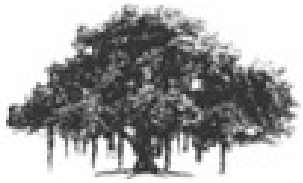


IIIT-H

Web Mining
Lecture 19: Entity Semantics Mining (Part 1)

Manish Gupta

5th Oct 2013

Slides borrowed (and modified) from
Kaushik Chakrabarti's tutorial "Simple Models, Lots of Data: Mining semantics about entities using Web-Scale Data" at Joint International Workshop on Entity-oriented and Semantic Search (JIWES) 2012 @ SIGIR Conference

Recap of Lecture 18: Mining Structured Information from the Web (Part 2)

- Extracting Top-K Lists from the Web
- Extracting Web Data Records Containing User-Generated Content
- Extracting Tables from the Web
  - WebTables: Exploring the Power of Tables on the Web
  - Answering Table Augmentation Queries from Unstructured Lists on the Web
  - Annotating Tables with Ontological Links
- Extracting Sets from the Web

Announcements

- Start working on projects

Today's Agenda

- Entity Synonyms
- Entity Attribute Discovery and Augmentation
- Entity Linking

Machines Perform Complex Tasks

- Today, computers are doing complex, "human" tasks
  - Machine translation
  - Speech recognition
  - Computer Vision
  - Spelling correction
  - Web Ranking
  - Spam Filtering
  - **Entity Semantics Mining**
  - …

- No neat, concise algorithms ☹
  - Cannot be explained by neat formulas like F=ma
  - Or elegant algorithm like Dijkstra's shortest path algorithm

Initial Attempts: Using Algorithms, Rules and Heuristics

- Machine translation
  - Hand-coded rules that capture underlying structure of language
- Speed recognition
  - Rules, heuristics and signal processing tricks
- Named entity recognition
  - Hand-crafted grammars by experienced computation linguists
- …

Initial Attempts: Using Algorithms, Rules and Heuristics

- Not driven by data
- Why?
  - Belief: human intelligence can be encapsulated in algorithms and hand-crafted rules
  - Data was not available, computers did not have capacity (kilobytes of memory, megabytes of disk)

Emergence of Data-driven Systems

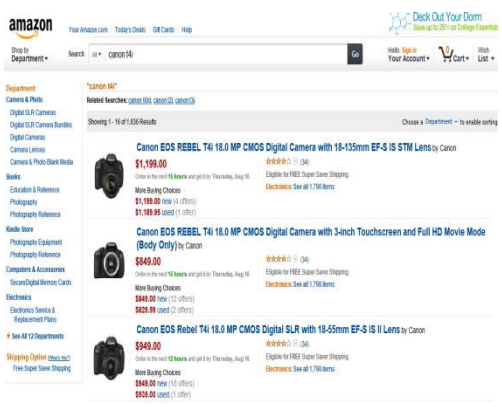
- Machine translation
  - Large training set of input-output pairs available
  - Initially, data-driven systems were worse (till early 2000s)
  - Rapid progress in quality: getting better with more training data, already exceeds best rule-based systems
  - Can expand to new domains rapidly
- Speech recognition
  - Large training sets available
  - Initially, data-driven methods (HMMs) were worse (till late 1980s)
  - Rapid progress in quality with more training data; have become the norm

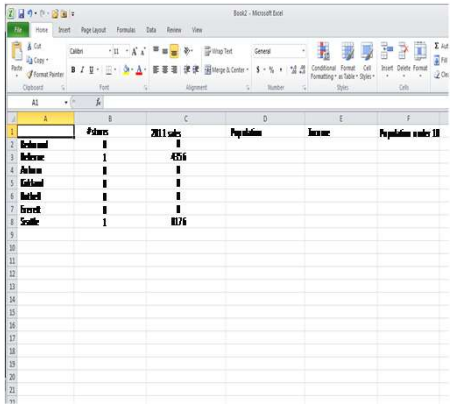
Why Mine Semantics about Entities?

- Entities are everywhere

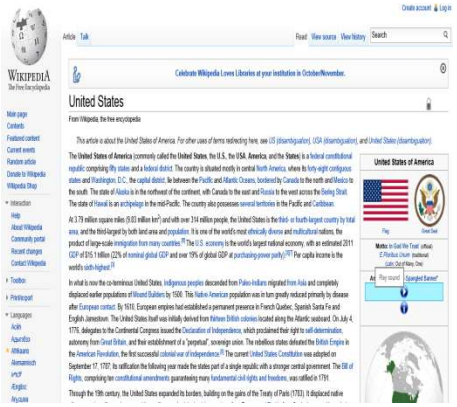
- Data
  - Most data on the web and inside the enterprise are about entities
  - Product catalogs in e-tailers
  - Tables on the web
  - Tables within an enterprise
  - Knowledge bases (Wikipedia, Freebase)
  - …

- Tasks
  - Most consumer and enterprise tasks are about entities
  - Searching for products, movies, places, hotels, restaurants, ….
  - Business Data Analysis
  - Social Media Analytics
  - …





City	2011 Estimate	2010 Census	2000 Census	County
Aberdeen	16,835	16,896	16,461	Grays Harbor
Airway Heights	6,138	6,114	4,500	Spokane
Algona	3,074	3,014	2,460	King
Anacortes	15,941	15,778	14,557	Skagit
Arlington	18,154	17,926	11,713	Snohomish
Asotin	1,270	1,251	1,095	Asotin
Auburn	71,517	70,180	40,314	King
Bainbridge Island	23,262	23,025	20,308	Kitsap
Battle Ground	17,893	17,571	9,296	Clark
Bellevue	124,798	122,363	109,569	King
Bellingham	81,862	80,885	67,171	Whatcom
Benton City	3,134	3,038	2,624	Benton
Bingen	724	712	672	Klickitat
Black Diamond	4,237	4,151	3,970	King
Blaine	4,744	4,684	3,770	Whatcom
Bonney Lake	17,579	17,374	9,687	Pierce
Bothell	34,055	33,505	30,150	King



Earlier Efforts Towards Entity Semantics

- Entity recognizers
  - Hand-crafted rules by experienced computational linguists
- Entity synonyms
  - Manually discovered by domain-experts
- Entity information gathering (attribute discovery, augmentation)
  - Writing wrappers for specific web sites
- Not data driven: used rules, heuristics rather than data

Entity Semantic Mining Tasks

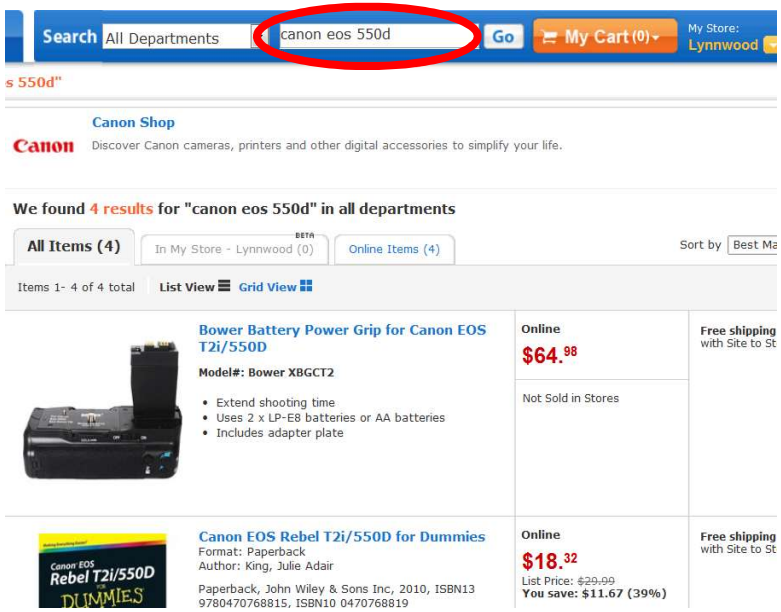
- Given an entity (or set of entities), what are other ways it is referred to?
  - **Entity Synonyms**
- Given a set of entities, what are the interesting attributes of these entities?
  - **Entity Attribute Discovery**
- Given a set of entities and an attribute, what are values of the entities on that attribute?
  - **Entity Augmentation**
- Given a set of entities and a text corpus, what are the semantic mentions of the entity in the corpus?
  - **Entity Linking**

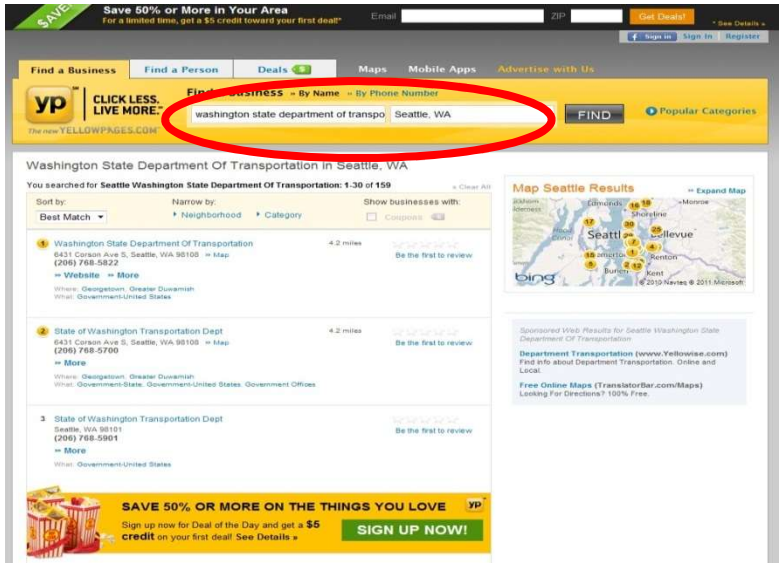
Today's Agenda

- **Entity Synonyms**
- Entity Attribute Discovery and Augmentation
- Entity Linking

Why Discover Entity Synonyms?

- Users search for entities:
  - E-tailers like Walmart:
    - allow users to search for products
  - Information providers like YellowPages:
    - allow users to search for local businesses
  - Online music stores like iTunes and Zune Marketplace:
    - allow users to search for songs/albums/artists





Why Discover Entity Synonyms?

- Powered by enterprise search engines like FAST and Endeca

Entity	Alternative ways to refer to that entity
canon 400d	canon rebel xti, cannon 400d
harry potter and the half blood prince	harry potter 6, half blood prince, harry potter and the half blooded prince
washington state department of licensing	wa dol, dol washington, wa dmv, wash state dept of licensing
university of washington	uw, uofw, univ of wa, univ of washington

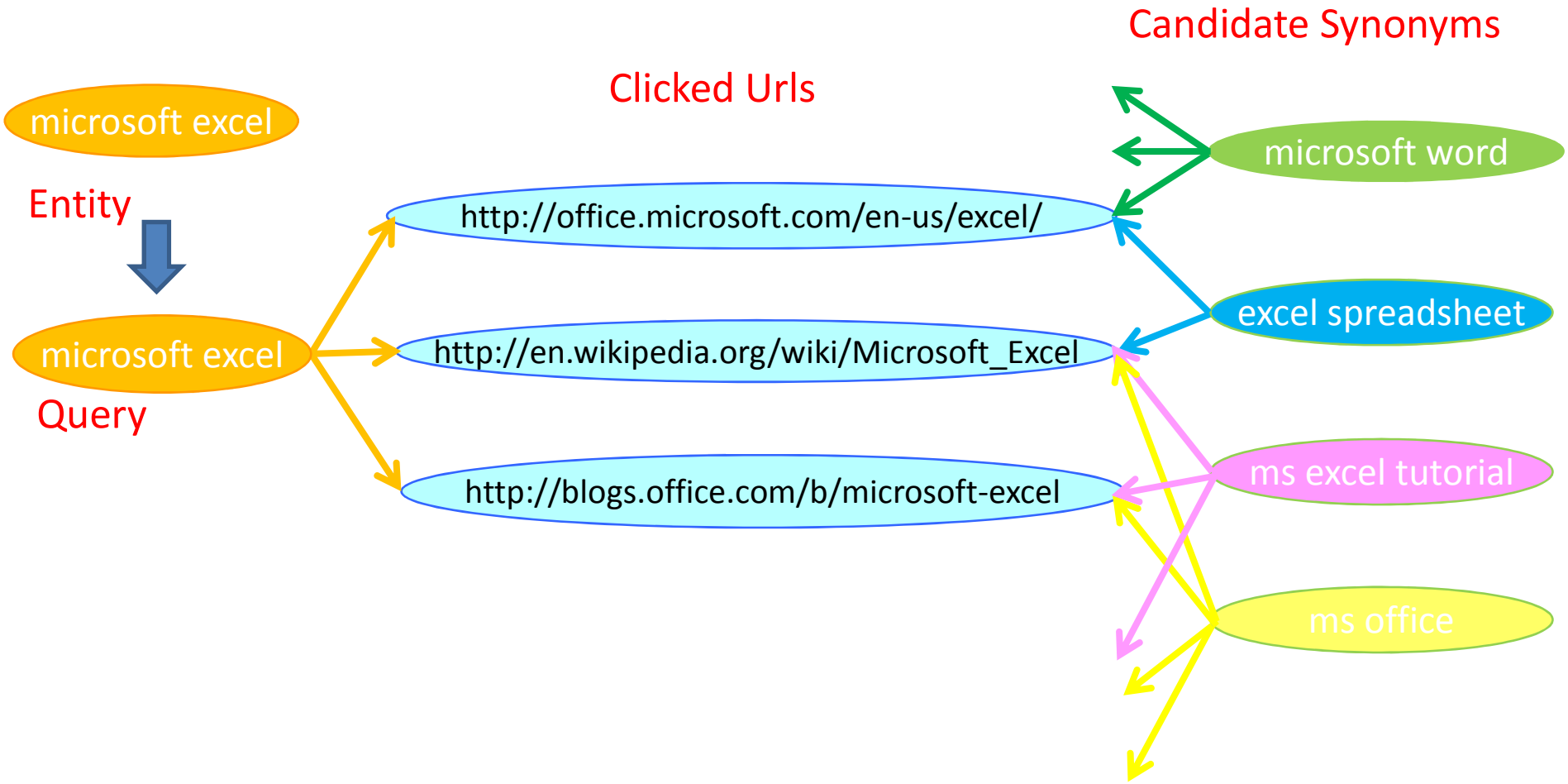
- Without this knowledge, search fails to return relevant results
  - Bad user experience, loss of web site conversions and sales, customer attrition
- Cannot treat them as strings, need to capture semantic relationship

Discovering Synonyms

- Previous approach: domain experts provide synonyms manually
  - Tedious, costly: e-tailers have thousands, even millions of entities
  - New entities keep coming, old entities get obsolete, hard to keep up

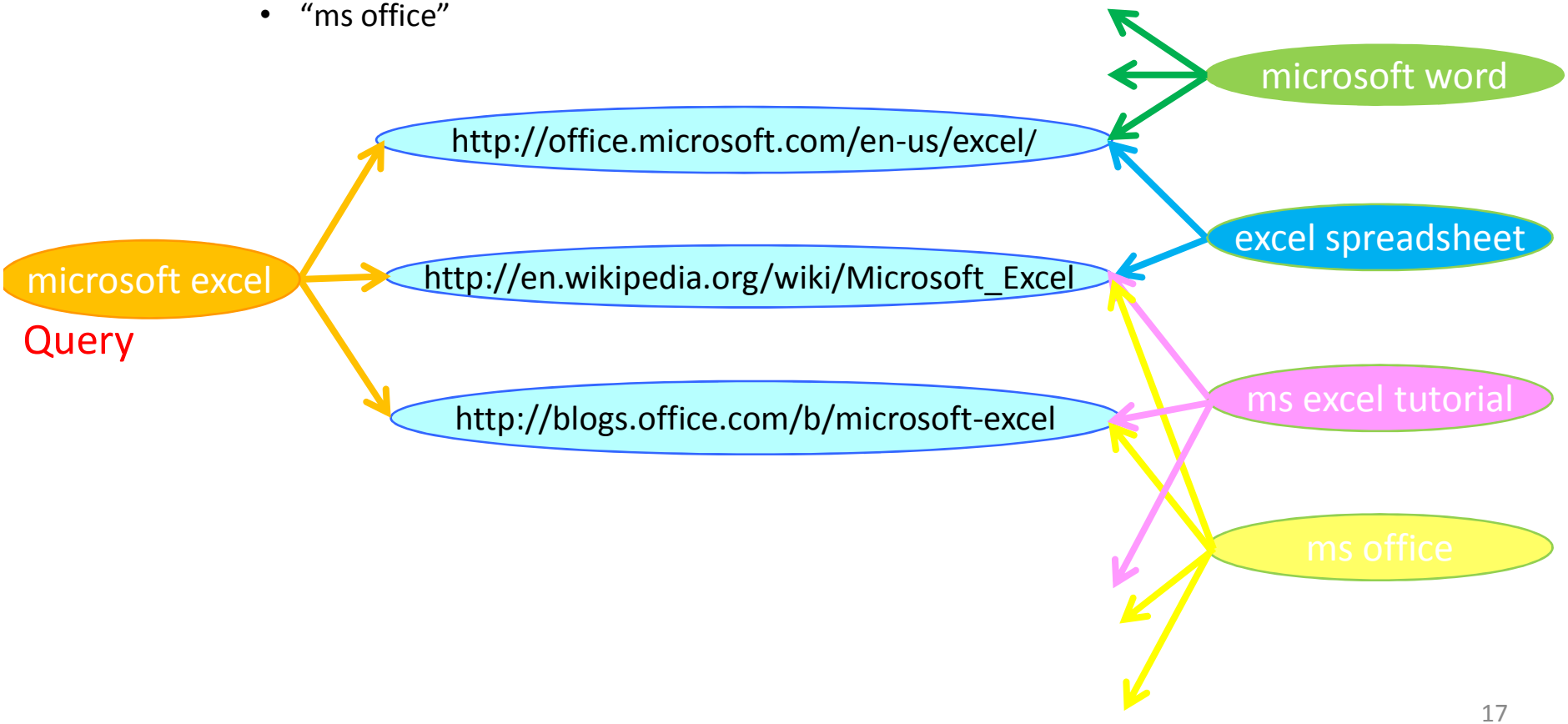
- Main Insight:
  - Search engine's query click log captures this => if two different search queries refer to same entity, significant overlap among the links clicked for the two queries.
  - Can we leverage it to find synonyms automatically?
  - Requirements:
    - Domain independent, high quality, scalable algorithm

Step 1: Generation of Candidate Synonyms

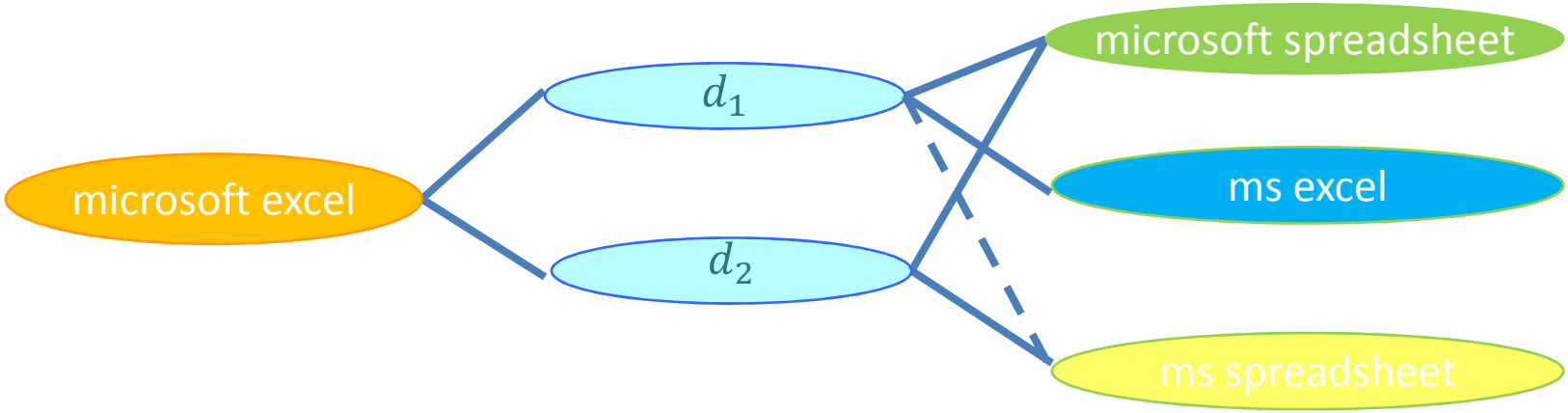


Step 2: Filtering of Candidate Synonyms

- Look for *strong* semantic relationship in *both directions*
  - Filter out candidates that are weakly related
    - "microsoft word"
  - Filter out candidates that are not exclusively related
    - "ms office"



Not good enough: Click log sparsity



- Pseudo-document similarity ("A Framework for Robust Discovery of Entity Synonyms", KDD 2012)

Not good enough: Not of same semantic type

- "microsoft excel" and "ms excel tutorial" are very strongly related but not of the same "type"
  - Software vs. tutorial

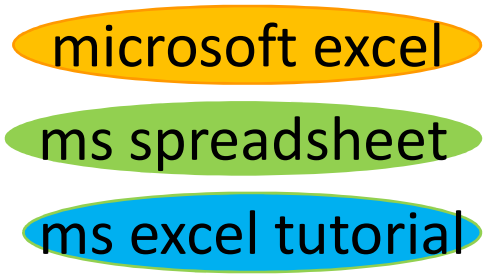
- A synonym (of an entity) should be
  - Strongly related in both directions (measured by pseudo document similarity)
  - Of the same type

Context Similarity

- Intuition
  - If X and Y are entities of the same type, X and Y are exchangeable under various contexts
- Where to find contexts?
  - Query logs
- Examples
  - microsoft excel help⇔ ms spreadsheet help ⇔ powerpoint help
  - microsoft excel download⇔ ms spreadsheet download ⇔ powerpoint download

Construct Context Vectors

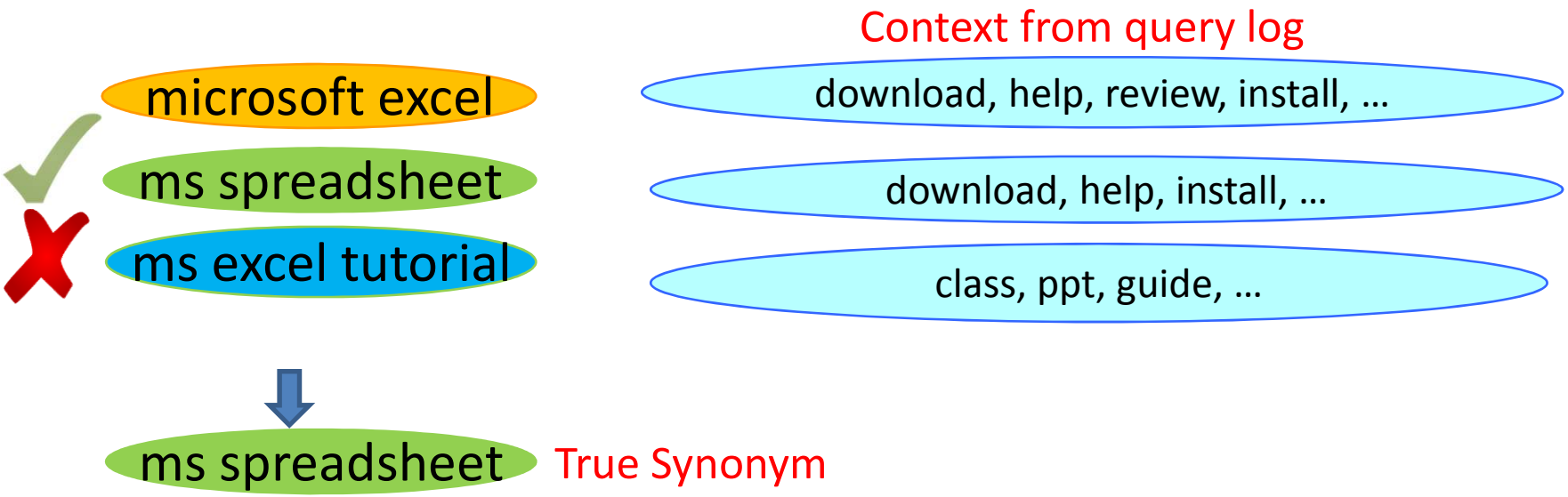
Entities and candidates



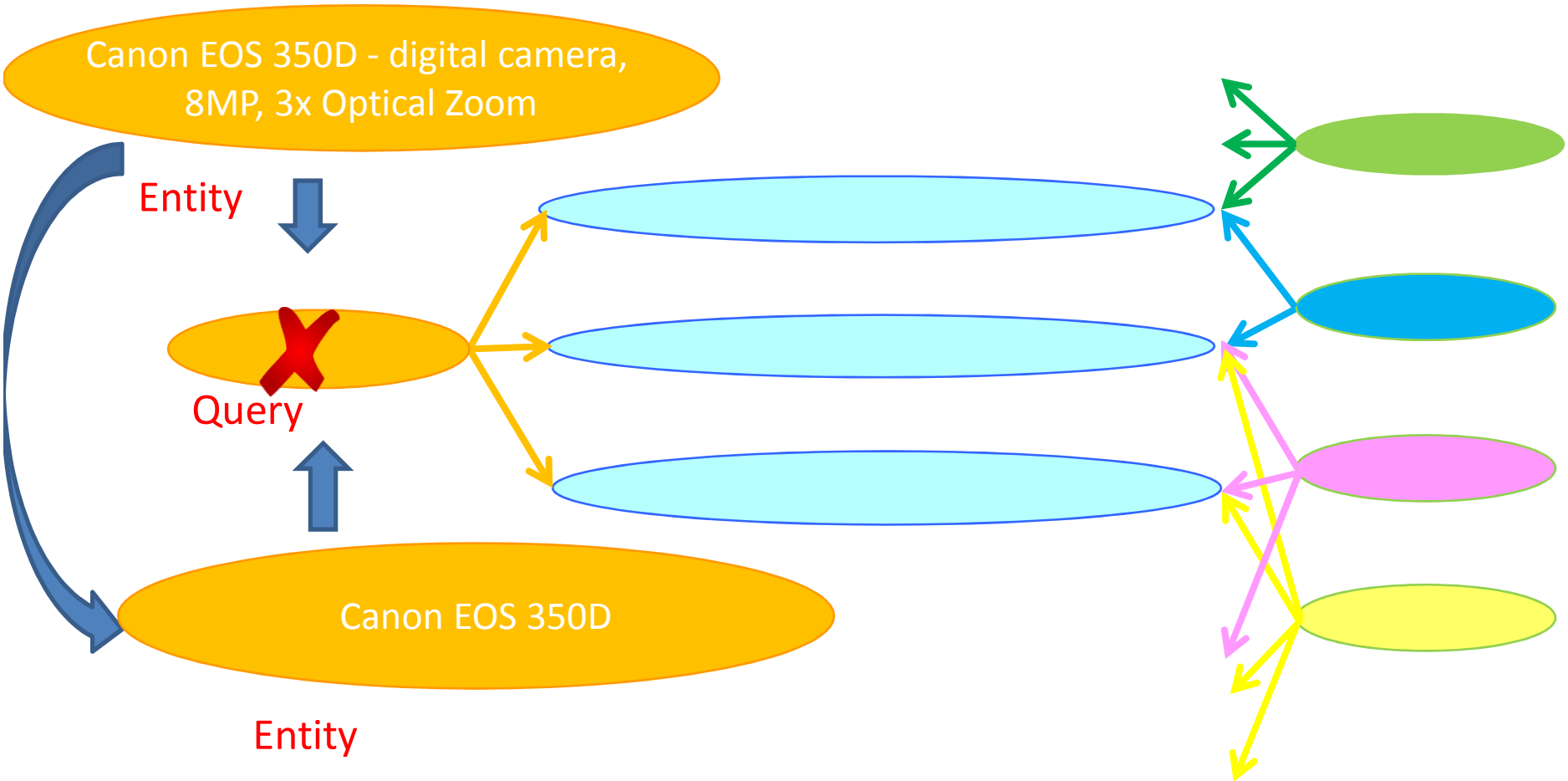
Other Queries
microsoft excel download
microsoft excel help
microsoft excel
ms spreadsheet download
ms spreadsheet help
ms excel tutorial class
ms excel tutorial ppt

Entities/ Candidates	Contexts
microsoft excel	download
microsoft excel	help
microsoft excel	review
ms spreadsheet	download
ms spreadsheet	help
ms excel tutorial	class
ms excel tutorial	ppt

Same type => High Context Similarity



Not good enough: Long entity names



Summary

- Lots of data available anyway
  - Similar to machine translation, speech recognition
  - The main insight is how to connect the problem with that data
- Insight is important, algorithms are relatively simple
  - A classifier with less than 10 features
  - Achieves ~90% precision, ~3 synonyms per entity for products
- Taking care of details important for high accuracy
  - Click log sparsity, context similarity, long entity names
- Need to scale to large query logs
  - MapReduce implementation

- Further details: "A Framework for Robust Discovery of Entity Synonyms", KDD 2012

Today's Agenda

- Entity Synonyms
- **Entity Attribute Discovery and Augmentation**
- Entity Linking

Information gathering

- Need to gather information for entities to make a decision
  - Consumer space:
    - Research for products (say, cameras, cars, appliances)
    - Research for stocks
    - ….
  - Enterprise space:
    - Information workers, business data analytics
    - Application developers, mashup creators
- Mostly manual today, very labor intensive
  - Can we automate it?
  - Semantic task: need to understand entities, attributes and values, not just strings

3 APIs for Information Gathering

Augmentation By Attribute Name (ABA)

Augmentation By Example (ABE)

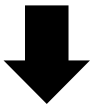
Attribute Discovery (AD)

Input Query Table

Model	Brand
S80	
A10	
GX-1S	
T1460	

S80	Nikon
A10	Canon
GX-1S	
T1460	

S80
A10
GX-1S
T1460



Output Table

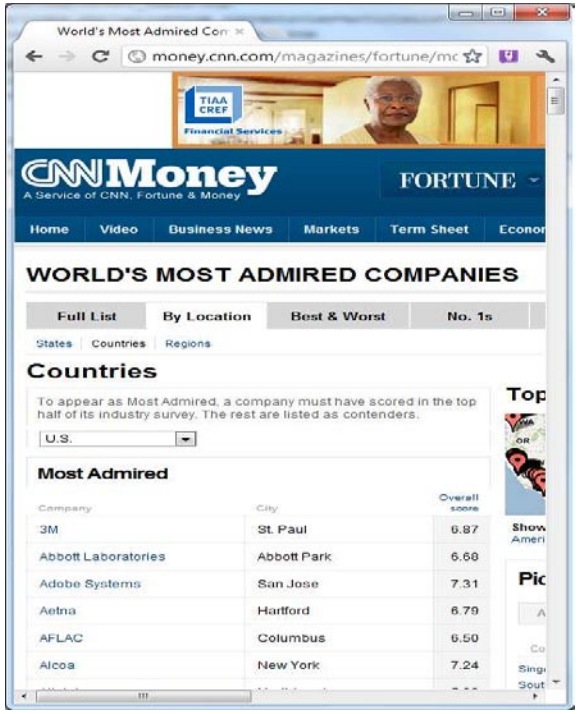
Model	Brand
S80	Nikon
A10	Canon
GX-1S	Samsung
T1460	Benq

S80	Nikon
A10	Canon
GX-1S	Samsung
T1460	Benq

brand | make | manufacturer | mfr
resolution | mp | megapixel | res
price | retail price | offer
zoom | optical zoom

Main Insight

- Tables on the web have a lot of structured information about entities and their attributes
- Can we leverage it?
- This data is available anyway
  - Similar to machine translation, speech recognition, entity synonyms
  - The main insight is to connect the problem with that data



City	2011 Estimate[1]	2010 Census[2]	2000 Census	County
Aberdeen	16,835	16,896	16,461	Grays Harbor
Airway Heights	6,138	6,114	4,500	Spokane
Algona	3,074	3,014	2,460	King
Anacortes	15,941	15,778	14,557	Skagit
Arlington	18,154	17,926	11,713	Snohomish
Asotin	1,270	1,251	1,095	Asotin
Auburn	71,517	70,180	40,314	King
Bainbridge Island	23,262	23,025	20,308	Kitsap
Battle Ground	17,893	17,571	9,296	Clark
Bellevue	124,798	122,363	109,569	King
Bellingham	81,862	80,885	67,171	Whatcom
Benton City	3,134	3,038	2,624	Benton
Bingen	724	712	672	Klickitat
Black Diamond	4,237	4,151	3,970	King
Blaine	4,744	4,684	3,770	Whatcom
Bonney Lake	17,579	17,374	9,687	Pierce
Bothell	34,055	33,505	30,150	King

Microsoft Corporation

Microsoft

Type	Public
Traded as	NASDAQ: MSFT SEHK: 4338 Dow Jones Industrial Average component NASDAQ-100 component S&P 500 component
Industry	Computer software Online services Video games
Founded	Albuquerque, New Mexico, United States (April 4, 1975)
Founder(s)	Bill Gates, Paul Allen
Headquarters	Microsoft Redmond Campus, Redmond, Washington, U.S.

Initial focus

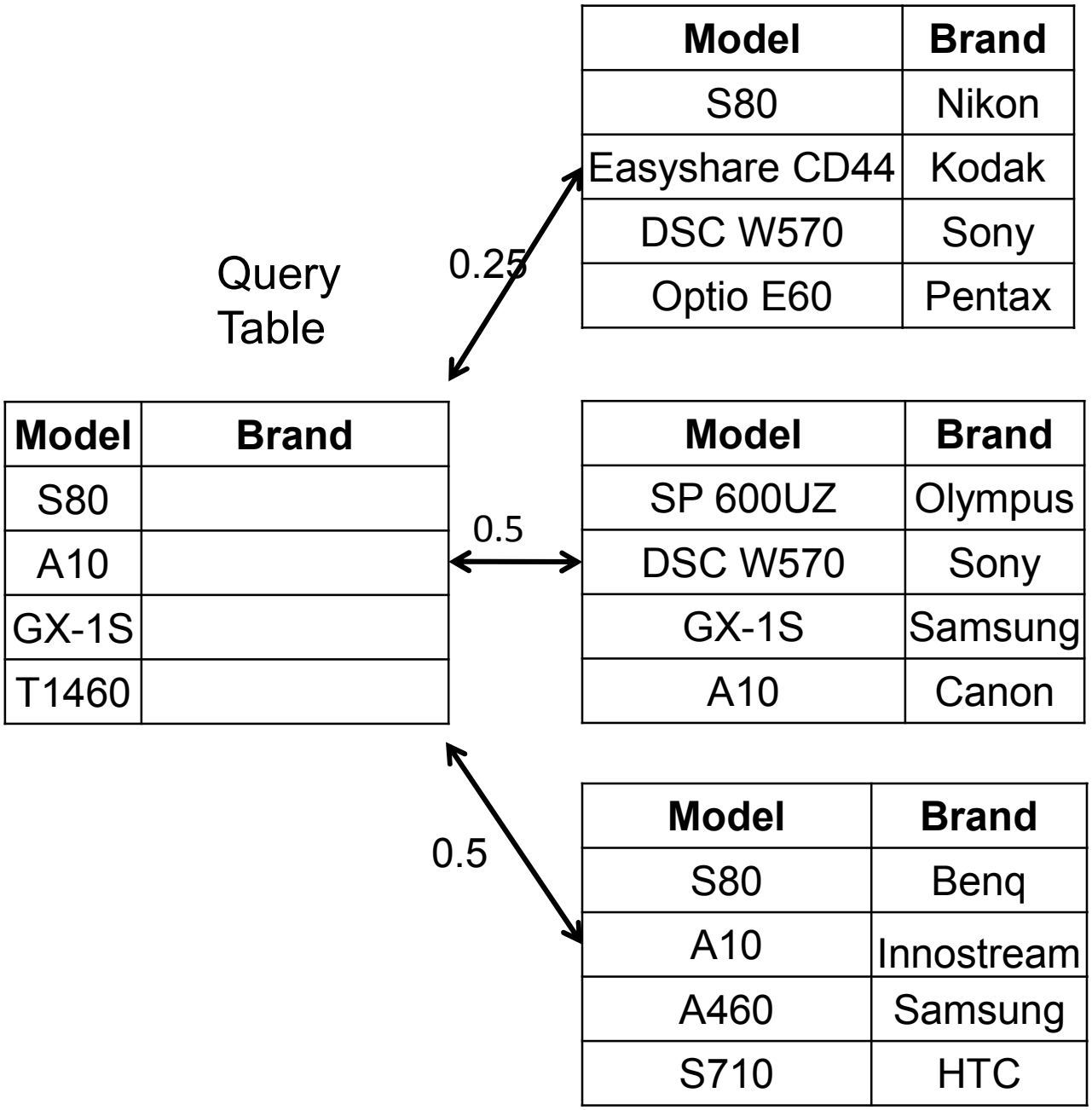
- Focus on relational tables
- Details in paper:
  - "InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables", SIGMOD 2012

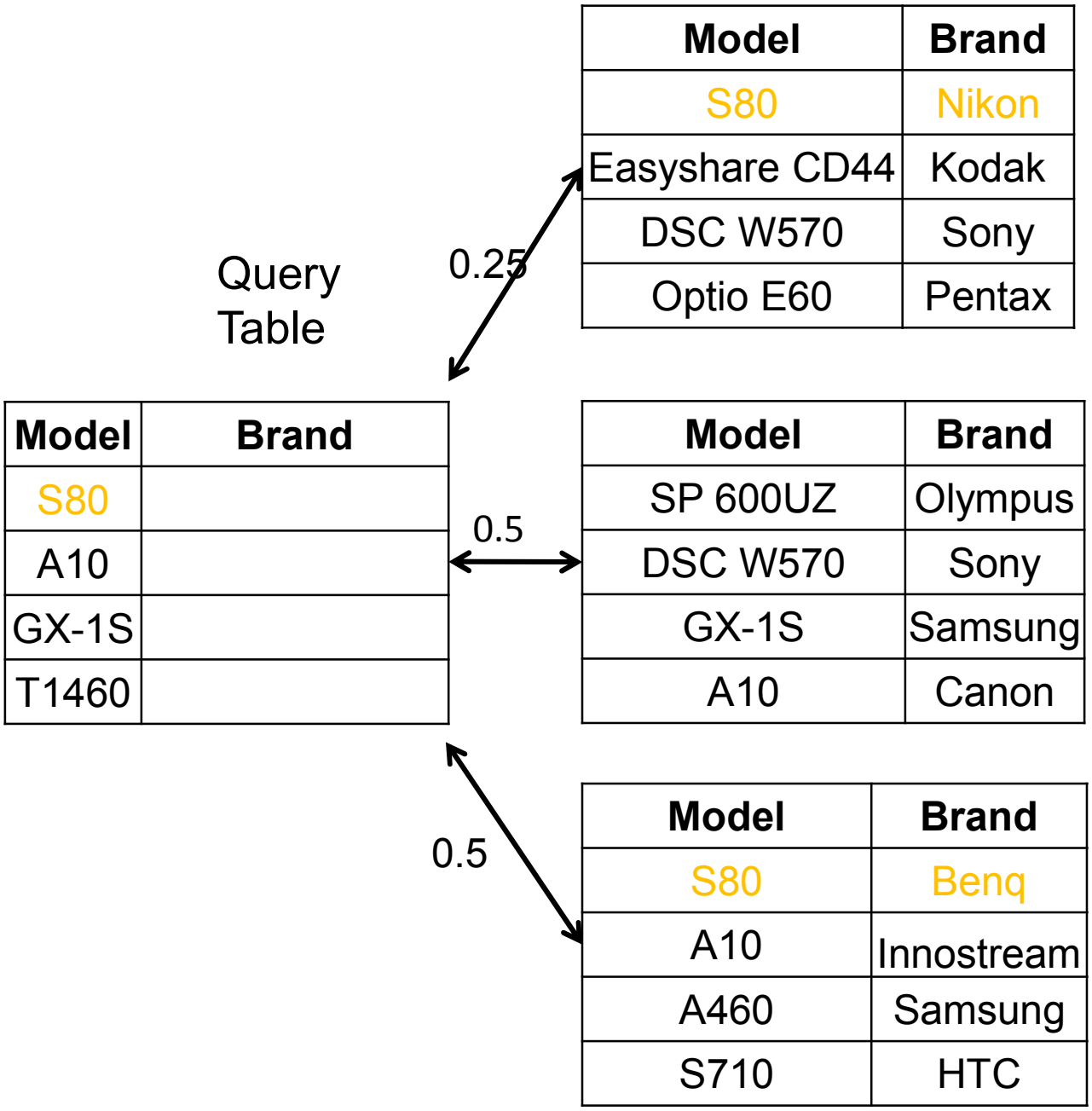
City	2011 Estimate[1]	2010 Census[2]	2000 Census	County
Aberdeen	16,835	16,896	16,461	Grays Harbor
Airway Heights	6,138	6,114	4,500	Spokane
Algona	3,074	3,014	2,460	King
Anacortes	15,941	15,778	14,557	Skagit
Arlington	18,154	17,926	11,713	Snohomish
Asotin	1,270	1,251	1,095	Asotin
Auburn	71,517	70	40,314	King
Bainbridge Island	23,262	025	20,308	Kitsap
Battle Ground	17,893	7,571	9,296	Clark
Bellevue	124,798	122,363	109,569	King
Bellingham	81,862	80,885	67,171	Whatcom
Benton City	3,134	3,038	2,624	Benton
Bingen	724	712	672	Klickitat
Black Diamond	4,237	4,151	3,970	King
Blaine	4,744	4,684	3,770	Whatcom
Bonney Lake	17,579	17,374	9,687	Pierce
Bothell	34,055	33,505	30,150	King

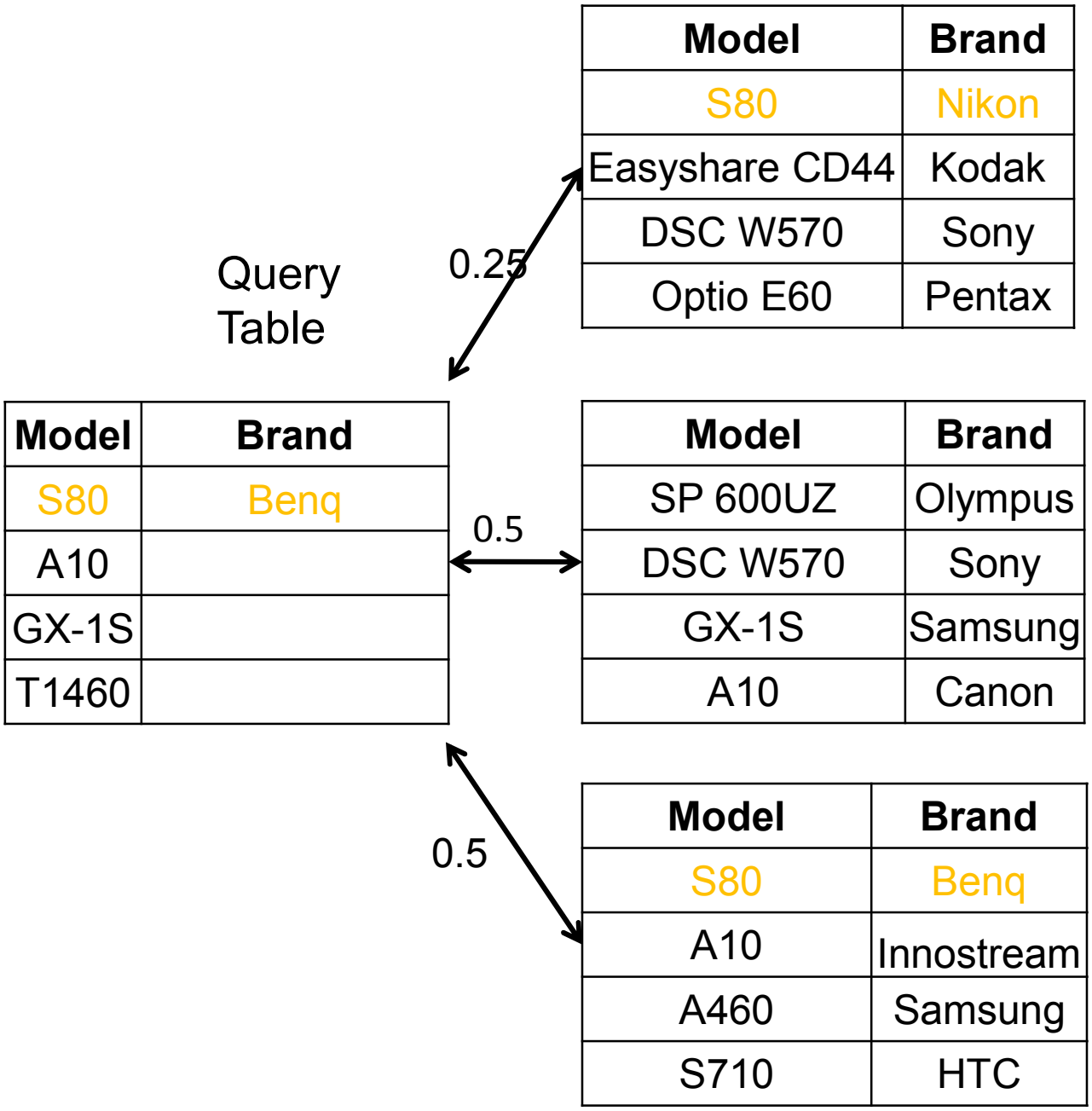
Microsoft Corporation

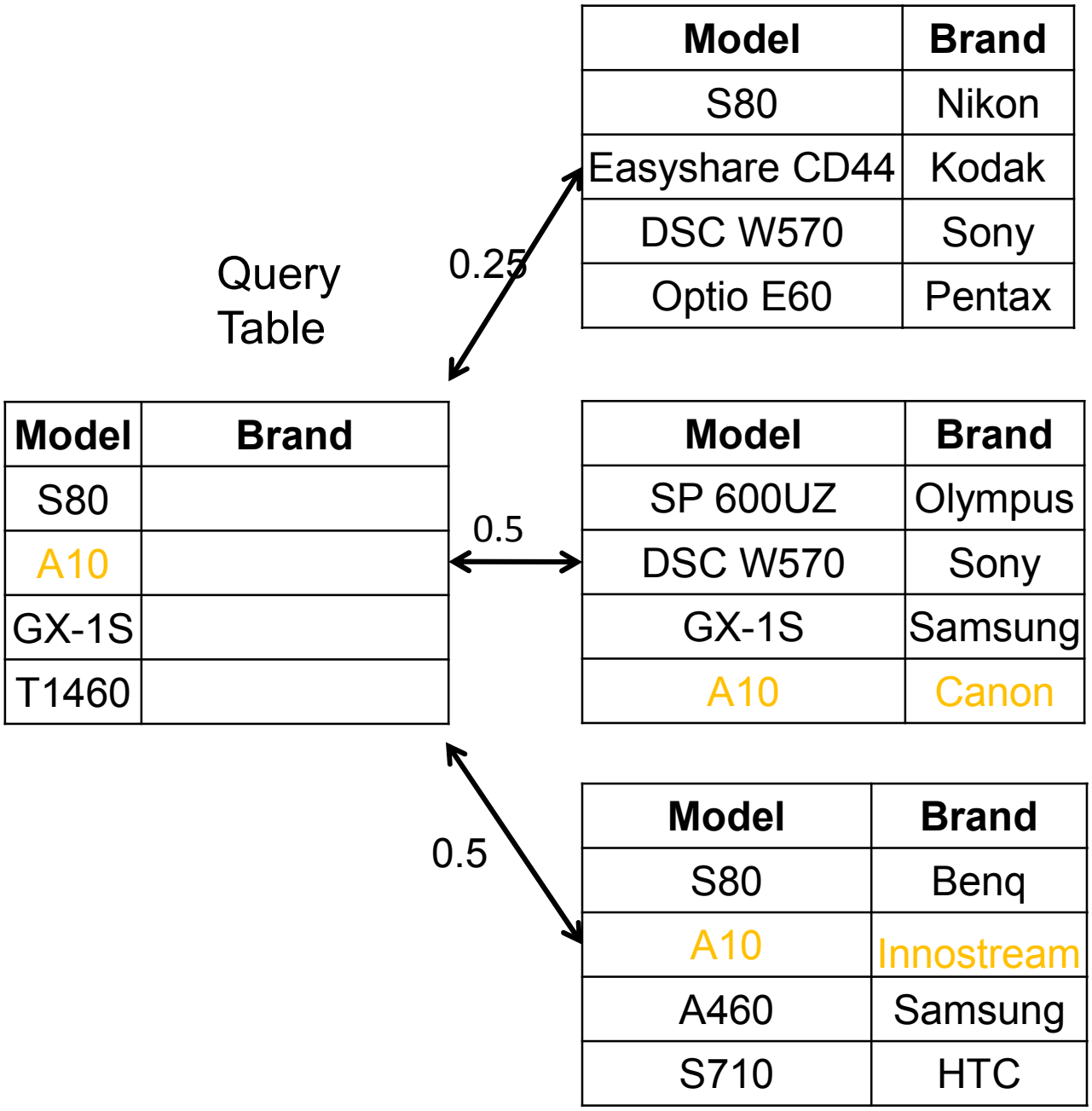
Microsoft

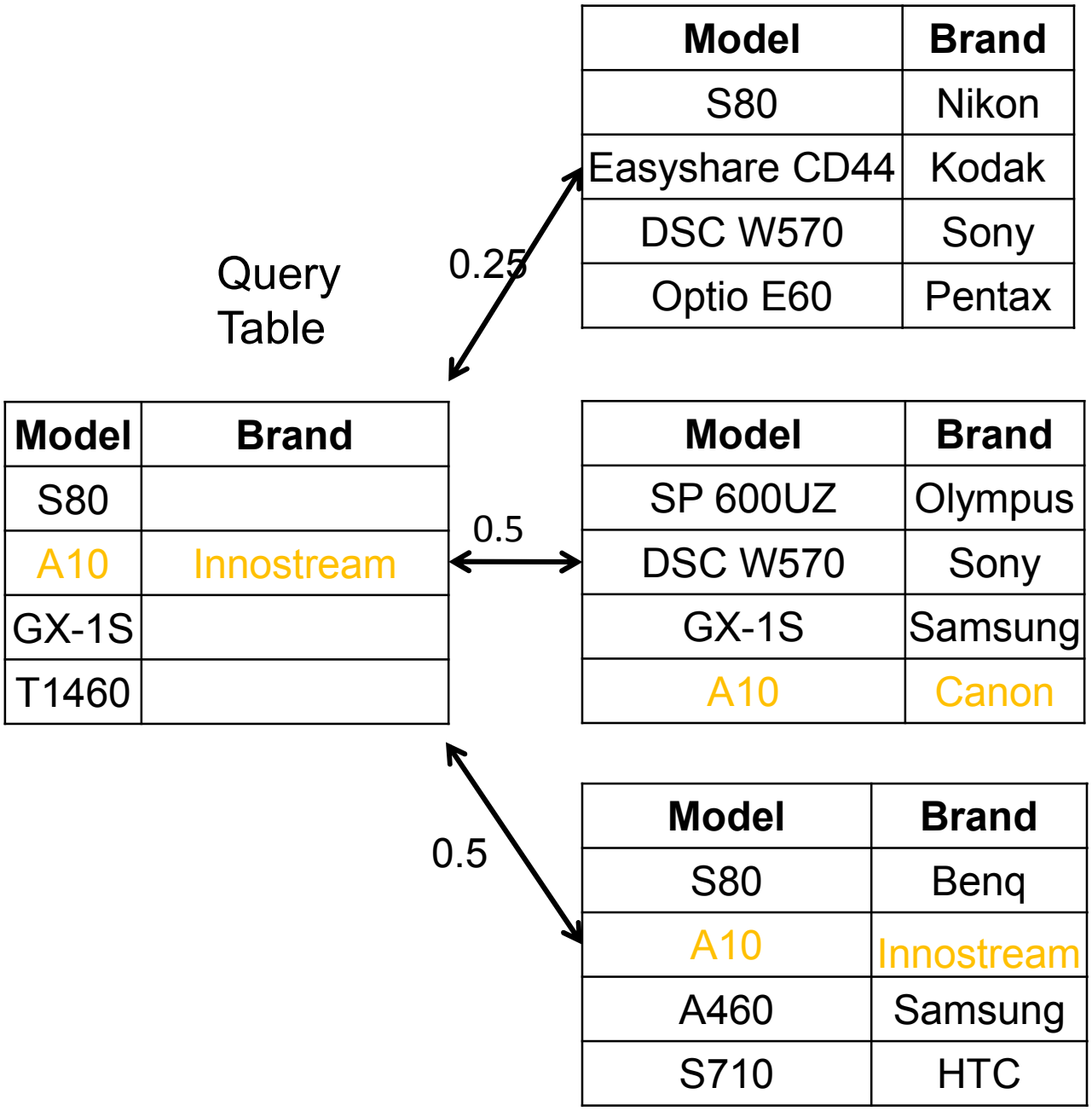
Type	Public
Traded as	NASDAQ: MSFT SEHK: 4338 Dow Jones Industrial Average component NASDAQ-100 component S&P 500 component
Industry	C er software O services ames
Founded	Albuquerque, New Mexico, United States (April 4, 1975)
Founder(s)	Bill Gates, Paul Allen
Headquarters	Microsoft Redmond Campus, Redmond, Washington, U.S.

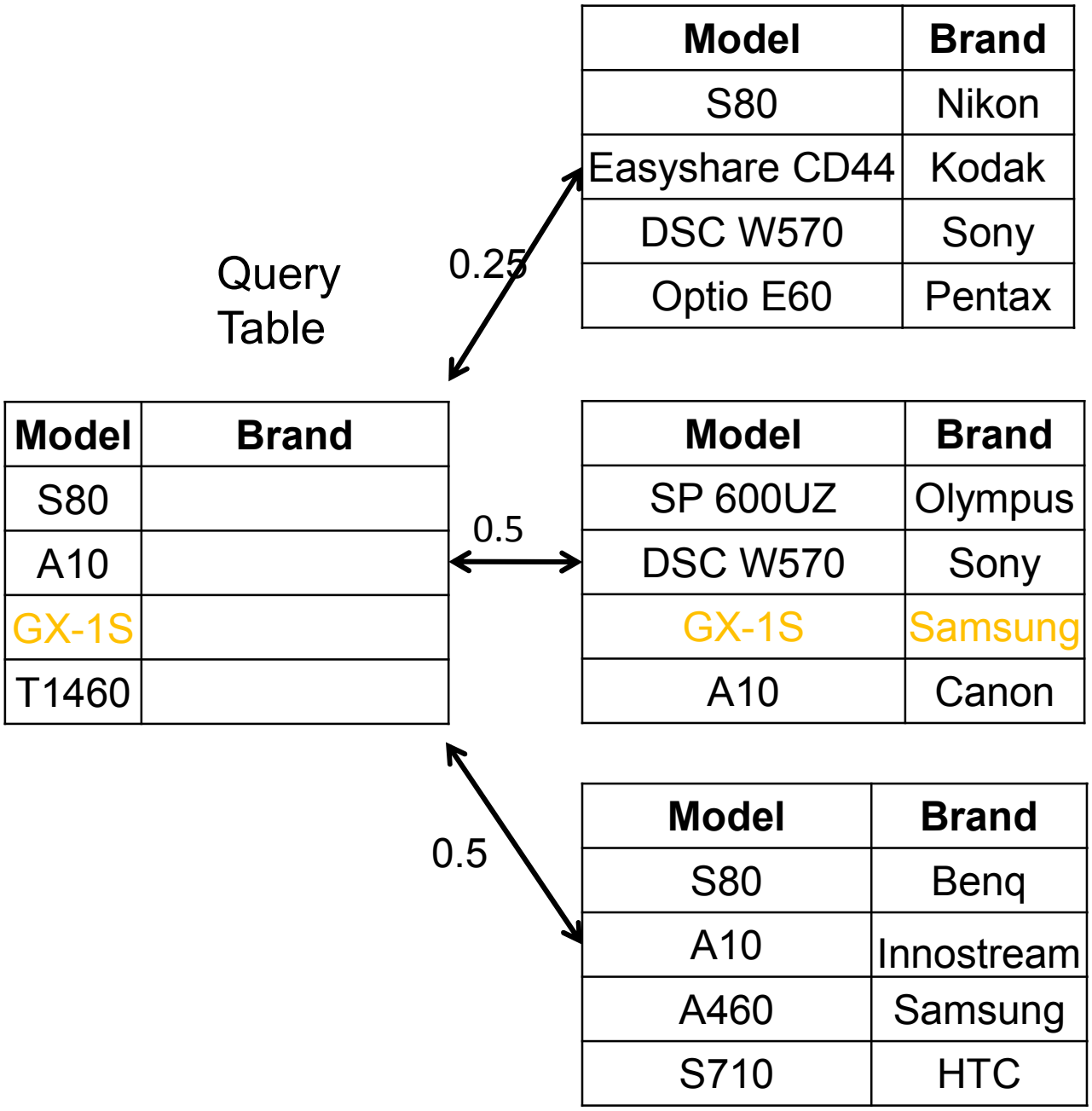

Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
0.25
Query Table
Model
Brand
S80
A10
GX-1S
T1460
0.5
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC

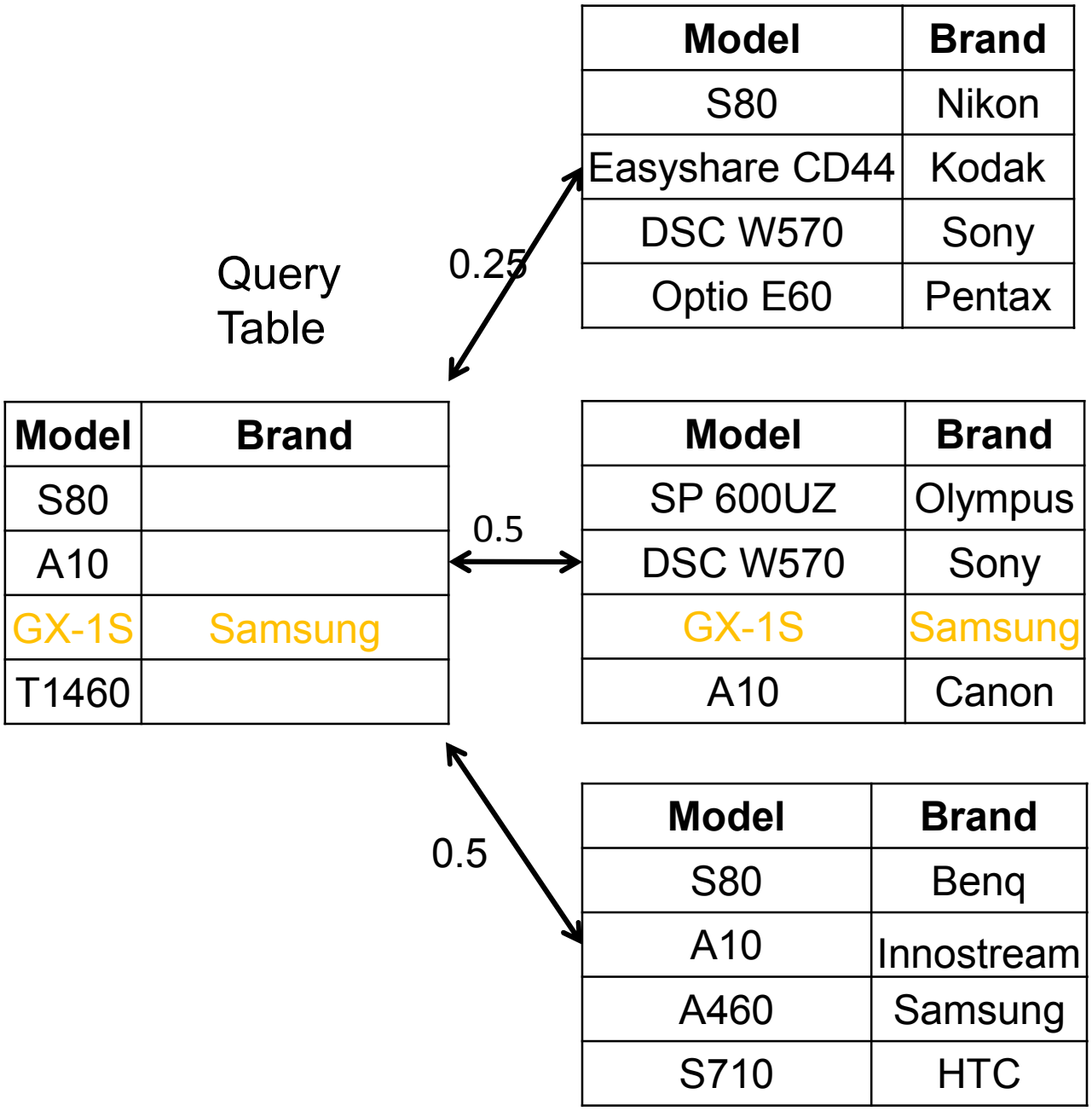

Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
Query Table
0.25
Model
Brand
S80
A10
GX-1S
T1460
0.5
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC


Query Table
0.25
0.5
0.5
Model
Brand
S80
Benq
A10
GX-1S
T1460
Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC


Query Table
0.25
0.5
0.5
Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
Model
Brand
S80
A10
GX-1S
T1460
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC


Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
0.25
Query Table
Model
Brand
S80
A10
Innostream
GX-1S
T1460
0.5
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC


Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
0.25
Query Table
Model
Brand
S80
A10
GX-1S
T1460
0.5
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC


Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
Query Table
0.25
Model
Brand
S80
A10
GX-1S
Samsung
T1460
0.5
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC

Direct Matching Approach (DMA)

Query Table

Model	Brand
S80	Benq
A10	Innostream
GX-1S	Samsung
T1460	

0.25

Model	Brand
S80	Nikon
Easyshare CD44	Kodak
DSC W570	Sony
Optio E60	Pentax

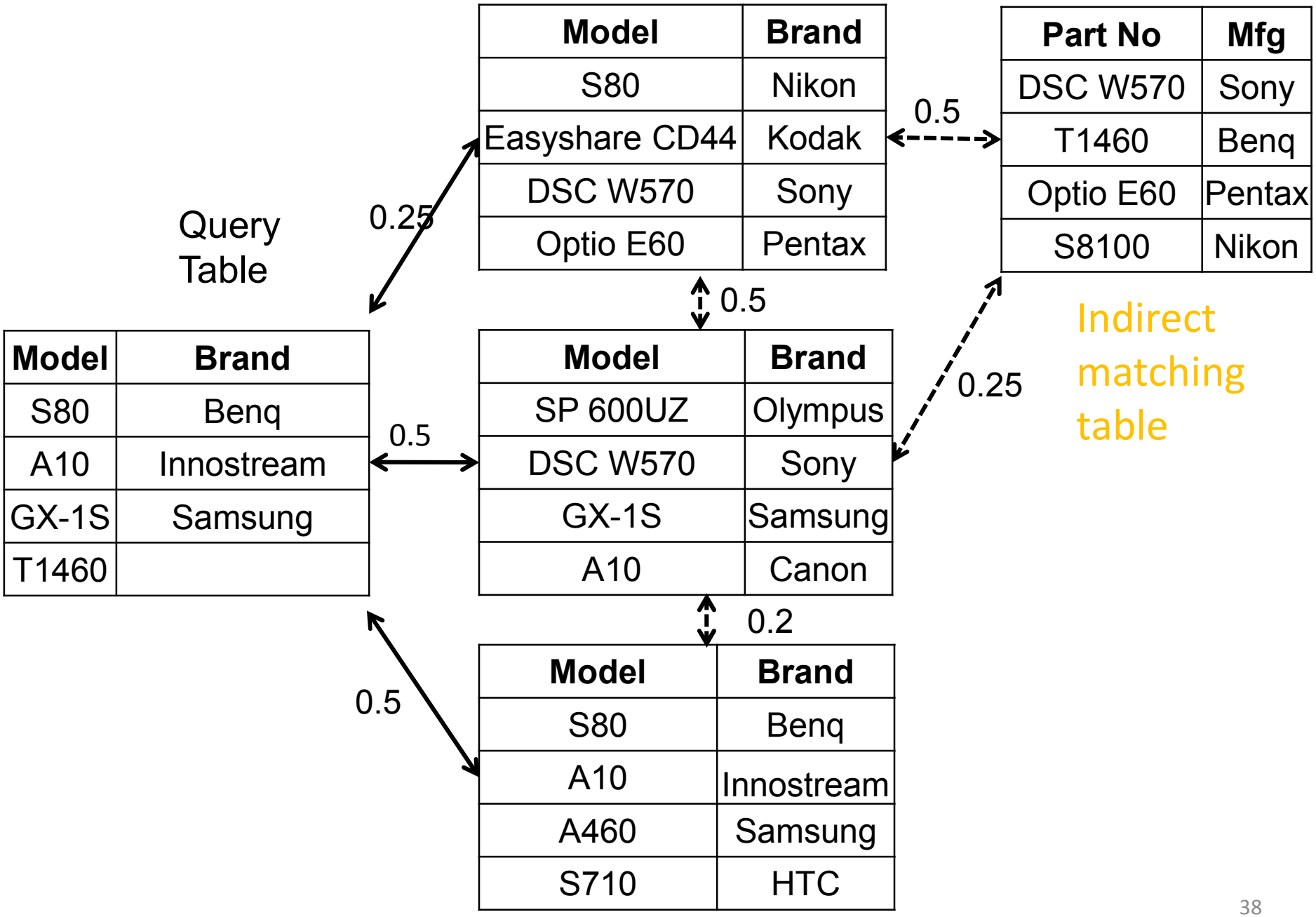
0.5

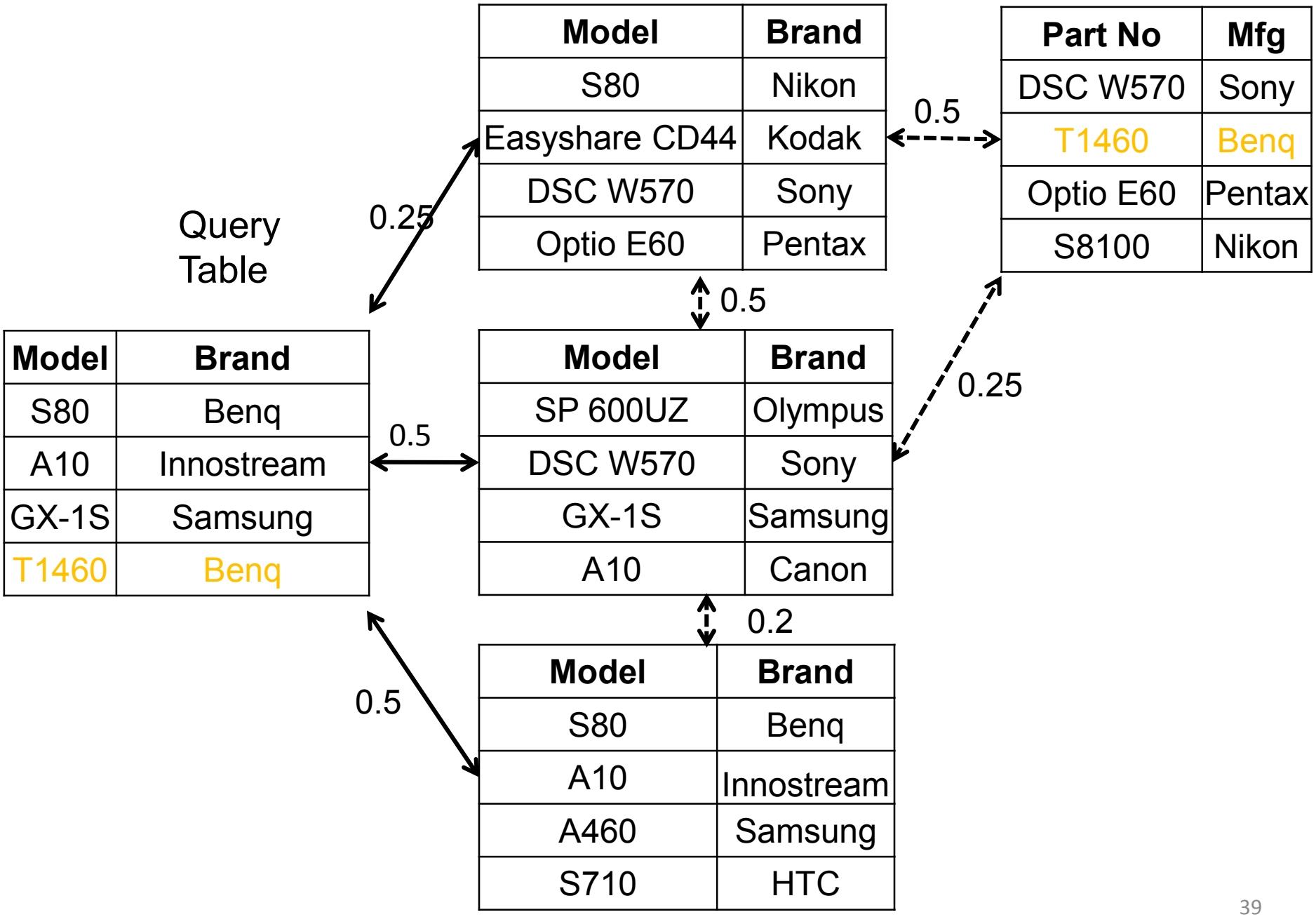
Model	Brand
SP 600UZ	Olympus
DSC W570	Sony
GX-1S	Samsung
A10	Canon

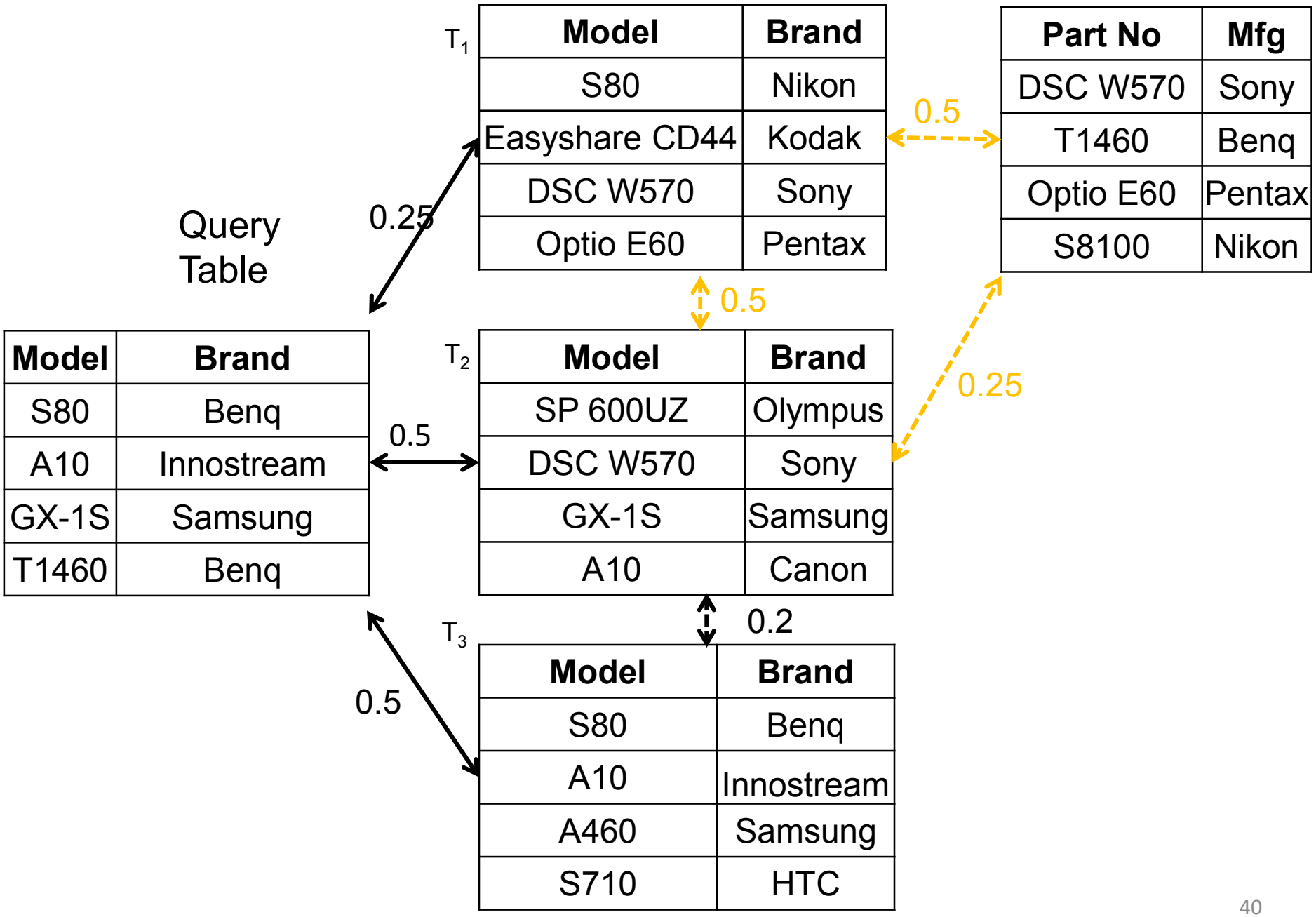
0.5

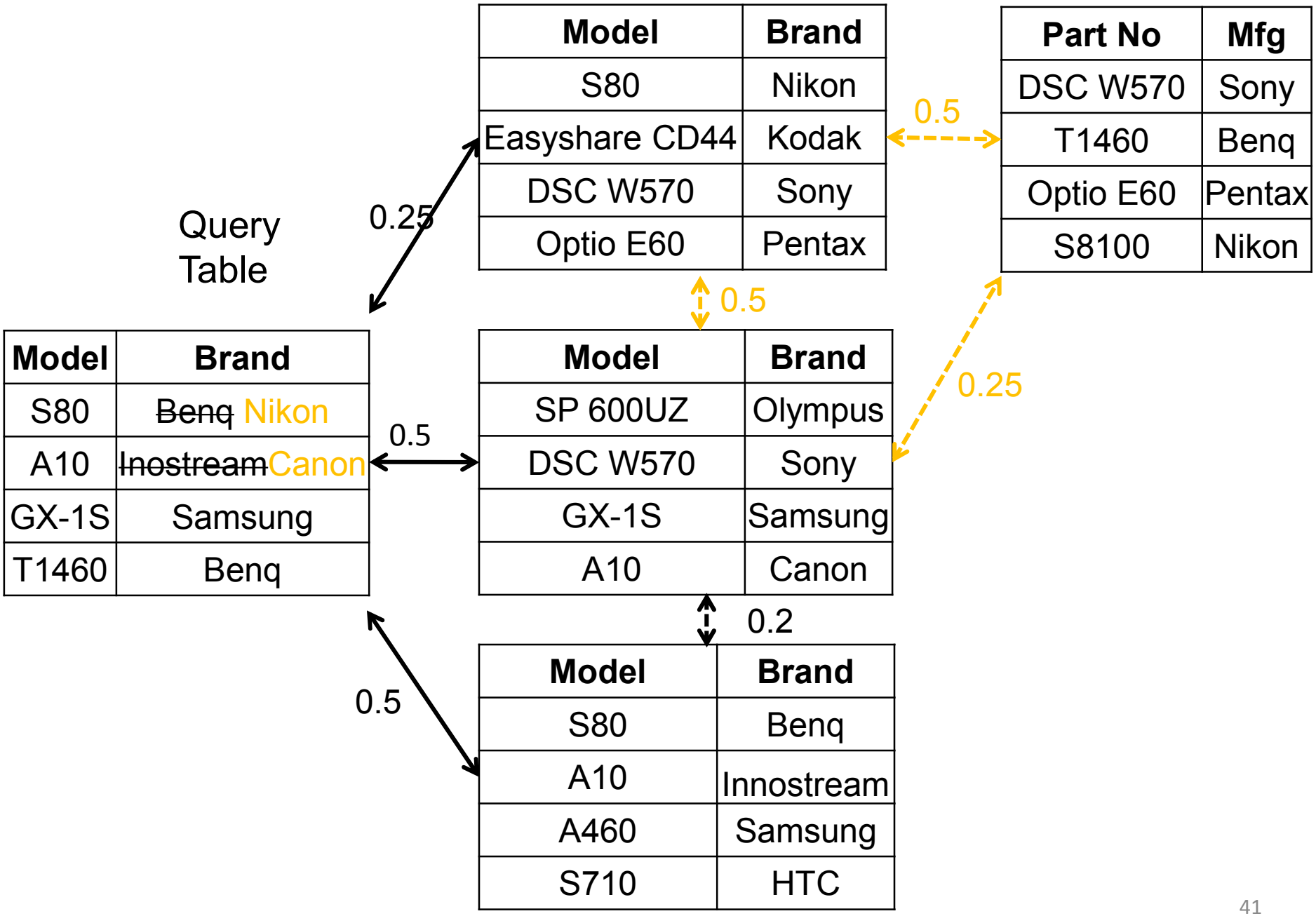
Model	Brand
S80	Benq
A10	Innostream
A460	Samsung
S710	HTC

Low precision
Low coverage


Query Table
Model
Brand
S80
Benq
A10
Innostream
GX-1S
Samsung
T1460
0.25
0.5
Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
0.5
Part No
Mfg
DSC W570
Sony
T1460
Benq
Optio E60
Pentax
S8100
Nikon
0.5
0.25
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.2
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC
Indirect matching table


Query Table
Model
Brand
S80
Benq
A10
Innostream
GX-1S
Samsung
T1460
Benq
0.25
Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
0.5
Part No
Mfg
DSC W570
Sony
T1460
Benq
Optio E60
Pentax
S8100
Nikon
0.5
0.5
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.25
0.2
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC


Query Table
T1
Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
Part No
Mfg
DSC W570
Sony
T1460
Benq
Optio E60
Pentax
S8100
Nikon
Model
Brand
S80
Benq
A10
Innostream
GX-1S
Samsung
T1460
Benq
T2
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
T3
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC
0.25
0.5
0.5
0.25
0.5
0.2
0.5


Query Table
Model
Brand
S80
Benq Nikon
A10
InostreamCanon
GX-1S
Samsung
T1460
Benq
0.25
0.5
Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
0.5
Part No
Mfg
DSC W570
Sony
T1460
Benq
Optio E60
Pentax
S8100
Nikon
0.5
0.25
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.2
0.5
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC

How to model holistic match?

- Topic Sensitive Pagerank (TSP)
  - Prepare the Schema matching graph among web tables (SMW)
  - Compute PPVs (Personalized PageRank Vectors) for every node
  - By linearity of PageRank, at query time, TSP can be expressed as a weighted linear combination of PPVs.

Data Model

- Entity-Attribute Binary (EAB) relations
- The Query table is an EAB relation
- Web tables are split into EAB relations
  - URL, title, context retained

Query Table

Model	Brand
S80	
A10	
GX-1S	
T1460	

Web Table

Model	Brand	Res	Optical Zoom
S80	Nikon	14.1mp	5x
Easyshare CD44	Kodak	12mp	3x
DSC W570	Sony	16.1	5x
Optio E60	Pentax	10.1	3x

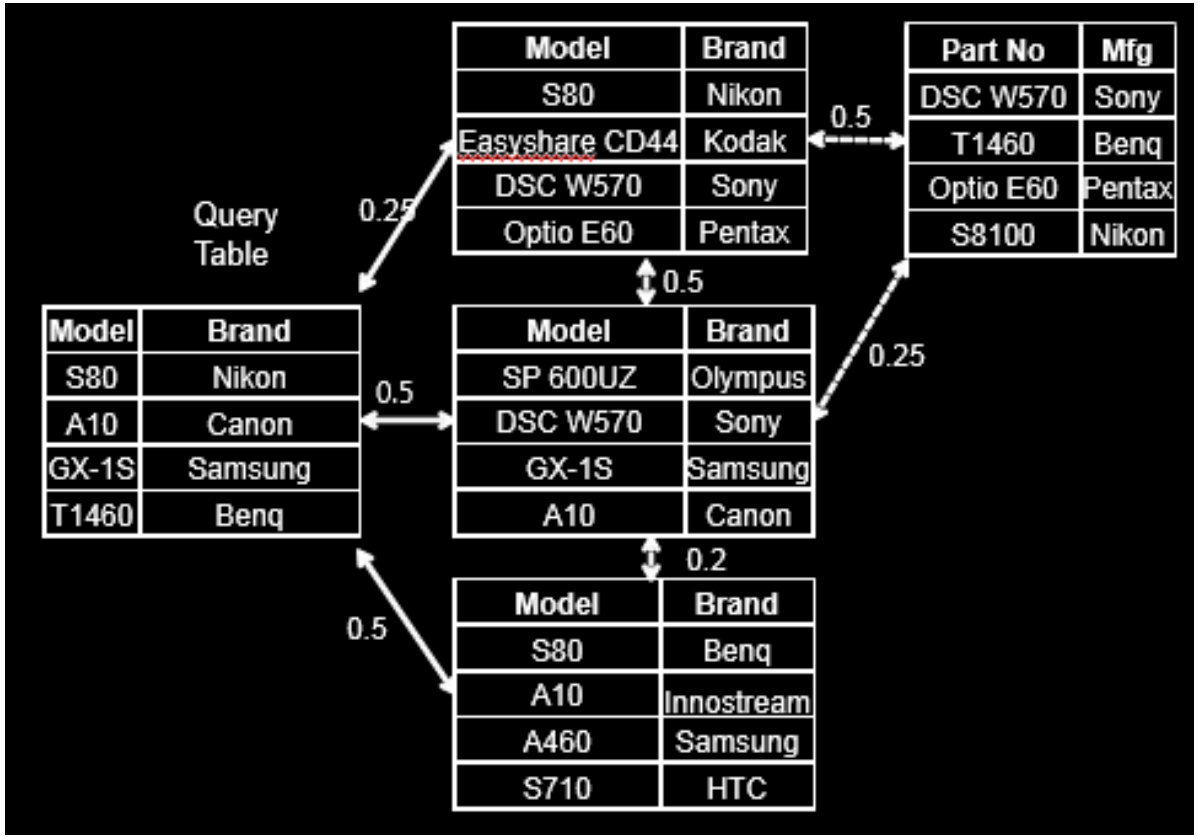
Model	Res
S80	14.1mp
Easyshare CD44	12mp
DSC W570	16.1
Optio E60	10.1

Model	Brand
S80	Nikon
Easyshare CD44	Kodak
DSC W570	Sony
Optio E60	Pentax

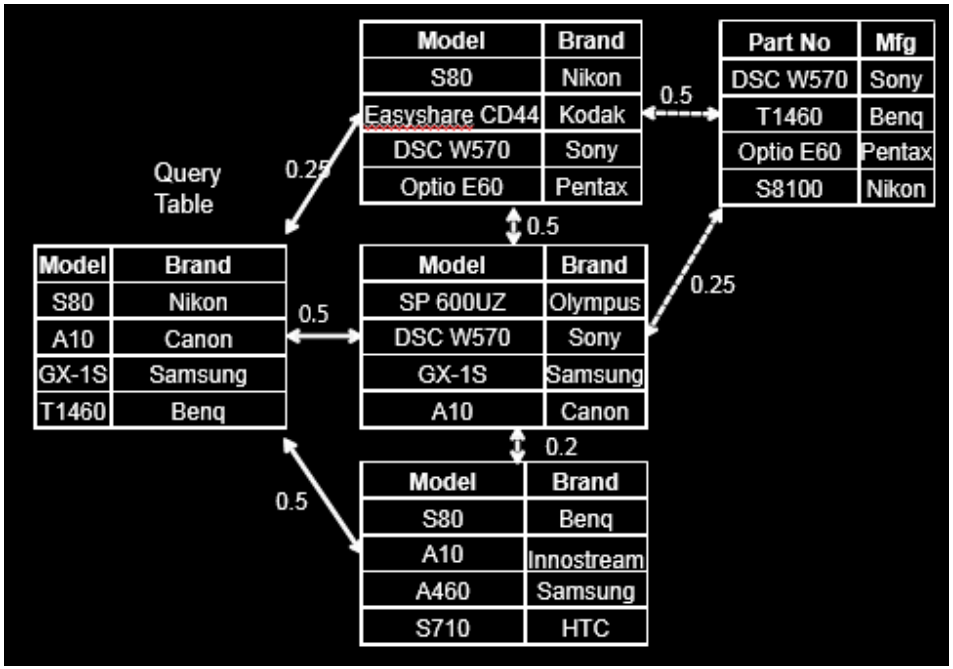
Model	Optical Zoom
S80	5x
Easyshare CD44	3x
DSC W570	5x
Optio E60	3x

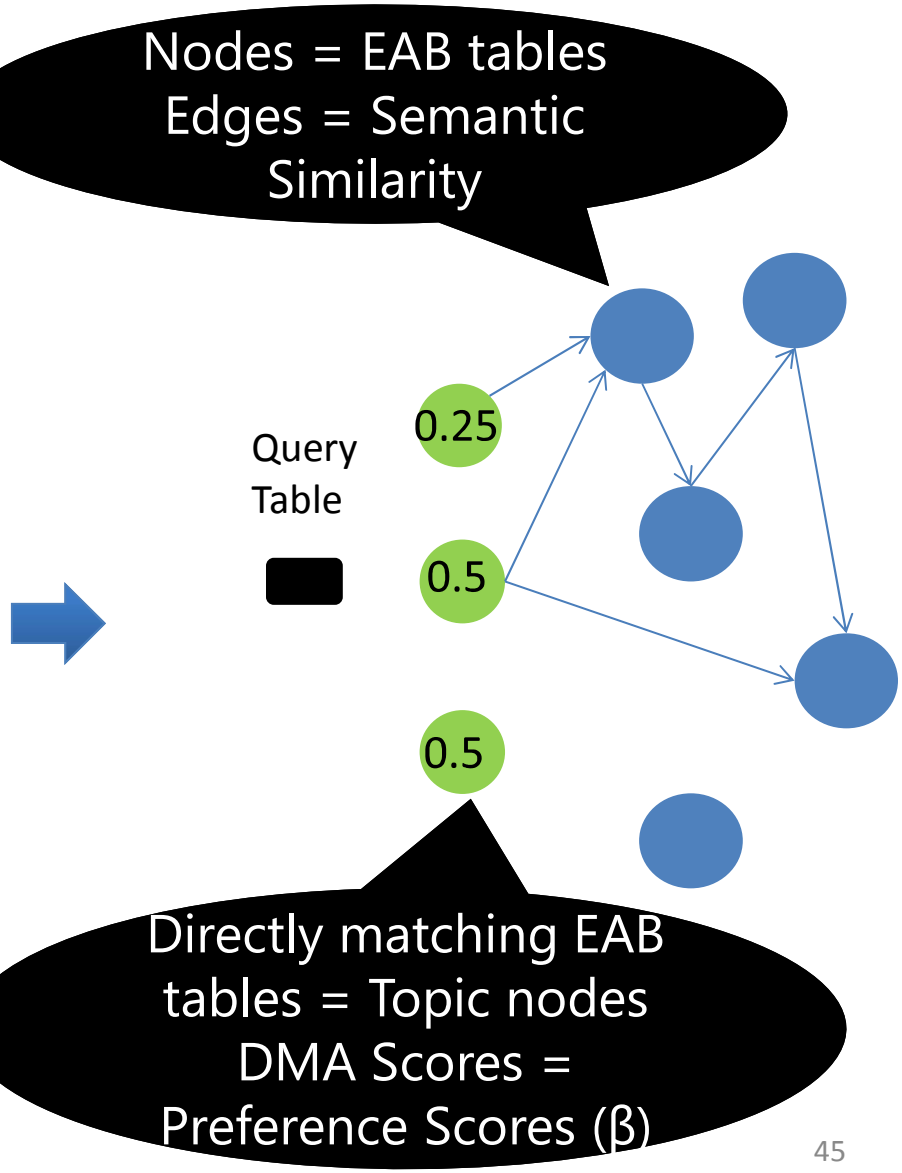
Augmentation Framework

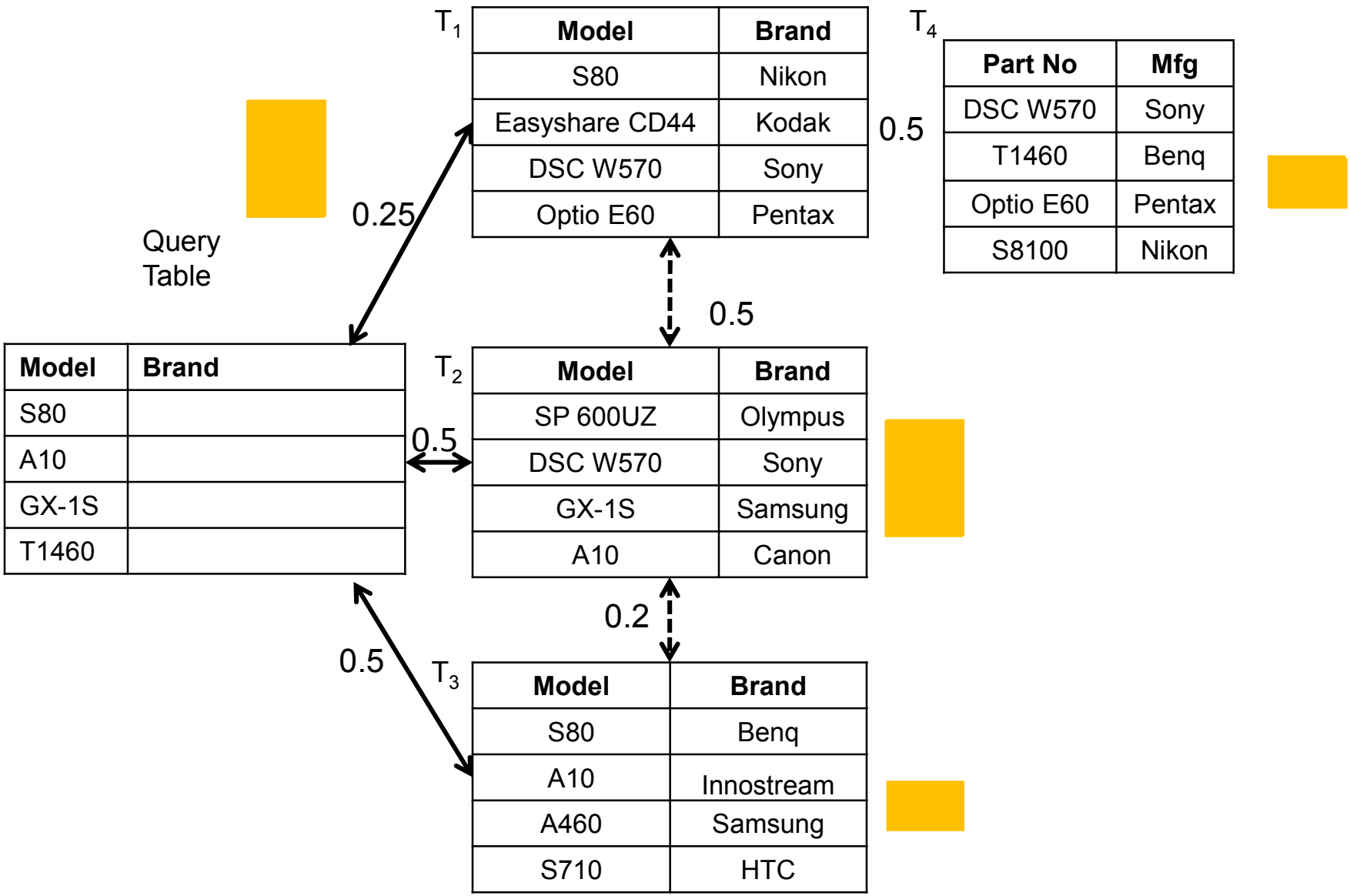
- 2 steps:
  - Identify Matching Tables
  - Predict Values Using Matching Tables



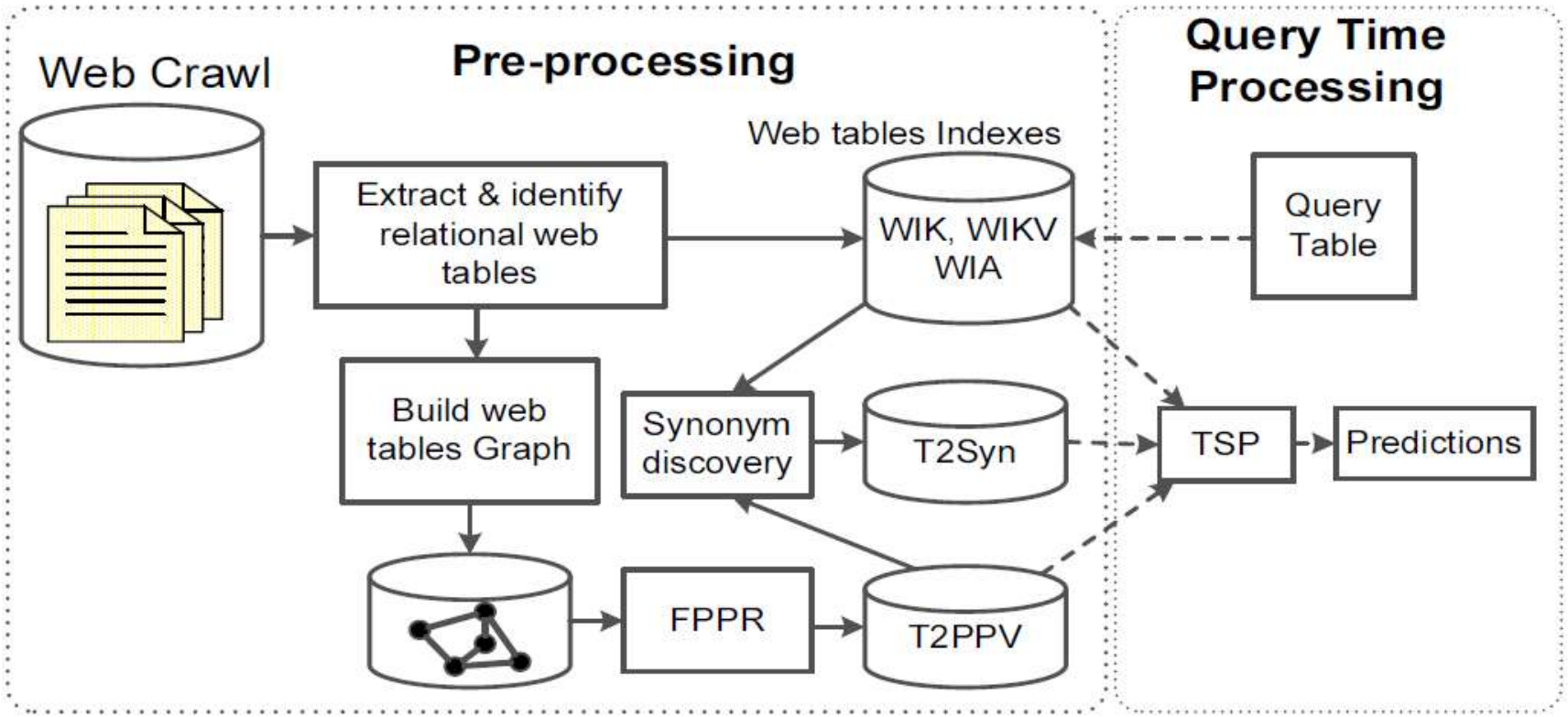
Problem to solve: Holistic matching of EAB Tables






T1
Model
Brand
S80
Nikon
Easyshare CD44
Kodak
DSC W570
Sony
Optio E60
Pentax
T4
Part No
Mfg
DSC W570
Sony
T1460
Benq
Optio E60
Pentax
S8100
Nikon
0.25
0.5
Query Table
Model
Brand
S80
A10
GX-1S
T1460
0.5
T2
Model
Brand
SP 600UZ
Olympus
DSC W570
Sony
GX-1S
Samsung
A10
Canon
0.2
0.5
T3
Model
Brand
S80
Benq
A10
Innostream
A460
Samsung
S710
HTC

InfoGather System Architecture



InfoGather System Architecture

- Indexes
  - WIK: index on the web tables' key attribute values
    - WIK(*Q*) returns the set of web tables that overlaps with *Q* on at least one of the keys.
  - WIKV: An index for the web tables complete records (that is key and value combined)
    - WIKV(*Q*) returns the set of web tables that contain at least one record from *Q*.
  - WIA: An index on the web tables attributes names
  - T2PPV: Index that stores PPV for every table T on SMW
  - T2Syn: Indexes synonyms of attribute B of table T

Building the SMW Graph

- Matching web tables
  - Should have the match for the key (type of entities) and the attribute column

Feature name	Document
Context	Terms in the text around the web table with idf weights
Table-to-Context	The table content as text and context text with idf weights
URL	The terms in the URL with idf weight computed from all URL set
Tuples	All the distinct table rows (or key-value pairs) form terms of a document with equal weights
Attributes name	The terms mentioned in the column names with equal weights
Column values	All the distinct values in a column form terms of a document with equal weights
Table-to-Table	The table content as text with idf weights

Summary

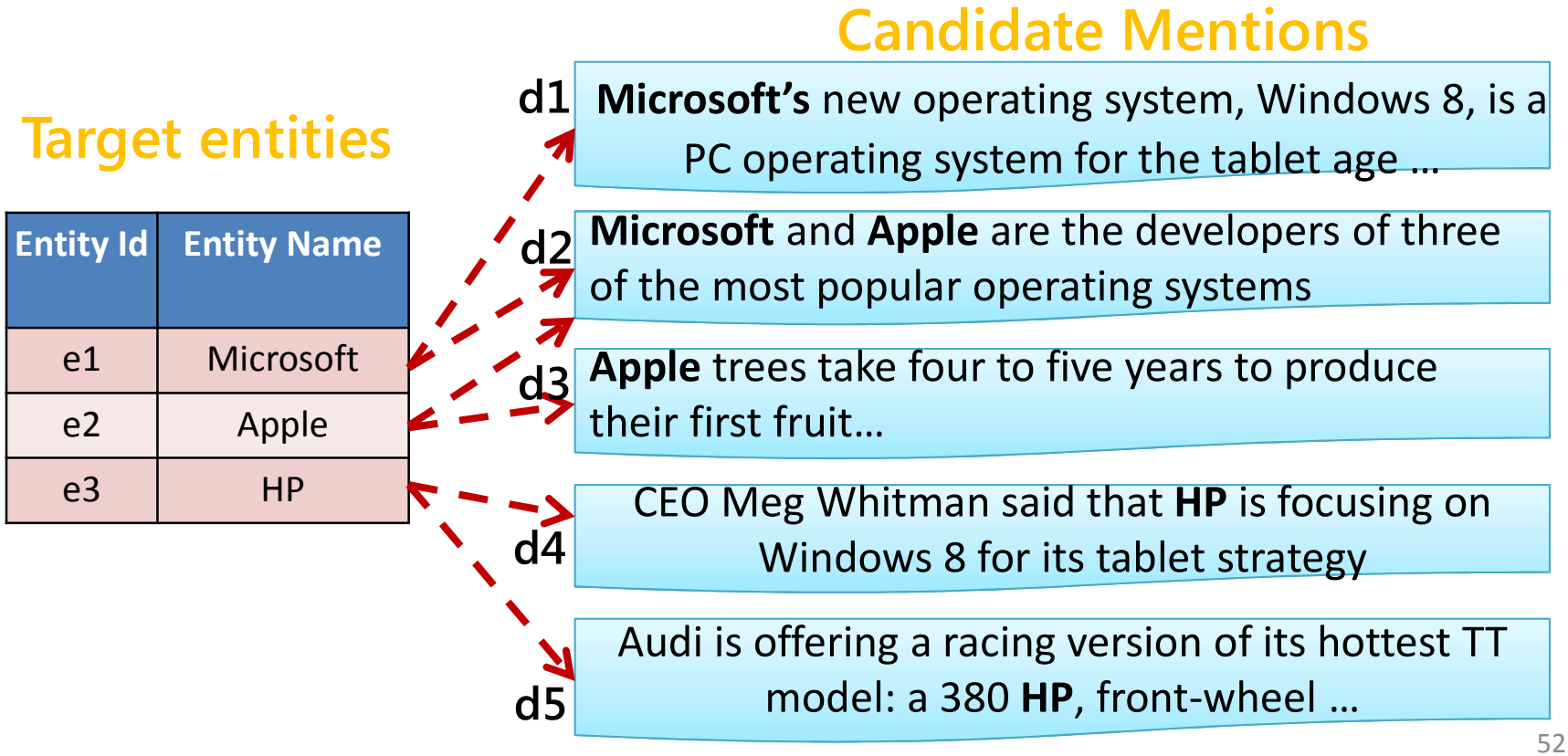
- Heavy use of preprocessing
  - Compute full personalized pagerank (FPPR) of semantic similarity graph upfront
  - Compute TSP scores on the fly
- Scalability is a huge challenge
  - How to build the SS graph?
    - > 1billion nodes
  - How to compute FPPR?
    - Use algorithm from Bahmani et. al. (SIGMOD 2011)
- Leveraging huge amount of data available anyway
  - As in MT, Speech Recognition
- Details in paper: "InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables", SIGMOD 2012

Today's Agenda

- Entity Synonyms
- Entity Attribute Discovery and Augmentation
- **Entity Linking**

Entity Linking Problem

- Given a "catalog" of entities and a set of documents, find mentions of those entities in those documents
  - Many applications
  - Main problem is identify true mentions among candidate mentions (refer it to as "targeted disambiguation")

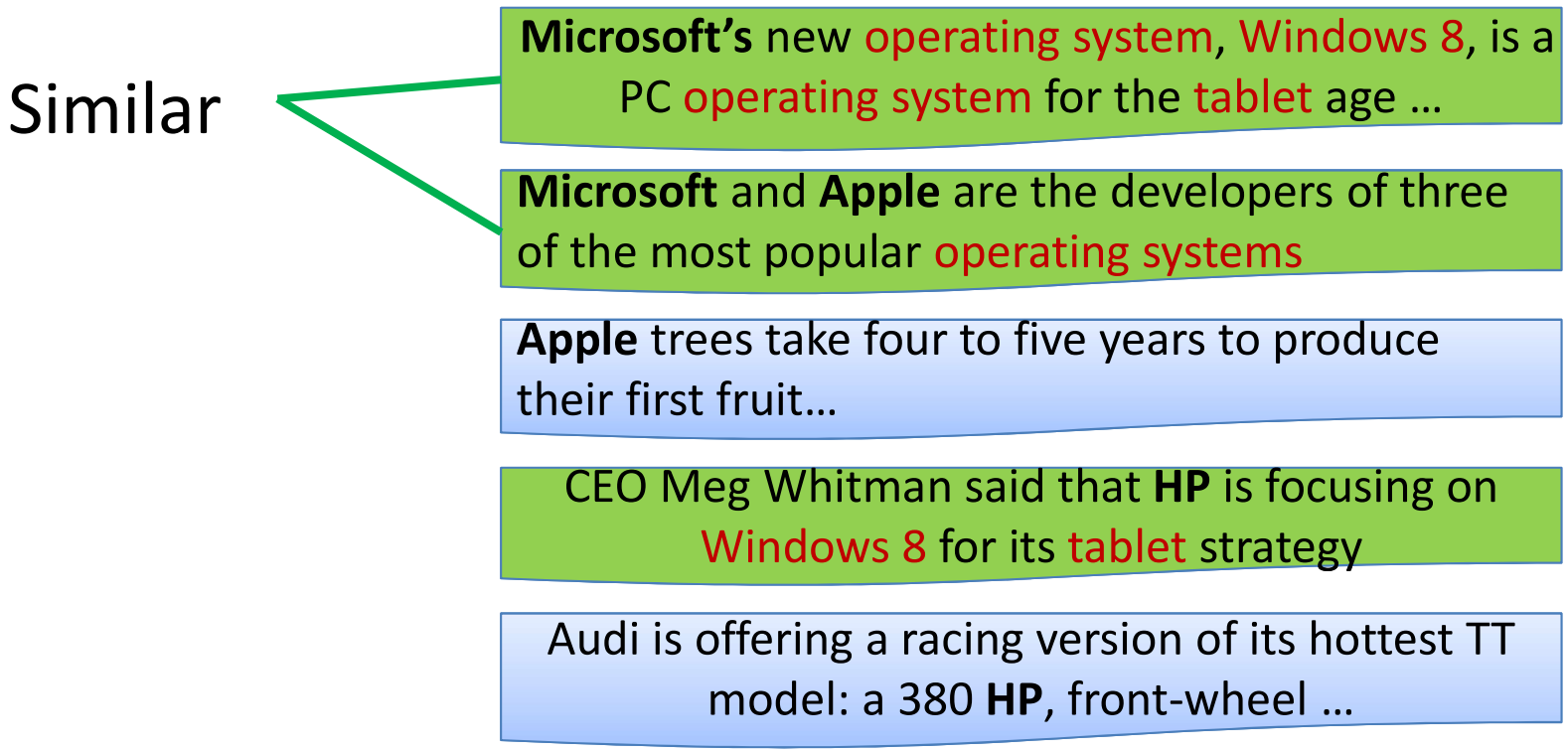


Existing solutions

- Solutions exist when entity catalog is “rich” with description about each entity
  - Wikipedia
- What if these entities are enterprise entities?
  - Not in Wikipedia, referred to as “ad hoc entities”
  - Documents can be enterprise documents or web documents
- Again, leverage the data!
  - Assume the entities in the catalog are homogeneous (of the same “type”)

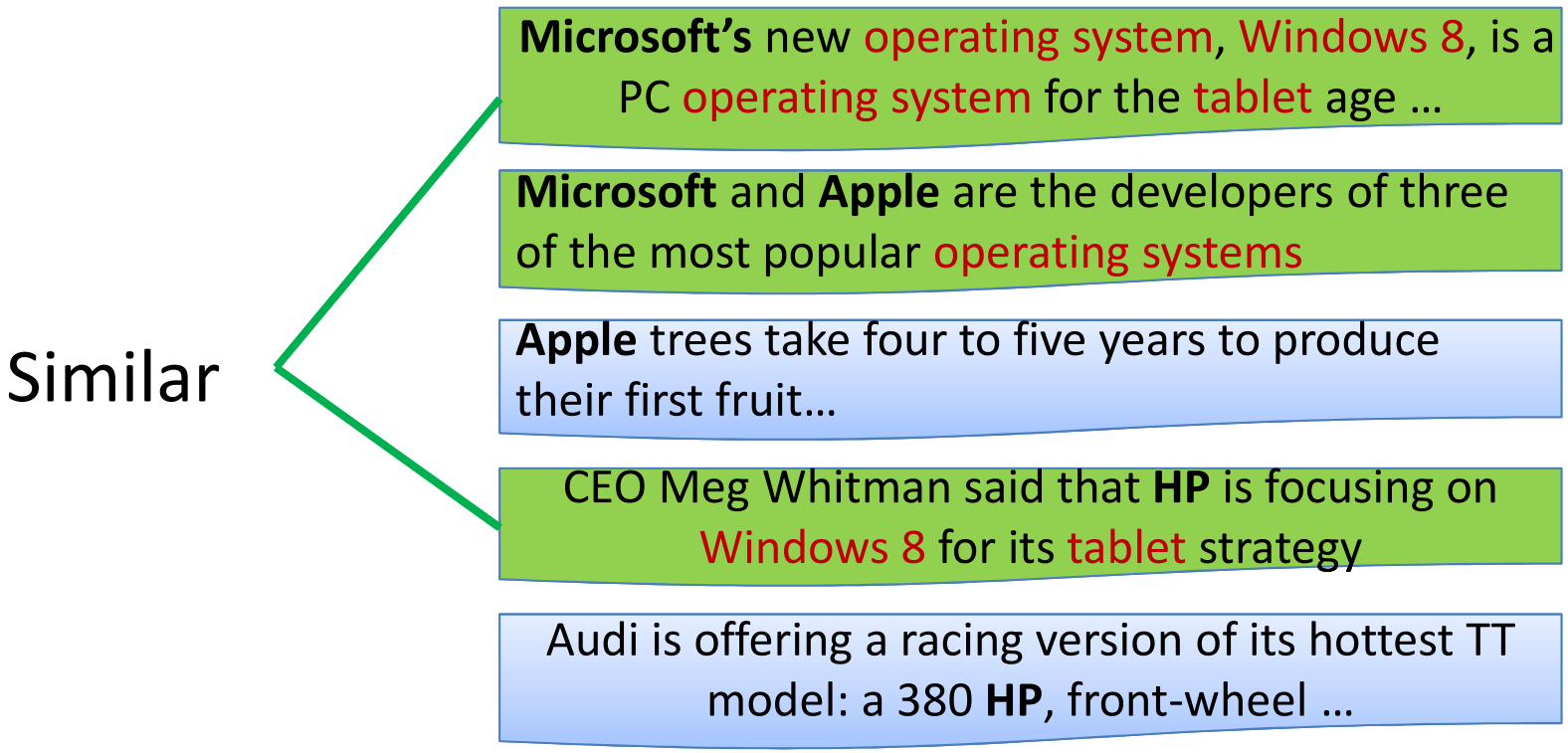
Insight – leverage homogeneity (1)

- Context Similarity



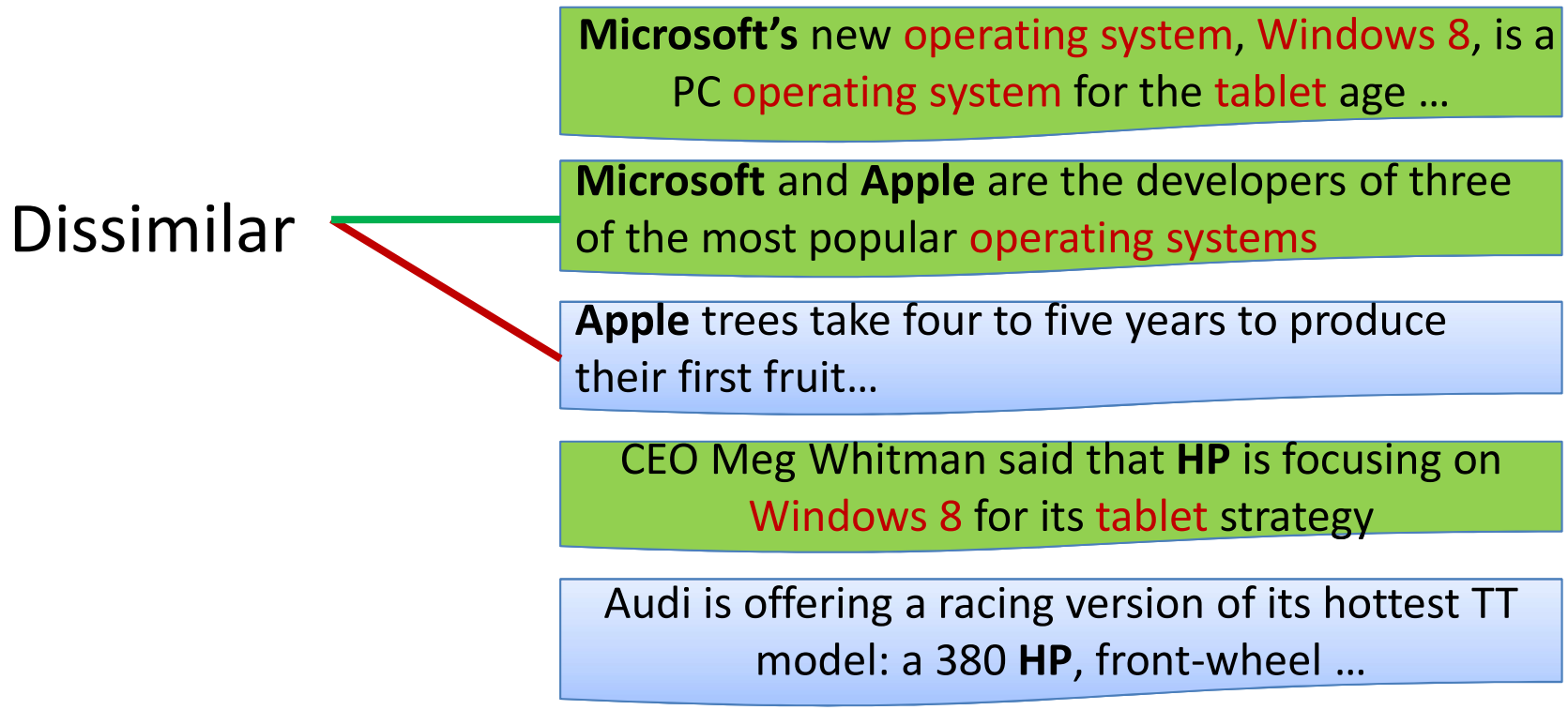
Insight – leverage homogeneity (1)

- Context Similarity



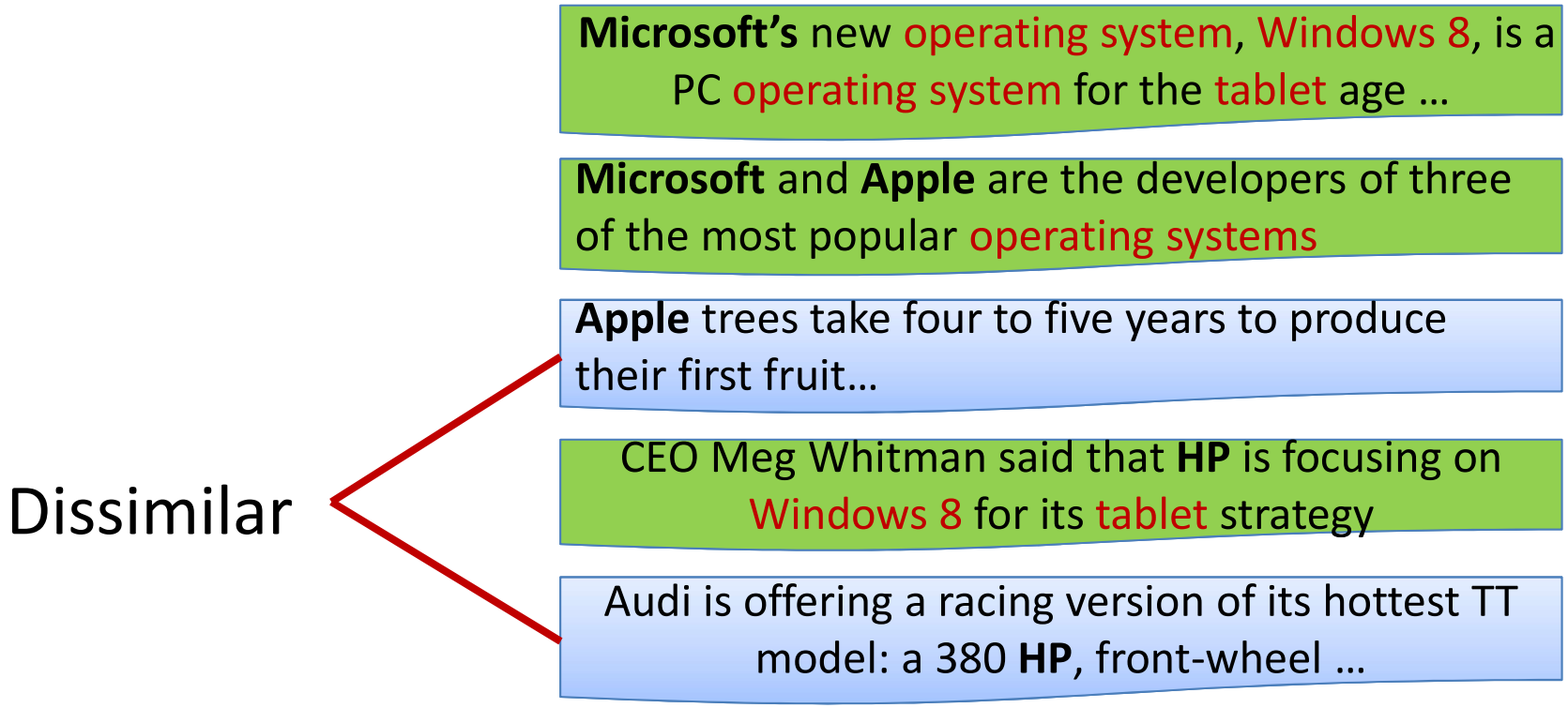
Insight – leverage homogeneity (1)

- Context Similarity

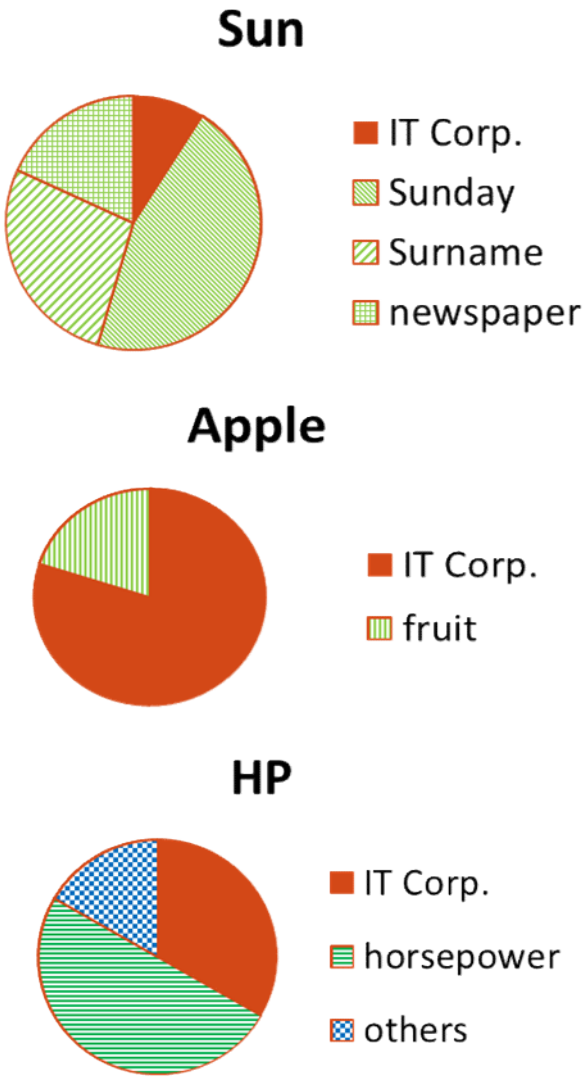


Insight – leverage homogeneity (1)

- Context Similarity



Insight – leverage homogeneity (1)



- Hypothesis:
- context between two true mentions (of any entities) is more similar than between two false mentions (of distinct entities) as well as between a true mention and a false mention

- Detail: the context of false mentions can be similar among themselves within an entity

Insight – leverage homogeneity (2)

- Co-mention

High confidence

Microsoft's new operating system, Windows 8, is a PC operating system for the tablet age …

Microsoft and **Apple** are the developers of three of the most popular operating systems

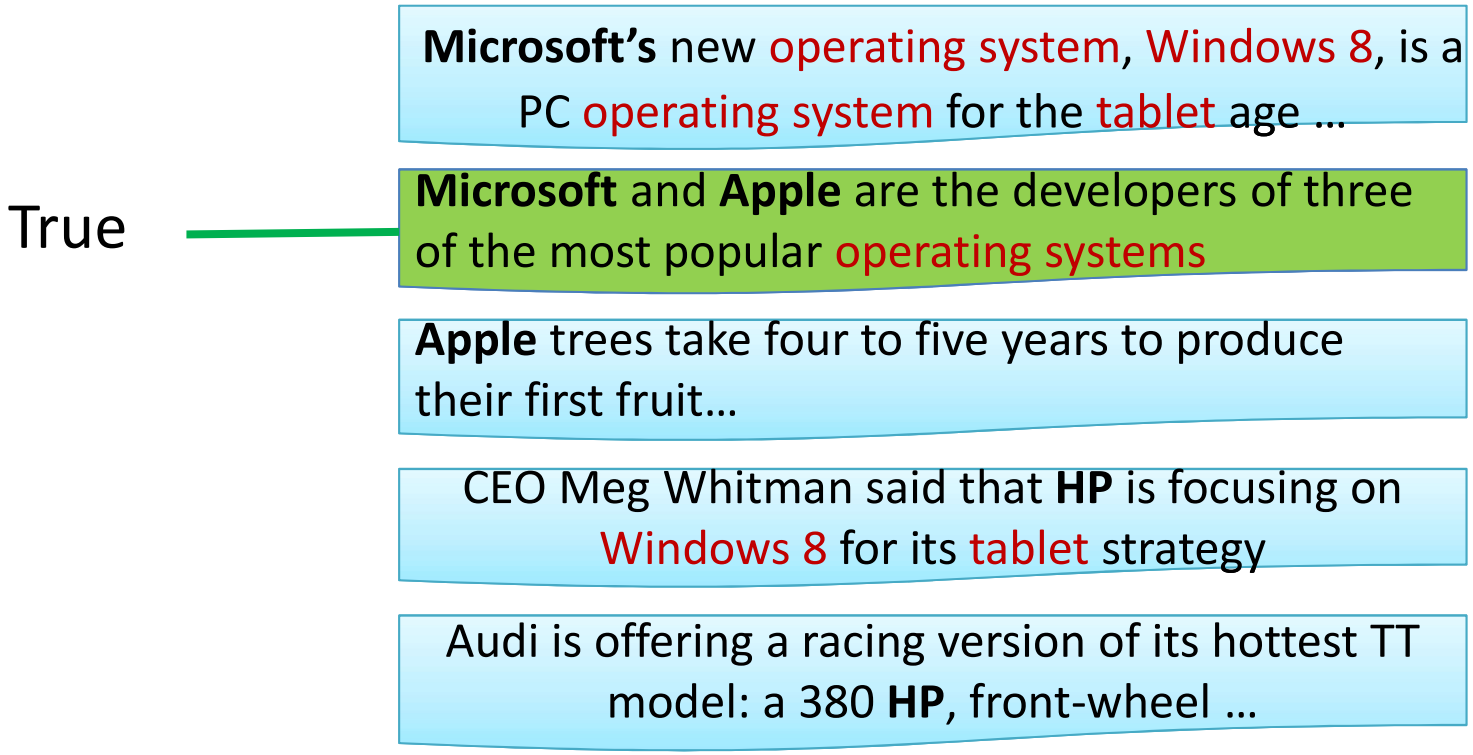
Apple trees take four to five years to produce their first fruit…

CEO Meg Whitman said that **HP** is focusing on Windows 8 for its tablet strategy

Audi is offering a racing version of its hottest TT model: a 380 **HP**, front-wheel …

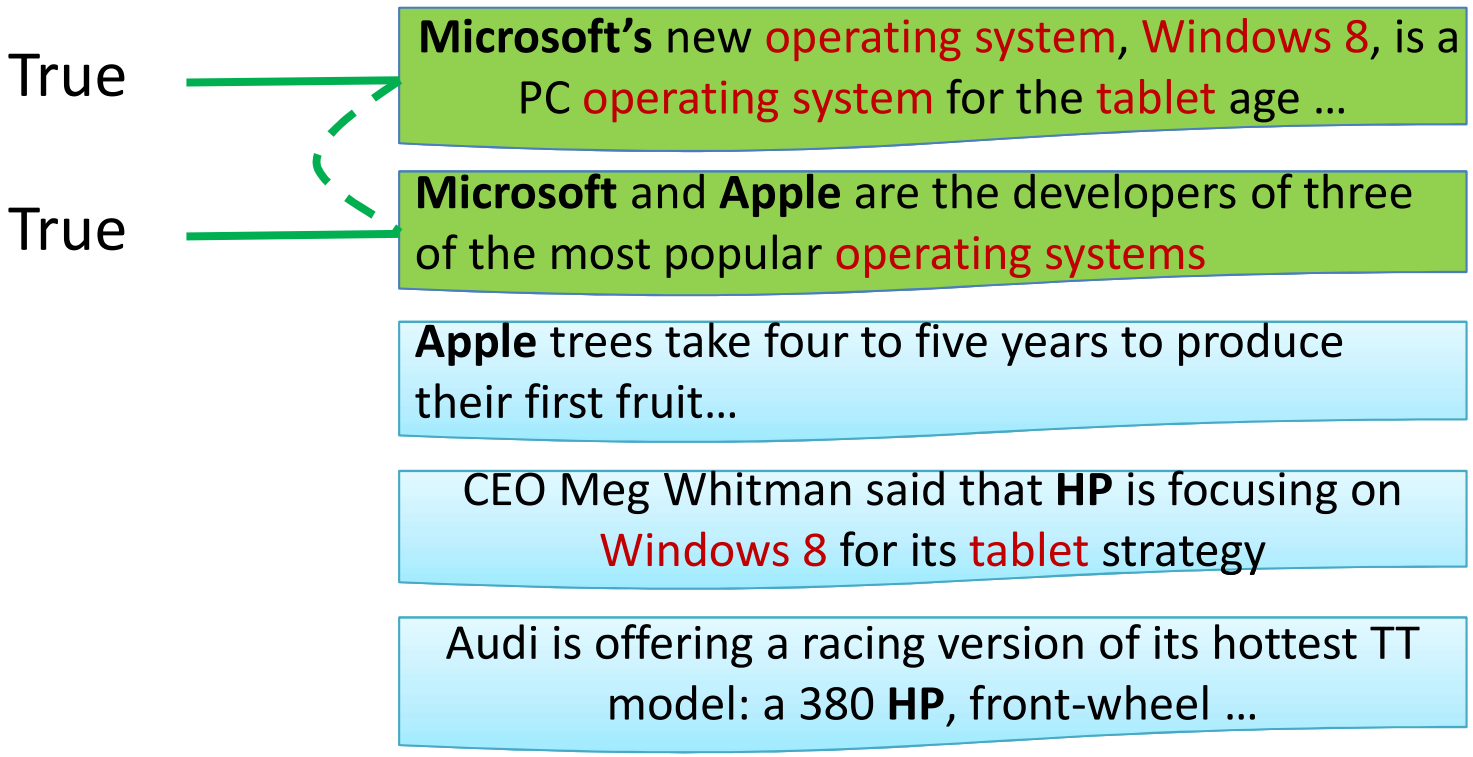
Insight – leverage homogeneity (3)

- Interdependency



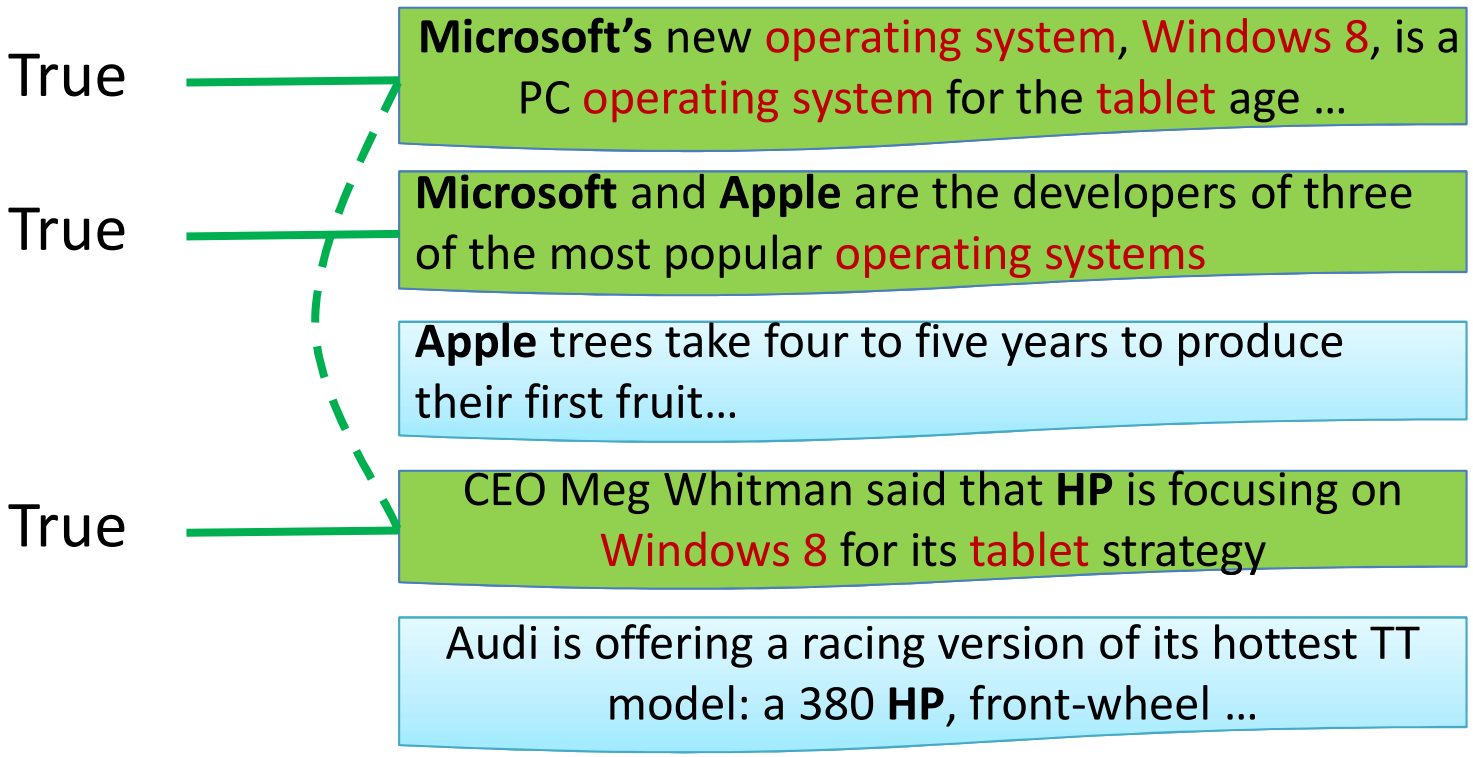
Insight – leverage homogeneity (3)

- Interdependency



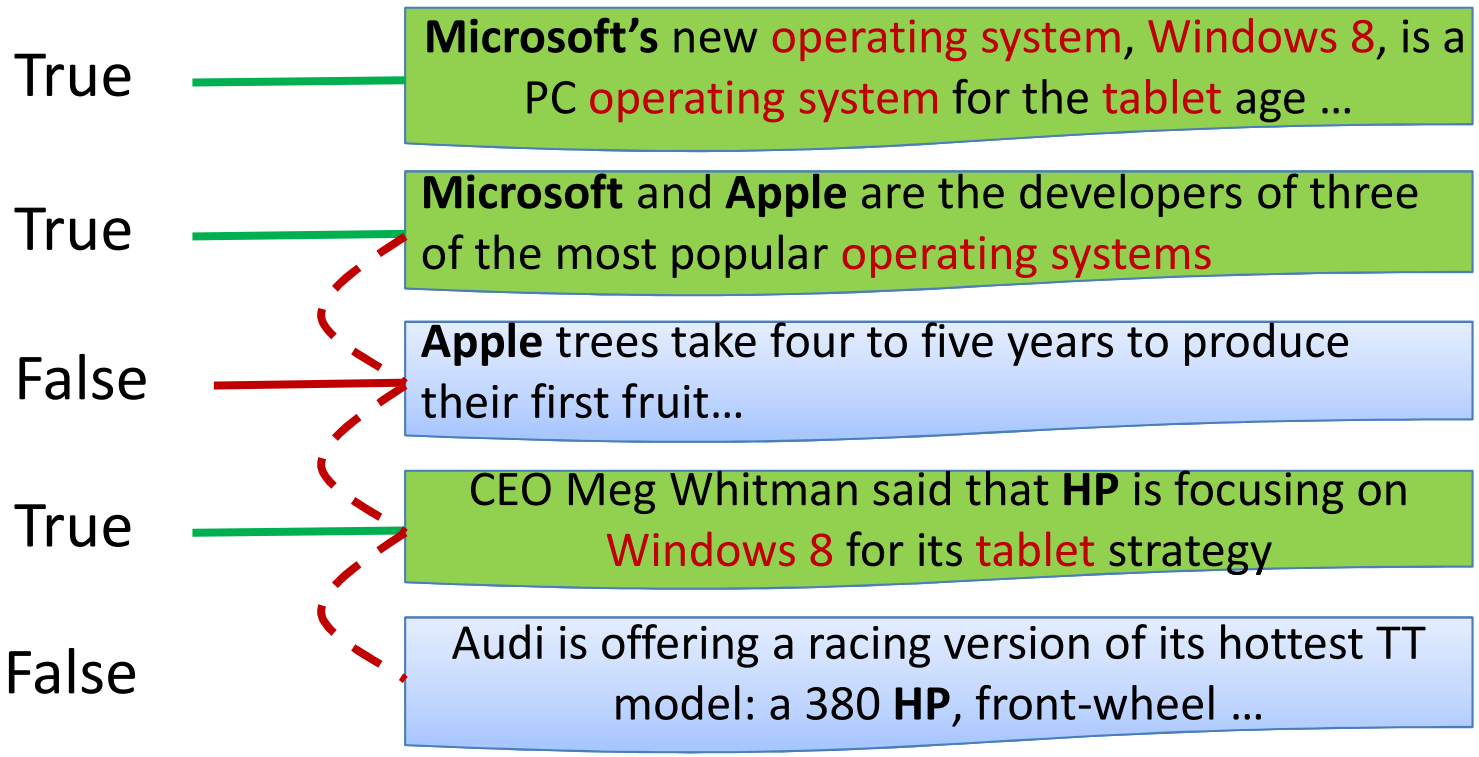
Insight – leverage homogeneity (3)

- Interdependency



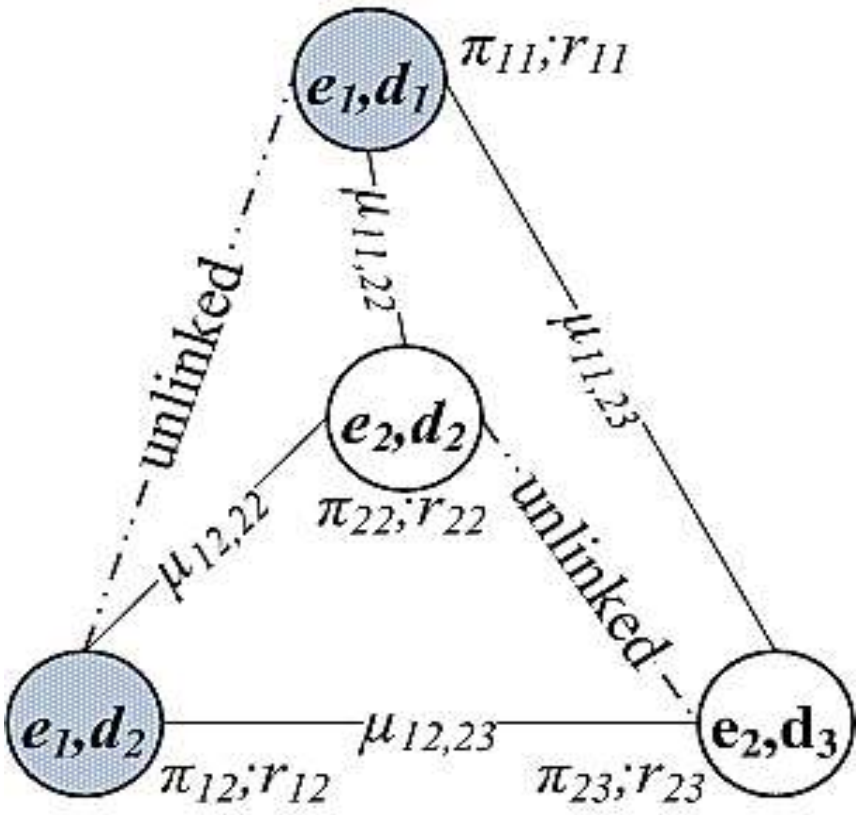
Insight – leverage homogeneity (3)

- Interdependency



Modeling using Graph: MentionRank (1)

- "Targeted Disambiguation of Ad-hoc, Homogeneous Sets of Named Entities", WWW 2012
- Each node is a mention (e_i, d_j) and has a prior score $\pi_{i,j}$
- Edge weight μ is based on context similarity
  - take a short context window of each occurrence of the entity name, and compute tf-idf similarity for every pair of such contexts, and then take the average
- r_{ij} is the final score of interest



Modeling using Graph: MentionRank (2)

- Prior score of each candidate mention π_{ij} is defined based on co-mention
  - Number of unique names of target entities occurred in d_i
  - $p_{ij} = \frac{\pi_{ij}}{\sum_{ij} \pi_{ij}}$
  - Finally, $r_{ij} = \lambda p_{ij} + (1 - \lambda) \sum_{i'j'} w_{iji'j'} r_{i'j'}$
- Interdependency modeling using PageRank-style propagation
  - Can be proved that power iterations work
  - But propagation is very different from PageRank

Modeling using Graph: MentionRank (2)

- Propagation weight
  - μ is un-normalized, so cannot use directly
- Although false mentions for an individual entity can be similar to each other, the false mentions across distinct entities belong to more heterogeneous domain than true mentions. Therefore, it is more reliable for a mention to be deemed true if it has similar context with mentions of many *different* entities than with many mentions of *the same* entity name.
  - unlinking – disallow the propagation between candidate mentions of the same entity
  - normalization – restrict the total contribution from mentions of an individual entity
- Since we perform context similarity over a short text window, similarity scores could be 0 or close to 0 for many pairs
  - Smoothing

$$
\begin{aligned}
w_{i'j',ij} &= \frac{z_{ij}}{k}, \text{ if } i = i' \\
&= \frac{\mu_{i'j',ij}}{V_i Z} + \frac{z_{ij}}{k}, \text{ otherwise} \\
z_{ij} &= 1 - \frac{\sum_{i' \neq i} \sum_{j'} \mu_{i'j',ij}}{V_i Z} \\
Z &= \max_{i,j} \frac{\sum_{i' \neq i} \sum_{j'} \mu_{i'j',ij}}{V_i}
\end{aligned}
$$

- V_i is the number of documents that have candidate mentions of e_i in the collection
- k is the total number of candidate mentions

Take-away Messages

- Mining semantics of entities is critical
  - Both on the web and within the enterprise
  - Lots of new entity tasks being studied
  - We discussed the following
    - Entity synonyms
    - Entity attribute discovery
    - Entity augmentation
    - Entity linking

Further Reading

- Kaushik Chakrabarti's tutorial "Simple Models, Lots of Data: Mining semantics about entities using Web-Scale Data" at Joint International Workshop on Entity-oriented and Semantic Search (JIWES) 2012 @ SIGIR Conference
- Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng, Dong Xin. A Framework for Robust Discovery of Entity Synonyms, KDD 2012.
- Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, Surajit Chaudhuri. InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables. SIGMOD 2012.
- Chi Wang, Kaushik Chakrabarti, Tao Cheng, Surajit Chaudhuri. Targeted Disambiguation of Ad-hoc, Homogeneous Sets of Named Entities. WWW 2012.

Preview of Lecture 20: Entity Semantics Mining (Part 2)

- Entity Set Expansion
- Entity Acronym Expansion
- Entity Actions
- Entity Tagging

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!