Scientific Text Mining and Knowledge Graphs

Chapter 2
Part 1: Taxonomy Construction

Presenter: Jingbo Shang

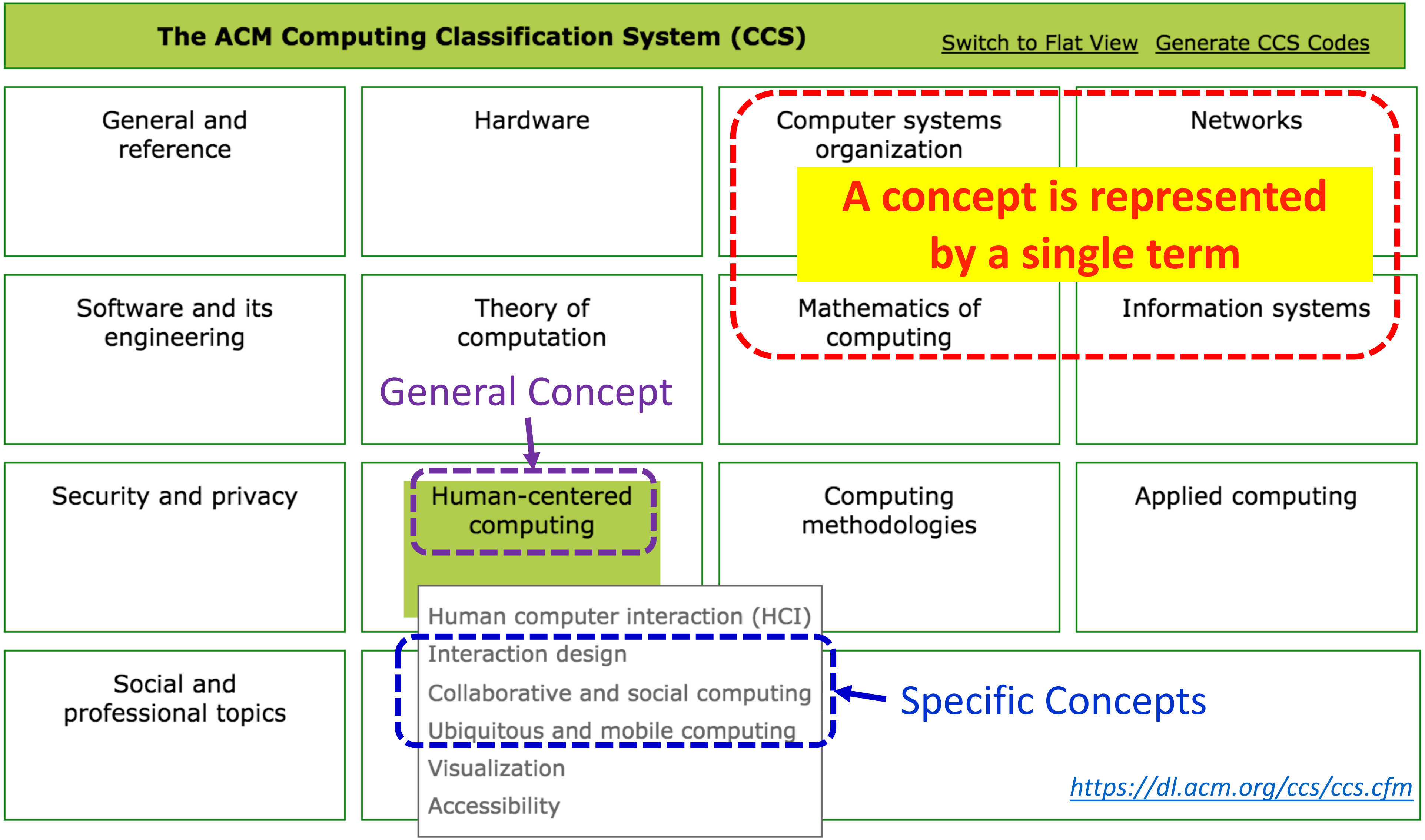
University of California, San Diego

jshang@ucsd.edu

What is a taxonomy?

Taxonomy is the practice and science of describing and organizing *concepts*

Example: ACM CCS Taxonomy



Example: eBay Product Taxonomy

All Categories

- eBay Motors
- Fashion
- **Books, Movies & Music**
- Electronics
- Collectibles & Art
- Home & Garden
- Sporting Goods
- Toys & Hobbies
- Business & Industrial
- Health
- Others

General concepts

A concept is represented by a small term cluster

Electronics

Specific concepts

Cell Phones, Smart Watches & Accessories

- Cell Phones & Smartphones
- Cell Phone Accessories
- Smart Watches
- Cell Phone & Smartphone Parts
- Cell Phone Cards & SIM Cards
- Smart Watch Accessories

View all in Cell Phones, Smart Watches & Accessories →

Computers, Tablets & Network Hardware

- Computer Components & Parts
- Laptops & Netbooks
- Tablets & eReaders
- Desktops & All-In-One Computers
- Computer Drives, Storage & Blank Media
- Computer Printers, Scanners & Supplies

View all in Computers, Tablets & Network Hardware →

Cameras & Photo

- Camera Lenses & Filters
- Digital Cameras
- Film Photography Equipment
- Camera Drones
- Camera & Photo Accessories
- Camcorders

View all in Cameras & Photo →

Most specific concepts

https://www.ebay.com/n/all-categories

Two General Types of Taxonomy

- Each concept in the taxonomy is called a ***taxon*** and based on how we represent a taxon, we categorize taxonomies into two types:
  - **Instance-based** Taxonomy – each taxon is a single term (+ strict synonyms)

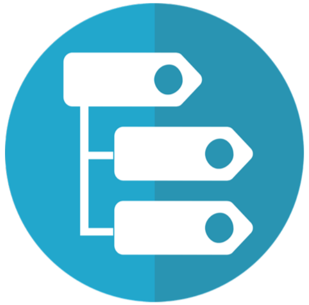
ACM CCS WORDNET MeSH ProBase

  - **Clustering-based** Taxonomy – each taxon is a topically related term cluster

ebay amazon Google Express wayfair

Taxonomy Underpins Digital Library

Scientific Papers



Taxonomy

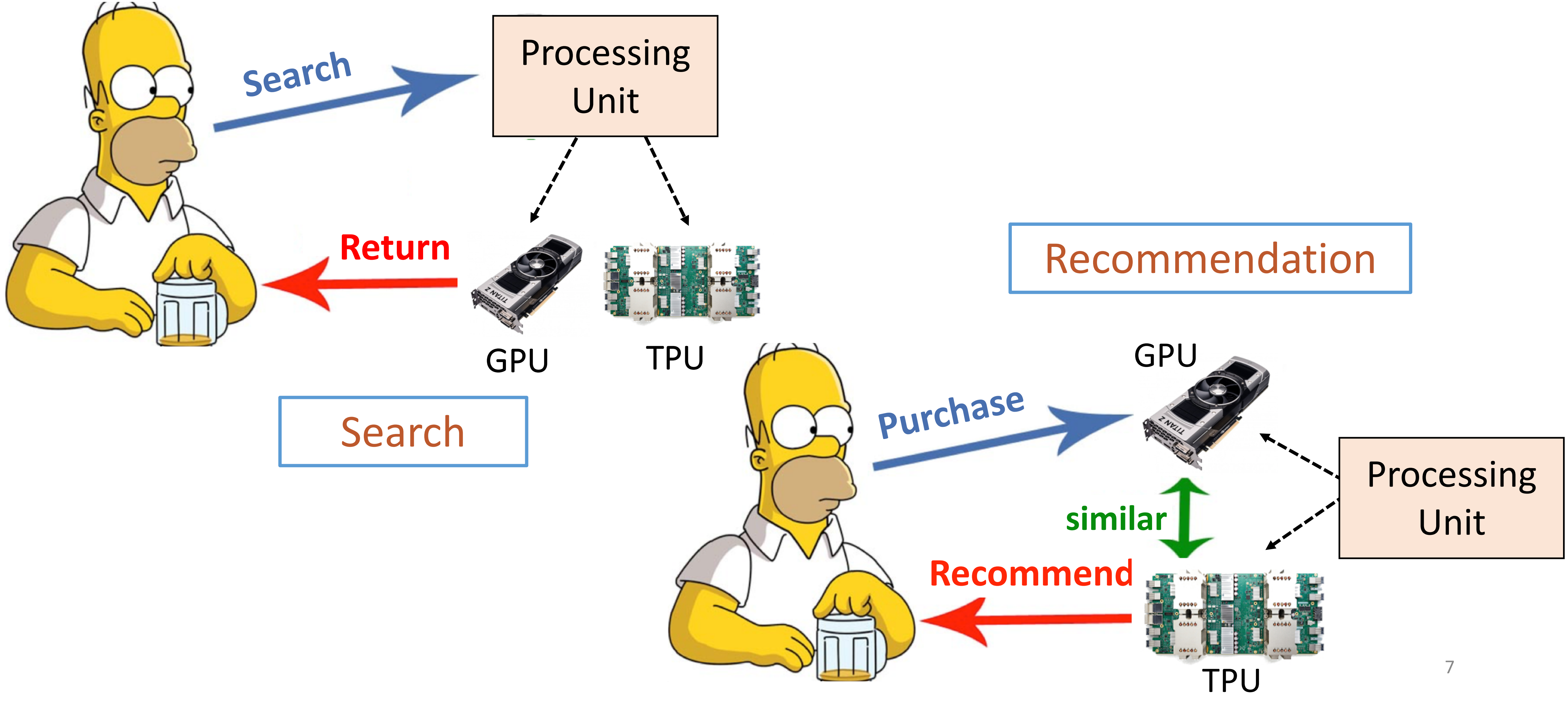
Document Categorization

Data Mining Papers

Theory Papers

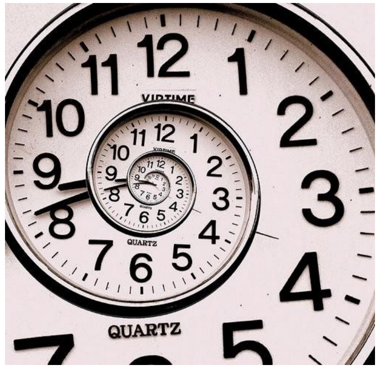
HCI Papers

Taxonomy Helps Search & Recommendation

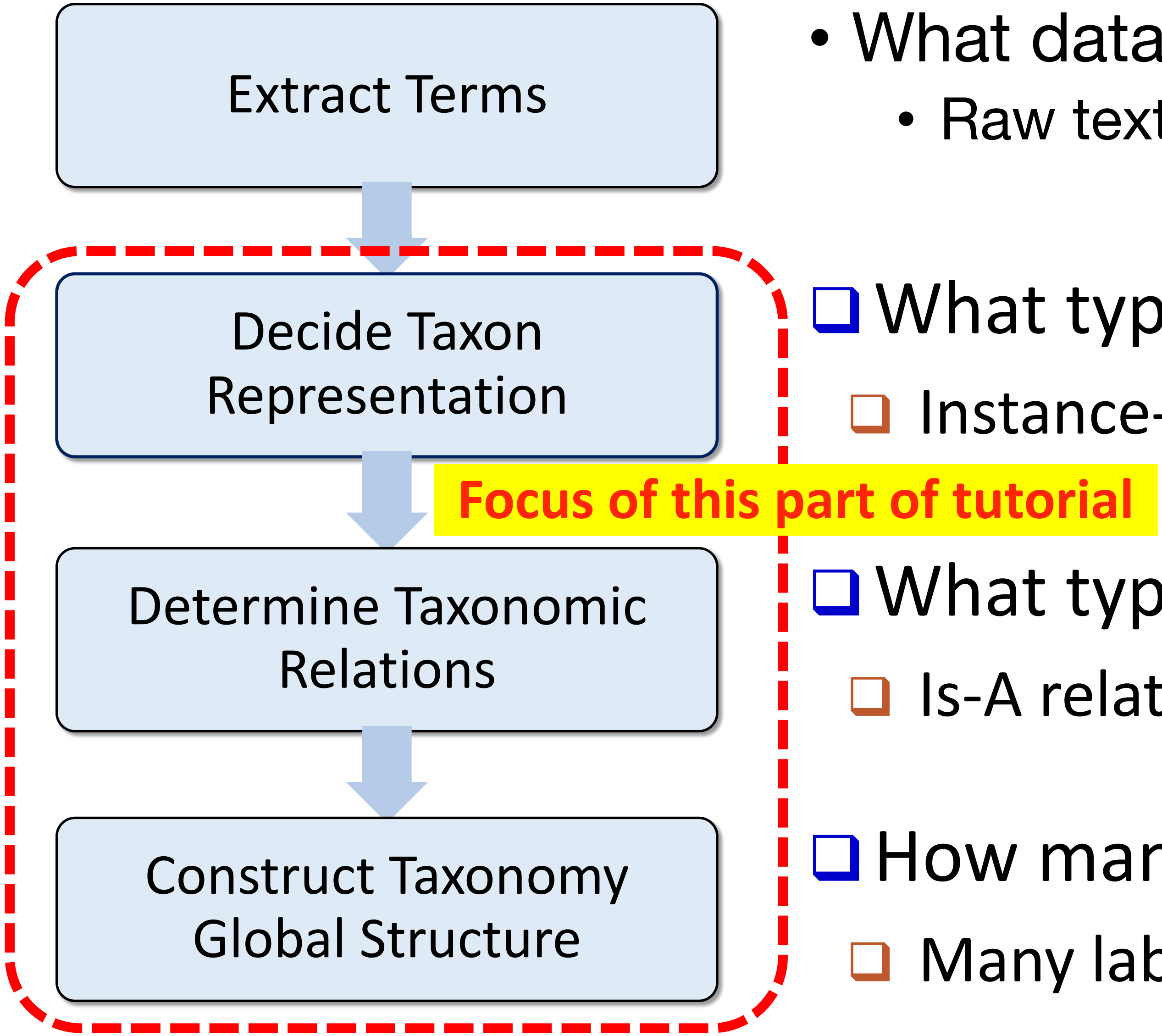


How to Build a Taxonomy?

- Manual curation
  - Time-consuming and expensive
  - Human (expert) labor-intensive
- Automated construction
  - Scalable and extensible



Typical Taxonomy Construction Process



- What data sources you have?
  - Raw text? Query logs? Clicks? Networks?

- What type of taxonomy you want?
  - Instance-based? Clustering-based?

- What type of relation you want?
  - Is-A relation? Part-of relation? Other relation?

- How many supervision you have?
  - Many labeled taxonomies? Several seed taxons?

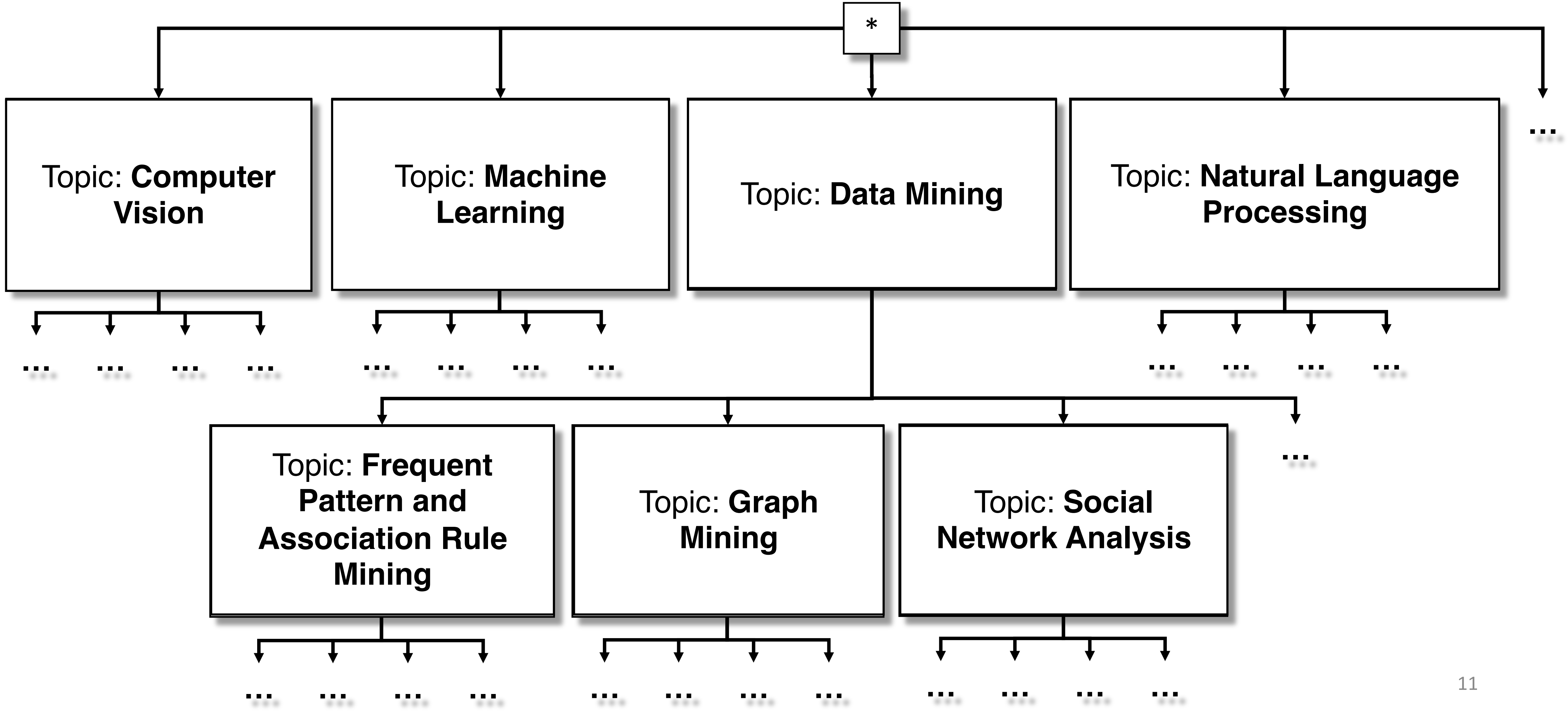
Taxonomy Construction Methods: A Landscape

For a more comprehensive landscape see:
https://github.com/mickeystroller/awesome-taxonomy

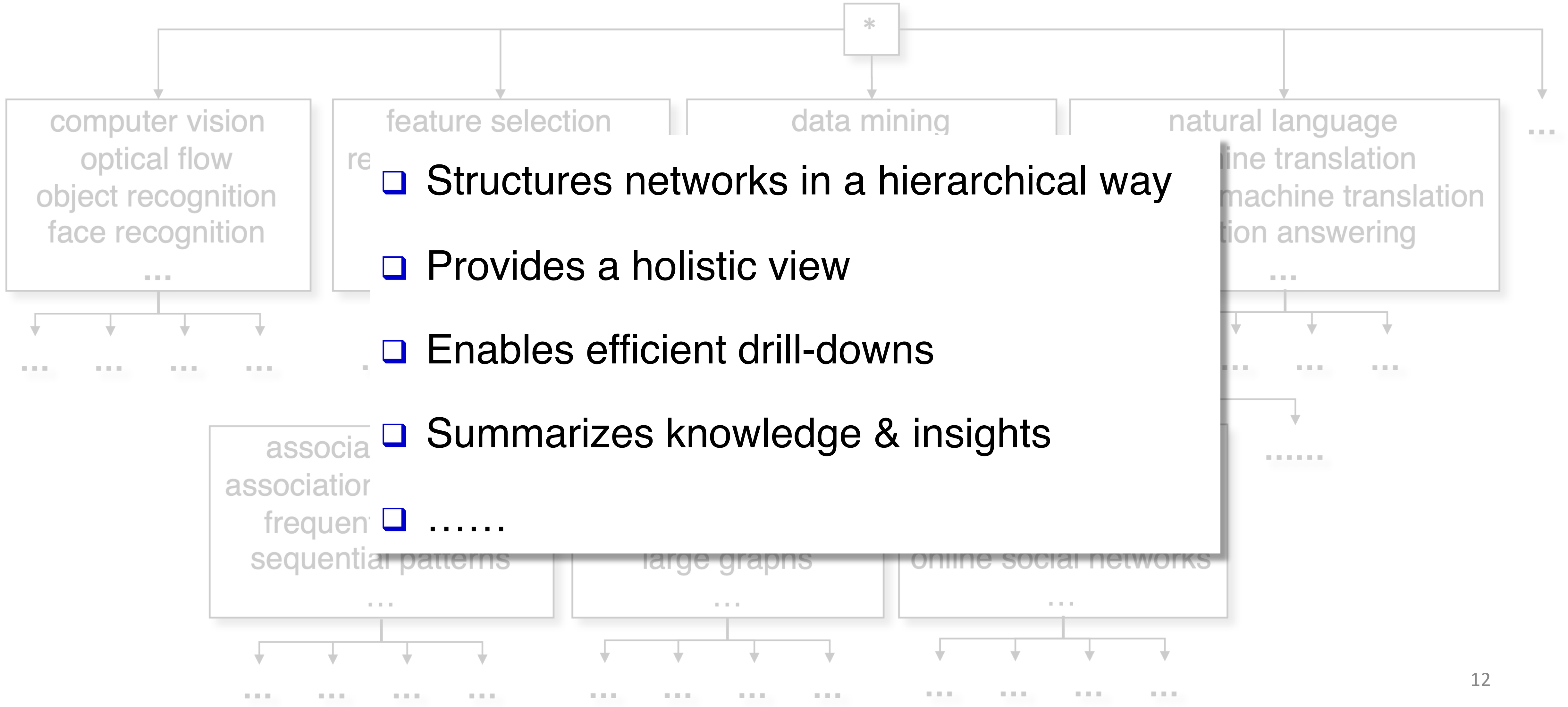
Focus of this tutorial

Structure level \ Label Amount	*No supervision*	*A few labeled supervision*	*Fully labeled relations/taxonomies*
Clustering-based taxonomy	hLDA [Blei et al.'03] CATHY [Wang et al.'13a] TaxoGen [Chao et al.'18] NetTaxo [Shang et al.'20]	SSHLDA [Mao et al.'12] HDM [Wang et al.'14]	HSLDA [Perotte et al.'11]
Instance-based taxonomy	Probase [Wu et al.'12] OntoLearn [Velardi et al.'13] WiBi [Flati et al.'14]	HiExpan [Shen et al.'18]	STI [Snow et al.'06] SL-MST [Bansal et al.'14] TaxoRL [Mao et al.'18]
Taxonomic relation	Hearst Pattern [Hearst'92] DIVE [Chang et al.'18] U-TEAL [Wang et al.'19]	SOL [Nakashole et al.'12] MetaPAD [Jiang et al.'17]	Piecewise-LP [Fu et al.'14] HypeNET [Shwartz et al.'16]

Constructed Topic Taxonomy: Example



Why Topic Taxonomy?



- Structures networks in a hierarchical way
- Provides a holistic view
- Enables efficient drill-downs
- Summarizes knowledge & insights
- ……

Hierarchical Topic Model

- Use a cluster of terms (i.e., a topic) to represent a concept and organize topics in a hierarchical way
- Pose different statistical assumptions on the data generation process
  - Nested Chinese Restaurant Process:
    - hLDA [Blei et al.’03], hLDA-nCRP [Blei et al.’ 10]
  - Pachinko Allocation Model:
    - PAM [Li and McCallum’06], hPAM [Mimno et al.’07]
  - Dirichlet Forest Model：
    - DF [Andrzejewski et al.’09], Guided HTM [Shin and Moon’17]

Example: hLDA

- Document generation from Chinese Restaurant Process

1. Let c_1 be the root restaurant.
2. For each level $\ell \in \{2, \ldots, L\}$:
   (a) Draw a table from restaurant $c_{\ell-1}$ using Eq. (1). Set c_ℓ to be the restaurant referred to by that table.
3. Draw an L-dimensional topic proportion vector θ from $\mathrm{Dir}(\alpha)$.
4. For each word $n \in \{1, \ldots, N\}$:
   (a) Draw $z \in \{1, \ldots, L\}$ from $\mathrm{Mult}(\theta)$.
   (b) Draw w_n from the topic associated with restaurant c_z.

Generates

We develop an approach to risk minimization and stochastic optimization that provides a convex surrogate for variance, allowing near-optimal and computationally efficient trading between approximation and estimation error.

“Observed” documents

Inference

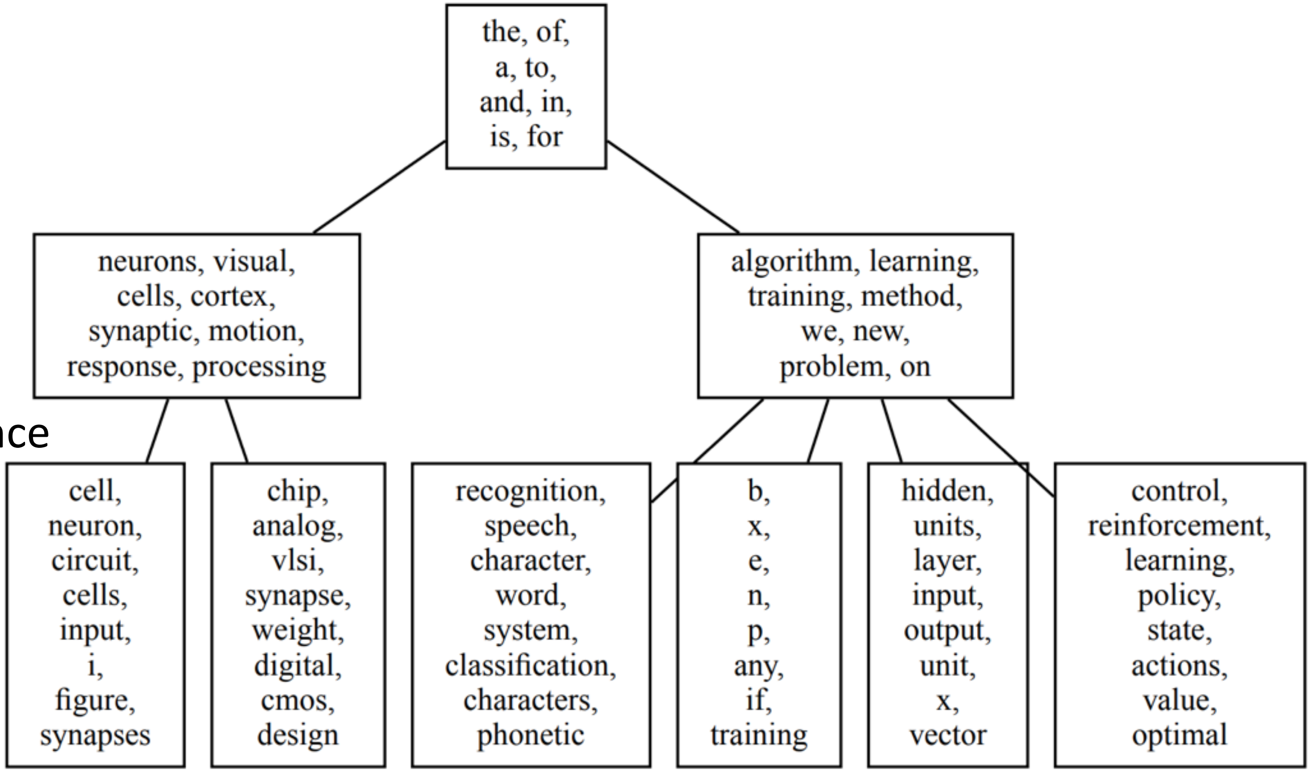


Figure credits to [Blei et al.’03]

Example: hPAM

- Document generation from Pachinko Allocation Model

1. For each document d, sample a distribution θ_0 over super-topics and a distribution θ_T over sub-topics for each super-topic.
2. For each word w,
   (a) Sample a super-topic z_T from θ_0.
   (b) Sample a sub-topic z_t from θ_{z_T}.
   (c) Sample a level ℓ from $\zeta_{z_T z_t}$.
   (d) Sample a word from ϕ_0 if $\ell = 1$, ϕ_{z_T} if $\ell = 2$, or ϕ_{z_t} if $\ell = 3$.

Generates

We develop an approach to risk minimization and stochastic optimization that provides a convex surrogate for variance, allowing near-optimal and computationally efficient trading between approximation and estimation error.

"Observed" documents

Inference

super-topic

sub-topic

- writes article don time apr
  - god jesus christ people christian
    - faith wrong read spiritual passage
    - agree reason matter statement means
    - history support community house involved
  - key government encryption president clipper
    - agree reason matter statement means
    - power arms president home vote
    - history support community house involved
  - israel jews israeli jewish arab
    - history support community house involved
    - side left happened committee region
    - agree reason matter statement means
  - turkish armenian armenians people turkey
    - side left happened committee region
    - history support community house involved
    - hundred clothes tyre bosnians origin
  - file ftp windows window image
    - bit fax manager lib uk
    - site dec sources key public
    - release size function appreciated box

Figure credits to [Mimno et al.'07]

Hierarchical Clustering

- Group terms into hierarchical clusters and each cluster represents an interested concept
- Top-down approaches:
  - CATHY [Wang et al.'13a]
  - CATHYHIN [Wang et al.'13b]
- Bottom-up approaches:
  - BRT [Liu et al.'12] [Song et al.' 15]

Example: CATHY [Wang et al.'13a]

- Step 1: Construct term co-occurrence network
- Step 2: Cluster co-occurrence network into subtopic's sub-networks and estimate each sub-topical phrase's frequency
- Step 3: Extract candidate phrases using topical frequency
- Step 4: Rank topical phrases based on topical frequency
- Step 5: Apply steps 2-5 to each subtopic recursively and construct the hierarchy in a top-down fashion

Example: BRT [Liu et al.'12]

- Agglomerative multi-branch clustering using Bayesian Rose Tree

Algorithm 1 Bayesian Rose Tree (BRT).

Input: A set of documents $\mathcal{D}$.

$\boxed{T_i \leftarrow \mathbf{x}_i \text{ for } i = 1, 2, \cdots, n}$

$c \leftarrow n$

while $c > 1$ **do**

1. Select T_i and T_j and merge them into T_m which maximizes

$$L(T_m) = \frac{p(\mathcal{D}_m|T_m)}{p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j)},$$

where the merge operation is join, absorb, or collapse.

2. Replace T_i and T_j with T_m in the tree.

3. $c \leftarrow c - 1$

end while

Join: $T_m = \{T_i, T_j\}$

Absorb: $T_m = \{\text{children}(T_i) \cup T_j\}$

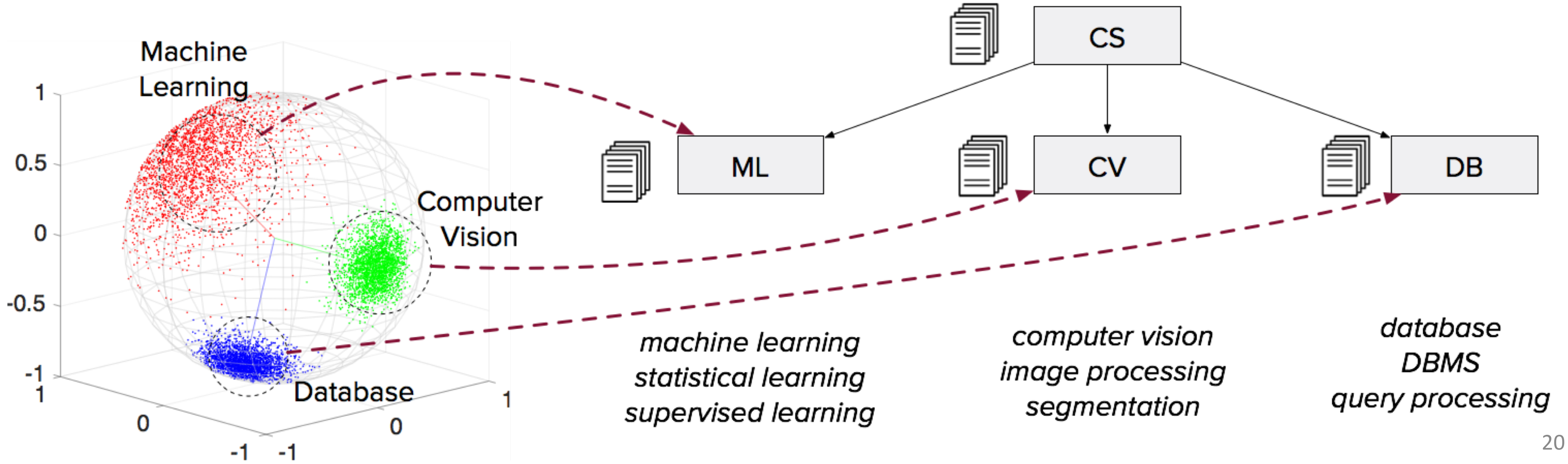
Collapse: $T_m = \{\text{children}(T_i) \cup \text{children}(T_j)\}$

Limitations of Previous Methods

- Too strong assumptions on document generation process
  - Bag-of-word document representation ignores word order information
  - Real-world data may not follow these statistical distributions/processes
- Computationally slow
  - Slow inference restricts their applications to large-scale data

Recent Methods: Uses Term Embedding

- Most existing work follows this idea: using term embedding to construct topic taxonomy based on hierarchical clustering
  - Learns term embedding to capture their semantic correlations
  - Constructs topic taxonomy in a recursive, top-down fashion



Limitation of Term Embedding: Example

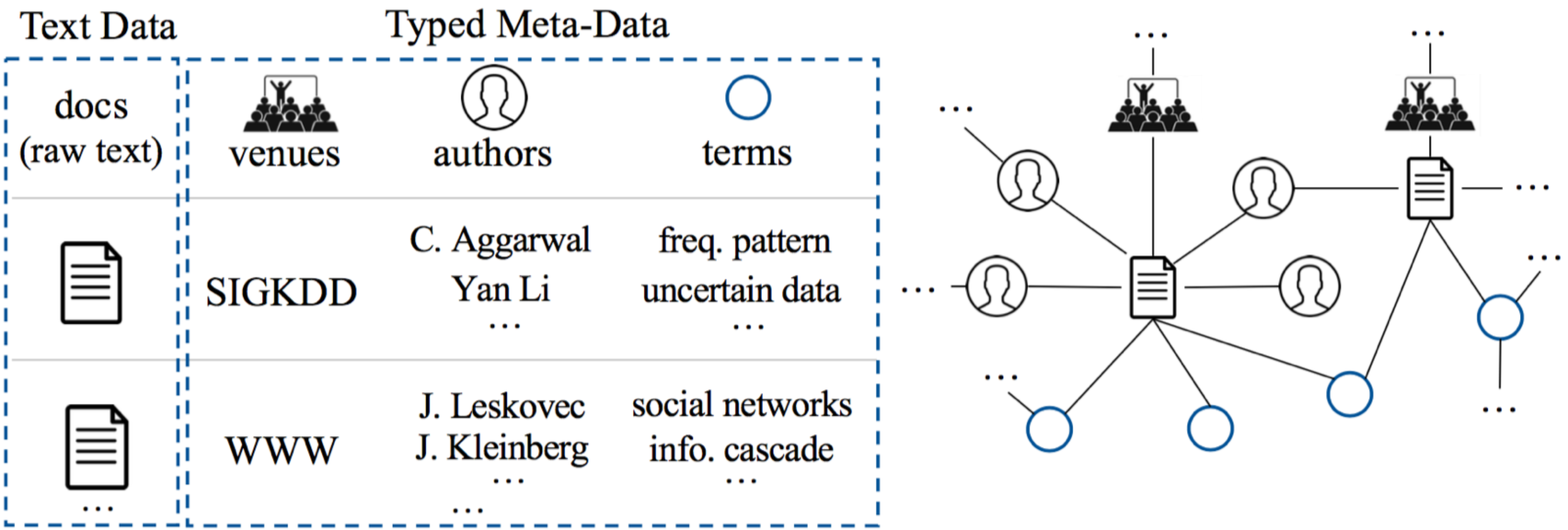
- Two terms in the Computer Science publications

SIGKDD & Data Mining Researchers —— "*Frequent Pattern*" vs. "*Transaction Database*" —— **SIGMOD & Database Researchers**

- From the taxonomy view
  - We should separate them into "*Data Mining*" and "*Database*"
- From the term embedding view
  - They are very similar due to similar contexts

Text-Rich Network: Text & Meta-Data

- Terms are extracted by AutoPhrase from raw texts (e.g., paper & review)

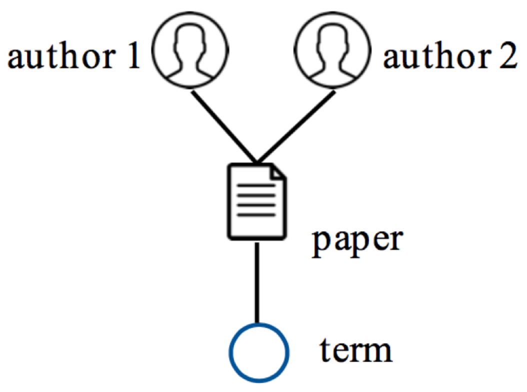


(a) An example digital collection of massive scientific papers.

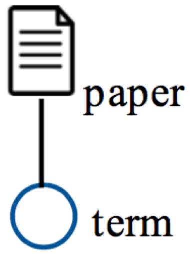
(b) An text-rich network view of the example digital collection.

Network Motifs: Contexts from Network

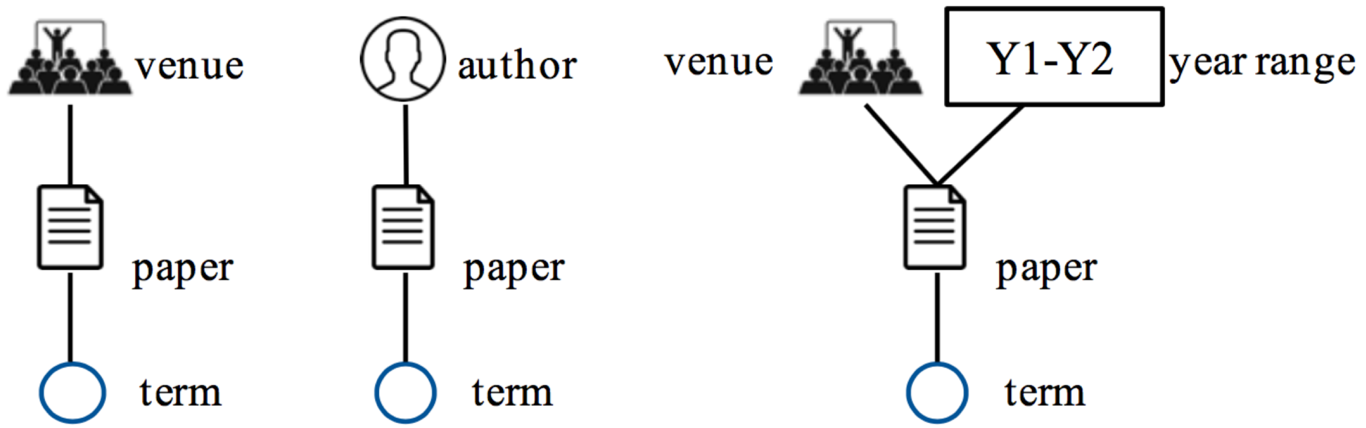
- Motif patterns capture subgraph contexts
  - Those nodes “connecting” two terms describes contexts
- Meta-path can be viewed as a special case of motif
  - T-P-T, T-P-V-P-T, T-P-A-P-T, …



Motif Pattern



Motif Instance

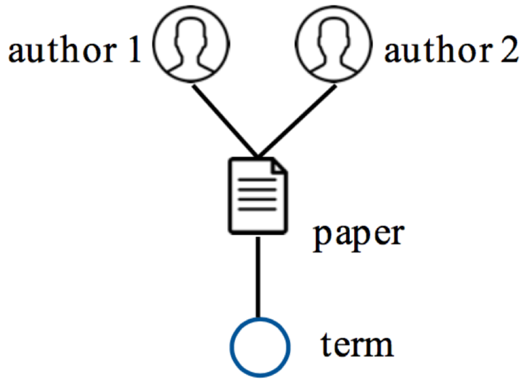


More Motif Patterns

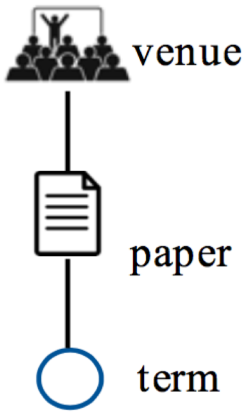
Text-Network Collaboration: **Challenges**

- Motif patterns are not created equally, even given by human experts

More effective at fine-grained levels

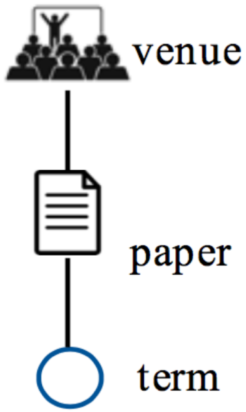


vs.



More effective at coarse-grained levels

- Motif instances of the same motif patterns are not equally informative



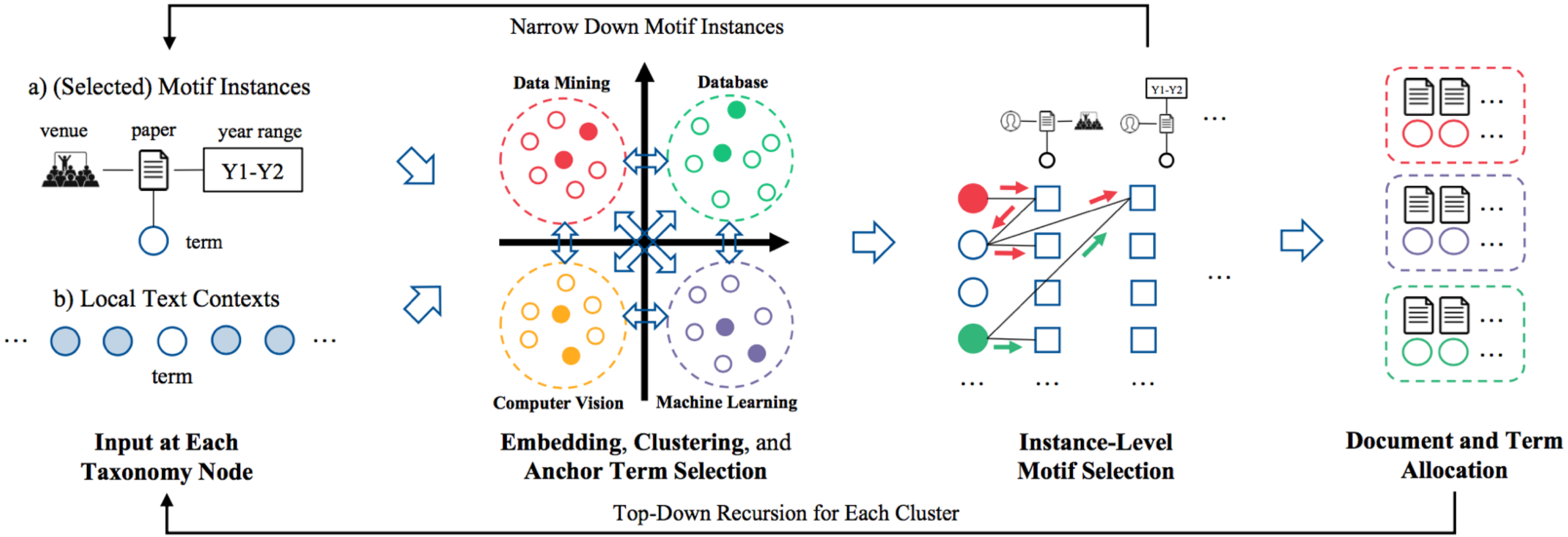
SIGKDD vs. **SIGMOD**

More diverse venues

More focused venues

NetTaxo: Instance-Level Motif Selection

- Starts from term embedding using textual contexts only
- Anchor terms makes the initial clustering results more robust
- Joint term embedding based on selected motif instances & textual contexts



NetTaxo: Joint Clustering Results (CS domain, 1st level)

*

data mining
knowledge discovery
social network
recommender systems
social media
social networks
collaborative filtering
link prediction
social network analysis
web search

natural language
machine translation
statistical machine translation
question answering
dependency parsing
word sense disambiguation
coreference resolution
natural language processing
word order
word alignment

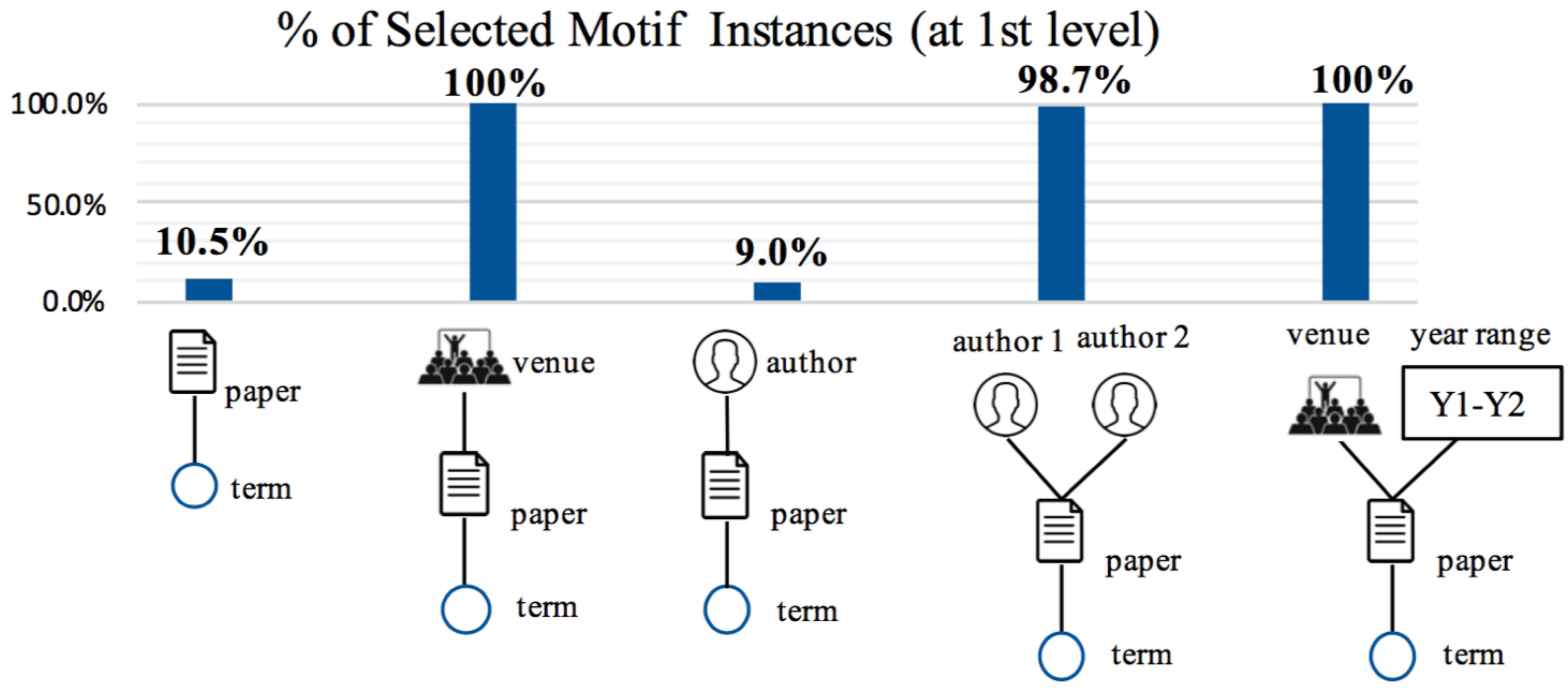
feature selection
reinforcement learning
active learning
model selection
gaussian process
neural networks
learning algorithms
decision trees
function approximation
gaussian processes

...

... ...

NetTaxo: Implicit Motif Pattern Selection

- Instance-level selection implicitly filters useless motif patterns too



Case Study: Selected Motif Instances I

- Interesting motif instances selected at different levels
- Author pairs with more focused research topics are selected at the 2nd level

Hua Wu - Zhanyi Liu

sentiment analysis
semantic features
semantic relations
textual similarity
sentiment words
sentiment classification
…

Roland Kuhn - George F. Foster

source language
bilingual corpora
bilingual word
machine translation
statistical machine translation
bleu score
…

Level 2: NLP -> Sub-Areas

Hua Wu - Zhanyi Liu
Omar F. Zaidan - Chris Callison-Burch
Boxing Chen - Roland Kuhn
Hua Wu - Haifeng Wang
Roland Kuhn - George F. Foster
Yoan Gutiérrez - Andrés Montoyo
John Makhoul - Richard M. Schwartz
…

Case Study: Selected Motif Instances II

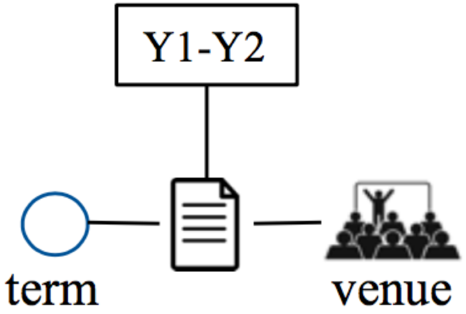
- Interesting motif instances selected at different levels
- Recent, diverse & early, focused venues may help at the 2nd level

CIKM 2010-2014

question answering
information extraction
language models
sentiment analysis
sentiment classification
knowledge base
. . .

Level 2: NLP -> Sub-Areas

ACL 1985-1989
COLING 1985-1989
EACL 1985-1989
COLING 1980-1984
ACL 1980-1984
CIKM 2010-2014
COLING 1990-1994
. . .



NetTaxo: Evaluation Metrics

- **Coherence Measure**
  - Are the terms at the same taxonomy node coherent?
- **Sibling Exclusiveness**
  - Are the terms at a taxonomy node more similar compared to the terms at its sibling nodes?
- **Parent-Child Relationship**
  - Are the relationships between the taxonomy nodes correct?

- All metrics are between 0 and 1; The bigger, the better.

NetTaxo: Experimental Results

- Two domains: CS papers & Yelp reviews
- Compared with ("++" means "enhanced by our phrase mining results")
  - TaxoGen (KDD'18) & HPAM++: Using text data only
  - CATHYHIN++ (ICDM'13, TKDE'18): Using network data only
  - HClusEmbed (NeurIPS'13, WWW'15): combines both term and node embeddings

	DBLP-5					**Yelp-5**				
	Coherence Measure	Sibling Exclusiveness	Parent-Child Relations Precision	Parent-Child Relations Recall	Parent-Child Relations F_1	Coherence Measure	Sibling Exclusiveness	Parent-Child Relations Precision	Parent-Child Relations Recall	Parent-Child Relations F_1
HPAM++	0.796	0.680	0.348	0.451	0.393	0.832	0.740	0.171	0.247	0.202
TaxoGen	0.840	0.740	0.780	0.713	0.745	0.920	0.800	0.650	0.618	0.633
CATHYHIN++	0.880	0.533	0.850	0.744	0.793	0.742	0.420	0.705	0.638	0.670
HClusEmbed	0.624	0.420	0.525	0.409	0.460	0.744	0.560	0.655	0.610	0.632
NetTaxo w/o Selection	0.908	0.680	0.895	0.808	0.849	0.816	0.540	0.668	0.681	0.674
NetTaxo	**0.912**	**0.880**	**0.898**	**0.810**	**0.852**	**0.928**	**0.854**	**0.790**	**0.825**	**0.807**