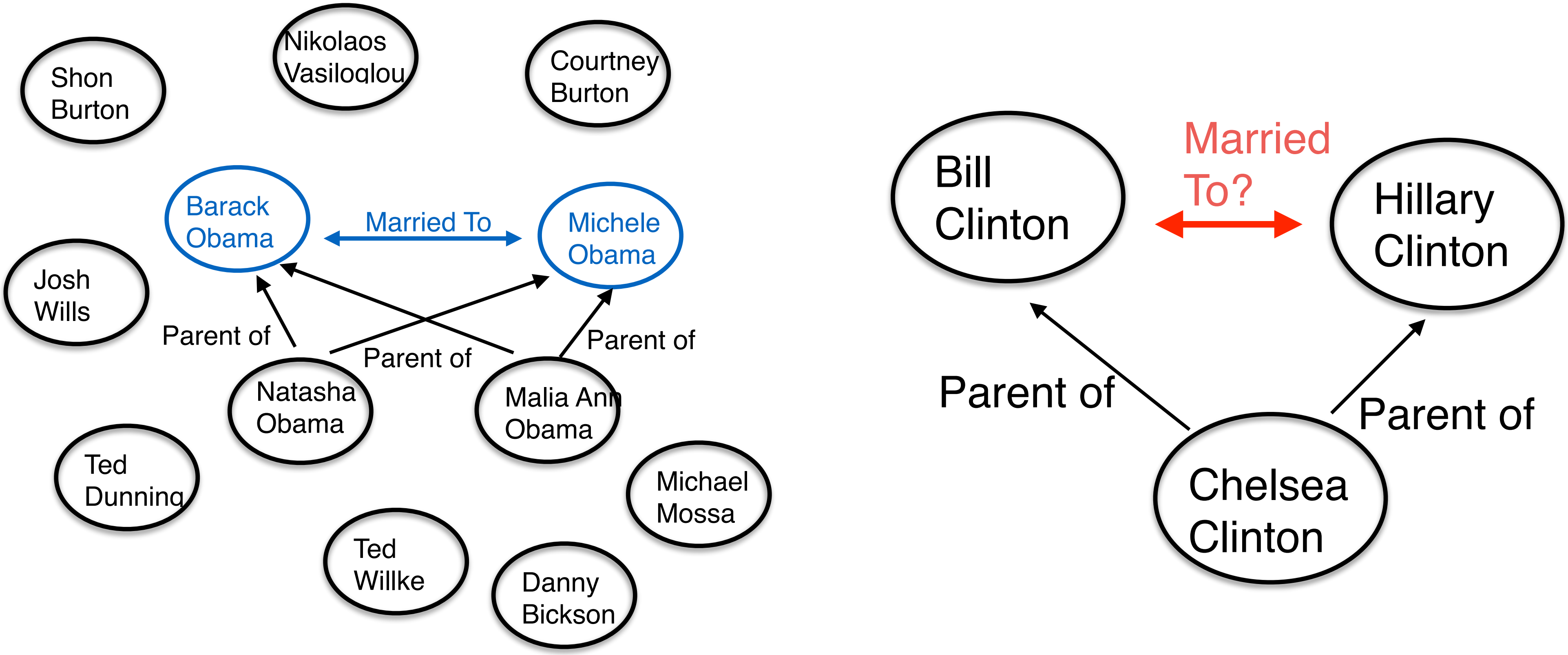
- Triplets (entity, predicate, entity) + score.
  - (</m/02mjmr, /people/person/place_of_birth, /m/02hrh0_>);
  - /m/02mjmr is the Freebase id for Barack Obama;
  - m/02hrh0_ is the id for Honolulu;
  - score is the KV probability of correctness.

Text Extraction

- **Tag text:** entity recognition, part of speech tagging, dependency parsing, co-reference resolution, …
- **Train relation extractors** using **distant supervision**:
  - given predicate, "married to":
  - find KG validated pairs for that predicate with this predicate (BarackObama, MichelleObama), (BillClinton, HillaryClinton);
  - find occurrences of the pairs, assuming they may express the predicate;
  - build classifier to score the predicate between the pairs based on the diverse occurrences.

Graph-Based Scores

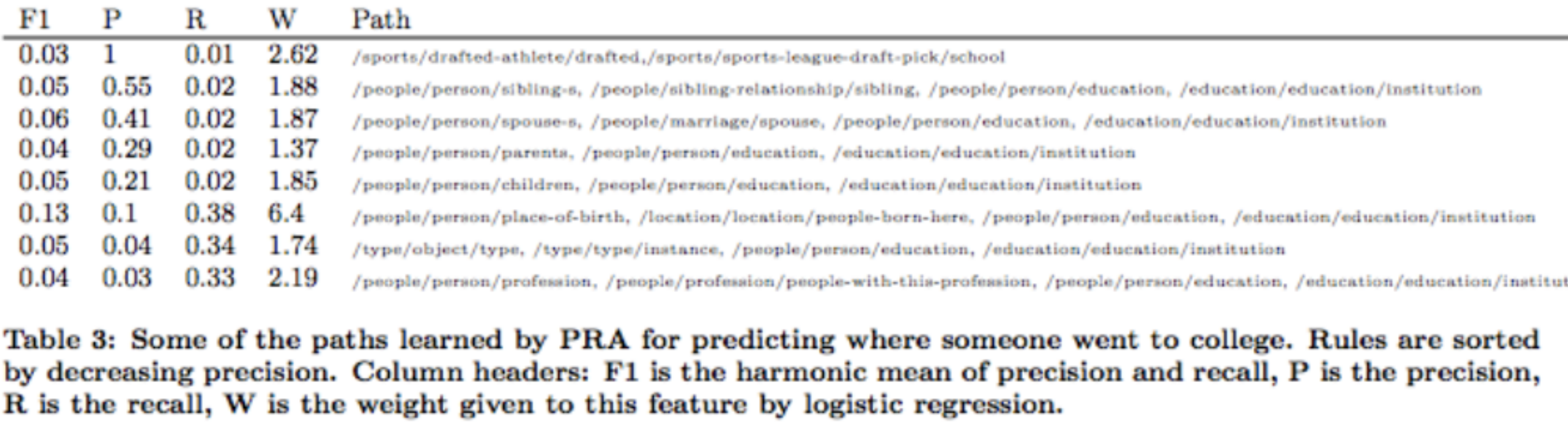
- Graph of entities



- Walk the graph to learn that two people that are both ParentOf may be married. Binary classifier based on evidence from different paths.

Knowledge Vault

- Example of learned graph-paths for 'went_to_college'



F1	P	R	W	Path
0.03	1	0.01	2.62	/sports/drafted-athlete/drafted,/sports/sports-league-draft-pick/school
0.05	0.55	0.02	1.88	/people/person/sibling-s, /people/sibling-relationship/sibling, /people/person/education, /education/education/institution
0.06	0.41	0.02	1.87	/people/person/spouse-s, /people/marriage/spouse, /people/person/education, /education/education/institution
0.04	0.29	0.02	1.37	/people/person/parents, /people/person/education, /education/education/institution
0.05	0.21	0.02	1.85	/people/person/children, /people/person/education, /education/education/institution
0.13	0.1	0.38	6.4	/people/person/place-of-birth, /location/location/people-born-here, /people/person/education, /education/education/institution
0.05	0.04	0.34	1.74	/type/object/type, /type/type/instance, /people/person/education, /education/education/institution
0.04	0.03	0.33	2.19	/people/person/profession, /people/profession/people-with-this-profession, /people/person/education, /education/education/institut

Table 3: Some of the paths learned by PRA for predicting where someone went to college. Rules are sorted by decreasing precision. Column headers: F1 is the harmonic mean of precision and recall, P is the precision, R is the recall, W is the weight given to this feature by logistic regression.

Knowledge Vault

- Train a classifier on the confidence scores from text- and graph-based extraction
  - Target values: curated facts
  - AUC= 0.911

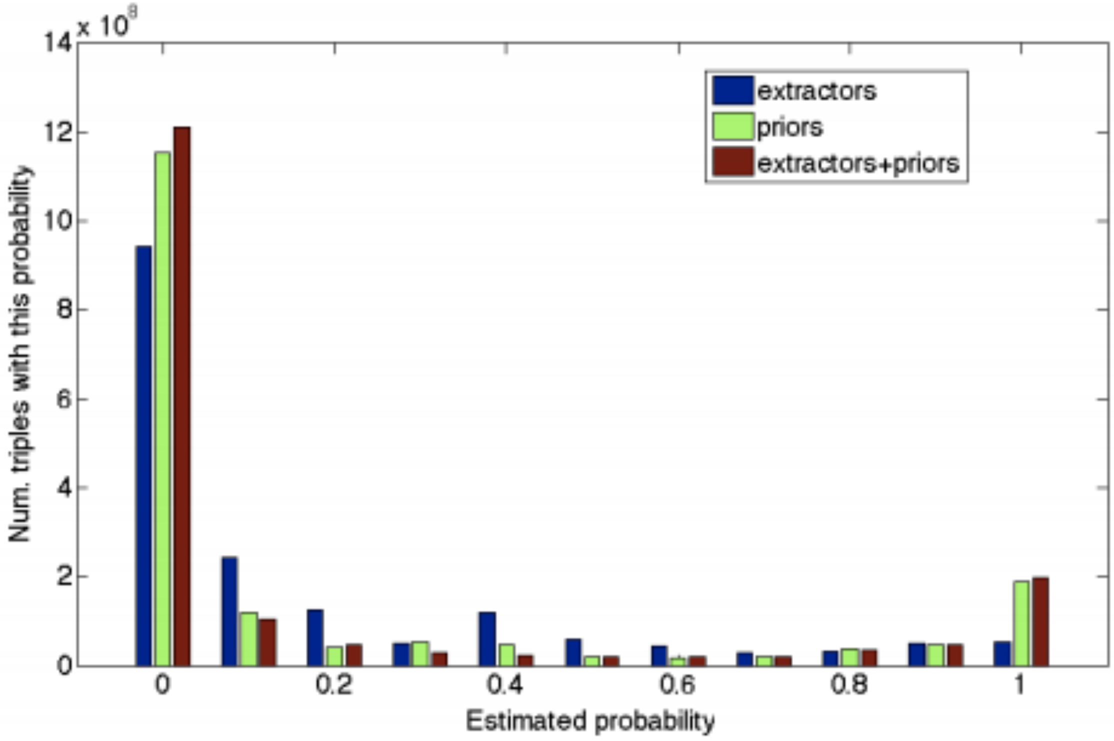


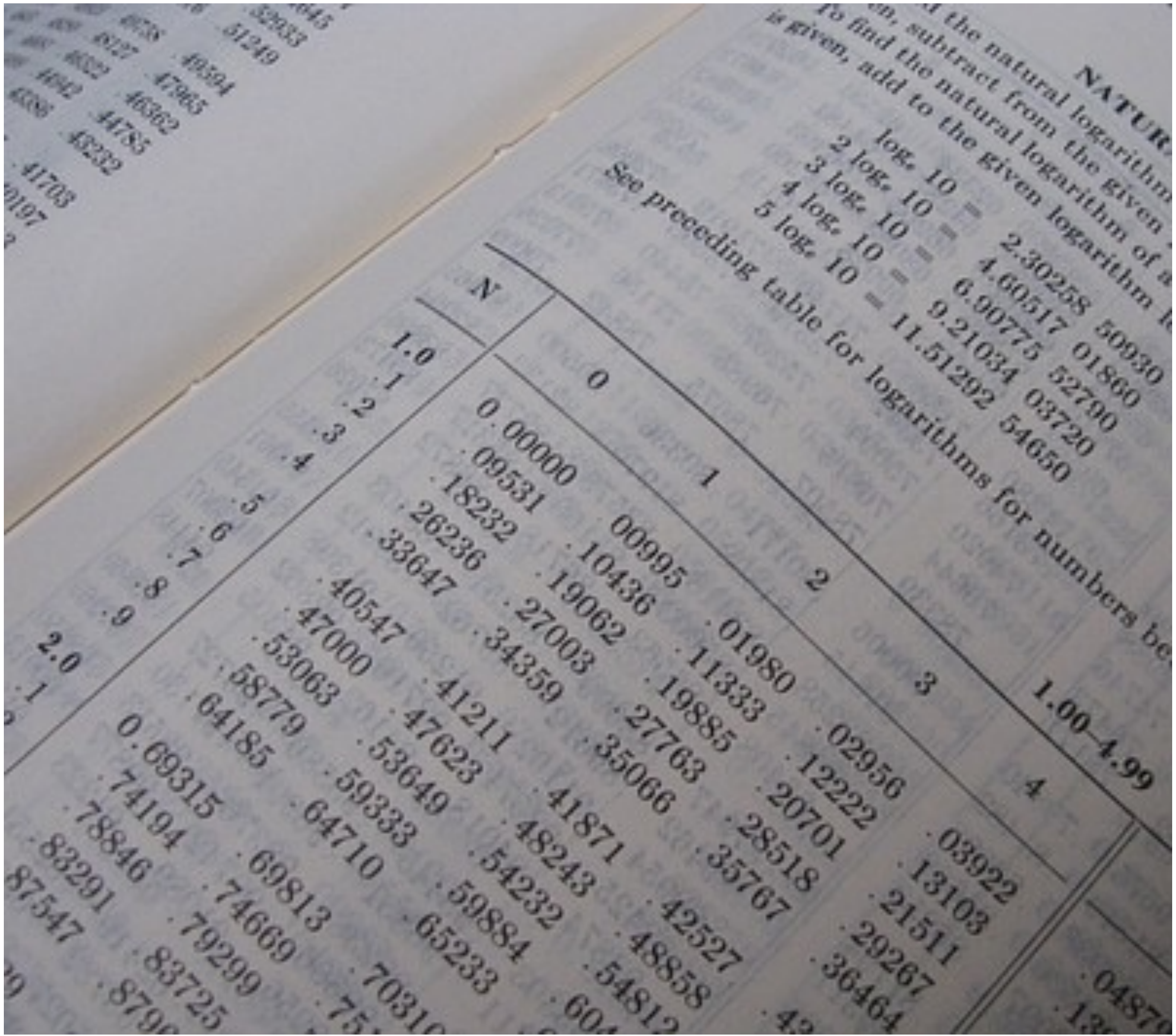
Figure 5: Number of triples in KV in each confidence bin.

Outline

- WebTables
  - How do we find the good tables on the web?
- Knowledge Vault
  - How do we find other quality facts on the web?
- Lattice Regression
  - How do we form interpretable non-linear models?

Lattice Regression

- Interpolated look-up tables as function class



NATUR

d the natural logarithm

en, subtract from the given

To find the natural logarithm of a

is given, add to the given logarithm

log_e 10 = 2.30258 50930
2 log_e 10 = 4.60517 01860
3 log_e 10 = 6.90775 52790
4 log_e 10 = 9.21034 03720
5 log_e 10 = 11.51292 54650

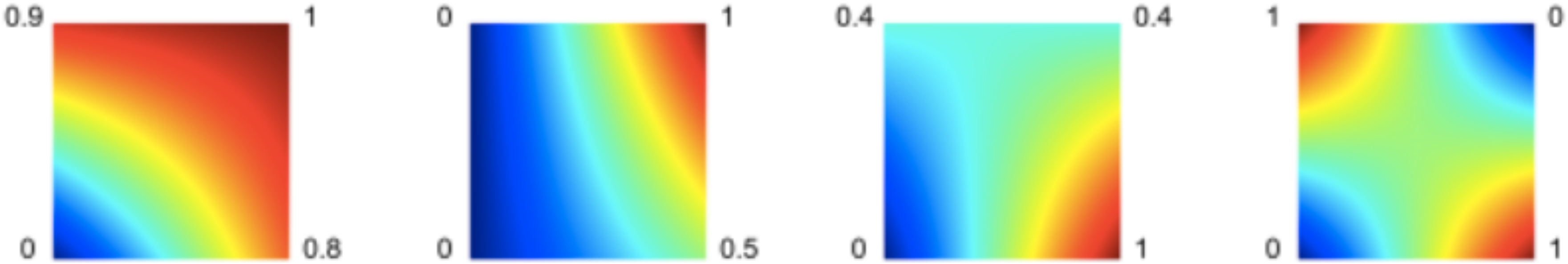
See preceding table for logarithms for numbers be

1.00–4.99

N	0	1	2	3	4	
1.0	0.00000	00995	·01980	·02956	·03922	·04870
·1	·09531	·10436	·11333	·12222	·13103	·13
·2	·18232	·19062	·19885	·20701	·21511	
·3	·26236	·27003	·27763	·28518	·29267	
·4	·33647	·34359	·35066	·35767	·36464	
·5	·40547	·41211	·41871	·42527	·43	
·6	·47000	·47623	·48243	·48858		
·7	·53063	·53649	·54232	·5481		
·8	·58779	·59333	·59884	·60		
·9	·64185	·64710	·65233			
2.0	0.69315	·69813	·7031			
·1	·74194	·74669	·75			
·2	·78846	·79299				
·3	·83291	·83725				
·4	·87547	·879				

Lattice Regression

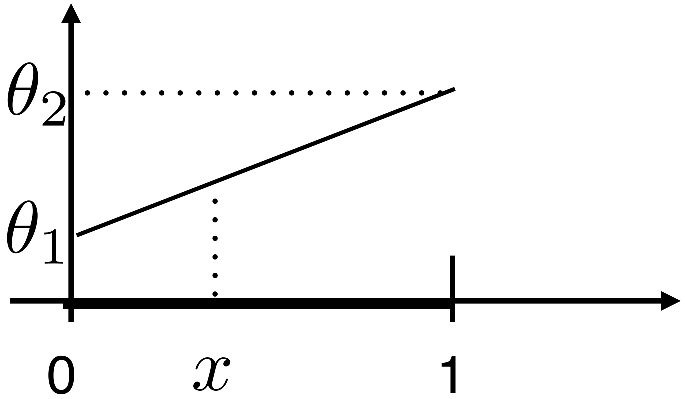
- Interpolated look-up tables as function class
  - low-dimensional functions
  - determine values at corners, linearly interpolate in between:



  - for draft paper on monotonic lattice regression see mayagupta.org

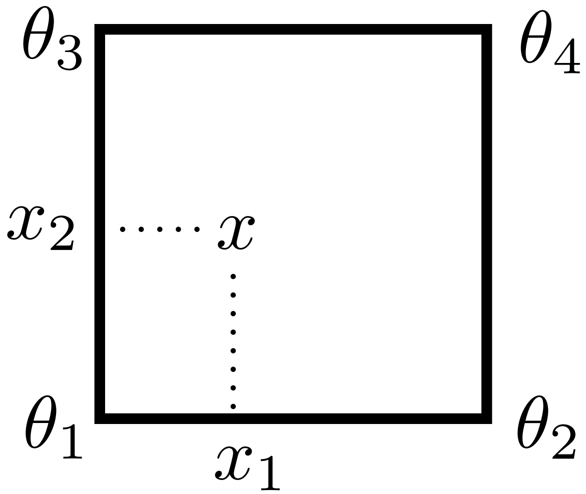
Linear Interpolating in D Dimensions

- Linear interpolating in 1 dimension



$$f(x) = \theta_1 + (\theta_2 - \theta_1)x = \theta_1(1-x) + \theta_2 x$$
$$= \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \begin{pmatrix} 1-x \\ x \end{pmatrix}$$

- In 2 dimensions

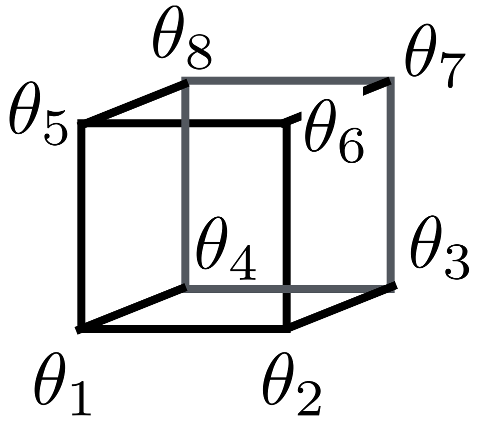


$$f(x) = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \begin{pmatrix} (1-x_1)(1-x_2) \\ x_1(1-x_2) \\ (1-x_1)x_2 \\ x_1 x_2 \end{pmatrix}$$

$$= \begin{pmatrix} \theta_1(1-x_2) + \theta_3 x_2 \\ \theta_2(1-x_2) + \theta_4 x_2 \end{pmatrix} \begin{pmatrix} 1-x_1 \\ x_1 \end{pmatrix}$$

Lattice Regression in D Dimensions

- In dimension D the lattice has 2^D vertices



$$f(\mathbf{x}) = \theta\phi(\mathbf{x}), \qquad \theta, \phi \in \mathbf{R}^{\mathbf{2^D}}$$

- Monotonicity:

$$\theta_2 \geq \theta_1 \qquad \theta_4 \geq \theta_1 \qquad \theta_5 \geq \theta_1$$

  - D constraints per vertex, 2^D vertices
  - $D2^{(D-1)}$ monotonicity constraints

Lattice Regression

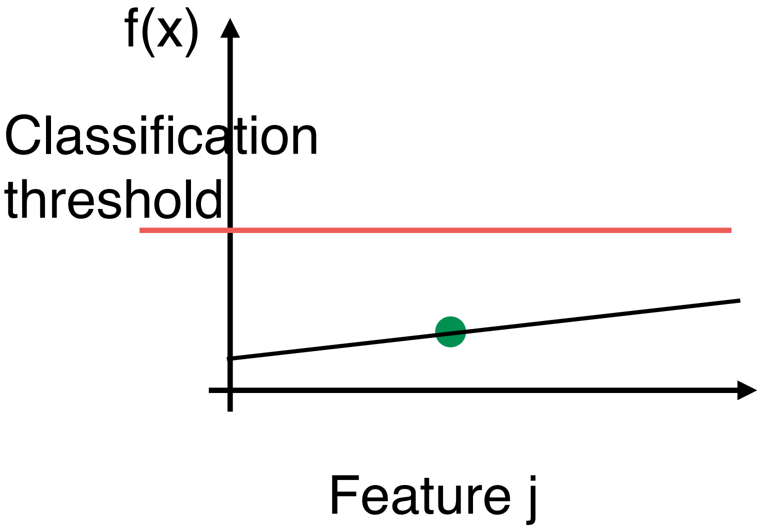
- Loss function

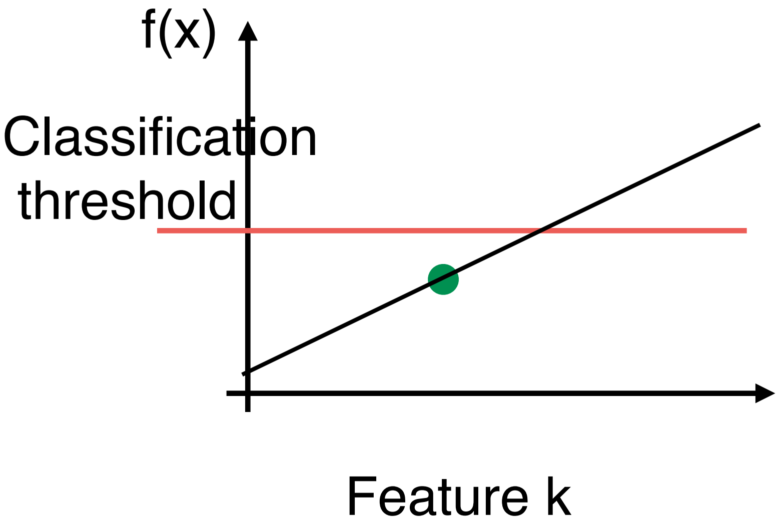
$$\operatorname*{argmin}_{\theta} \sum_{i=1}^{m} l(y_i, \theta\phi(x_i)) + R(\theta), \qquad \text{s.t.} \qquad A\theta \leq b$$

  - cost measured in Mean Squared Error
  - R is a regularizer
  - A represents all the constraints

Lattice Regression

- Examining the output function for a misclassified point:
  - what score is it worth improving?





Outline

- WebTables
  - How do we find the good tables on the web?
- Knowledge Vault
  - How do we find other quality facts on the web?
- Lattice Regression
  - How do we form interpretable non-linear models?