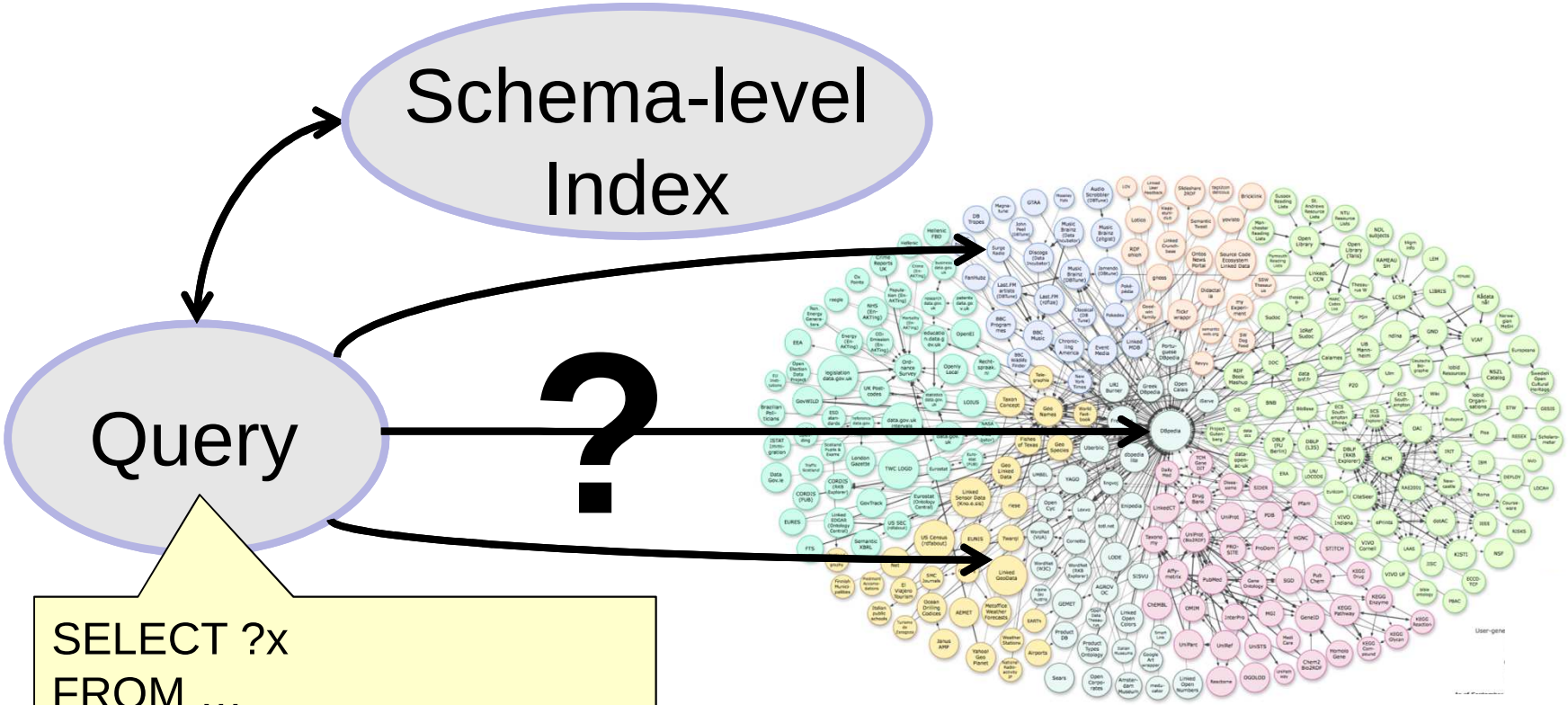
Analyses of Schema Structures on the Linked Open Data Cloud

A. Scherp, T. Gottron, M. Knauf, R. Dividino, G. Gröner
asc@informatik.uni-kiel.de

FG Datenbanken

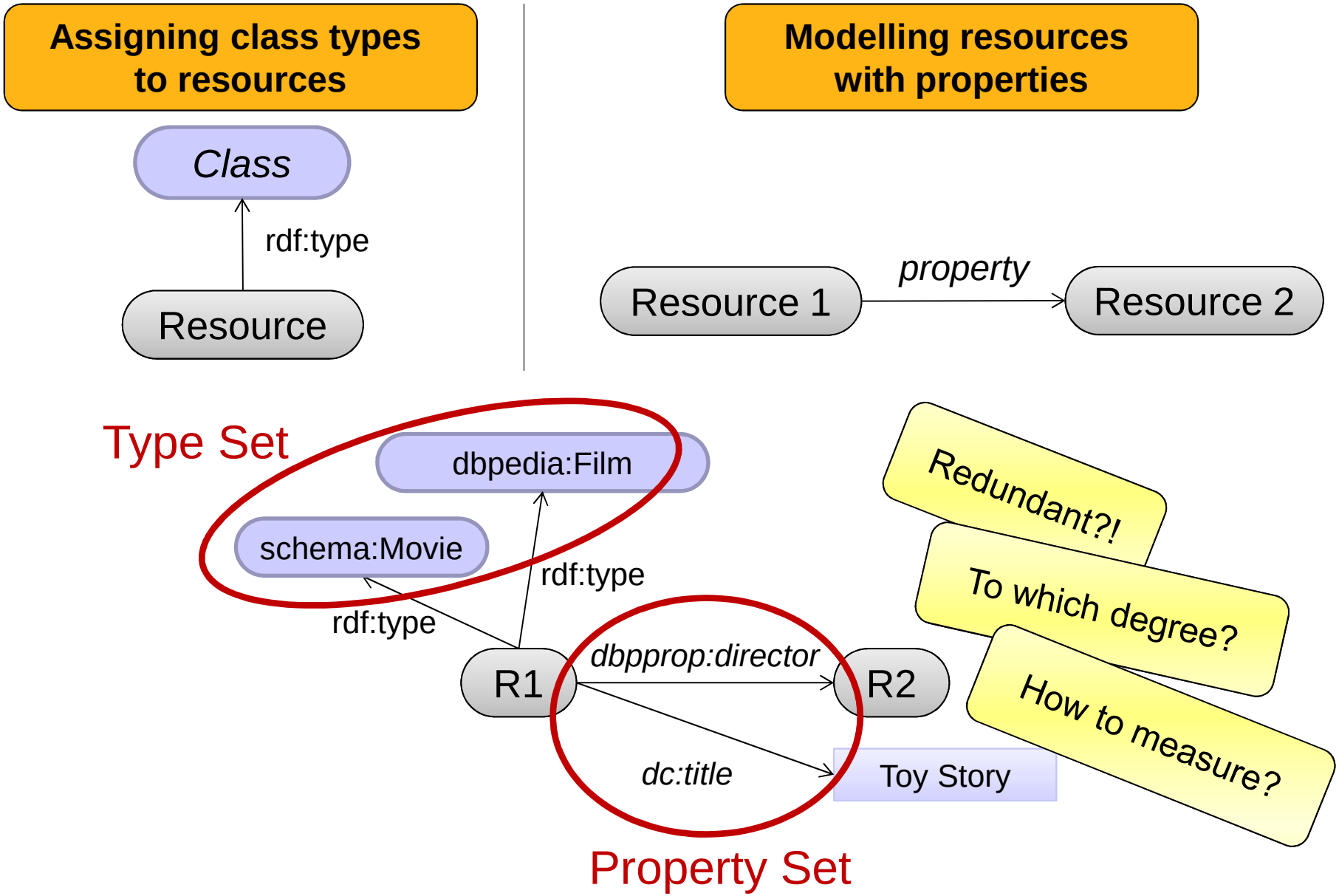
Braunschweig, March 2014

Motivation: Search for LOD Sources



[KGS+12] M. Konrath, T. Gottron, S. Staab, A. Scherp: SchemEX - Efficient construction of a data catalogue by stream-based indexing of linked data. J. Web Sem. 16: 52-58 (2012)

Schema-Modeling with Types and Properties



Overview: Analysis of Schema Structures

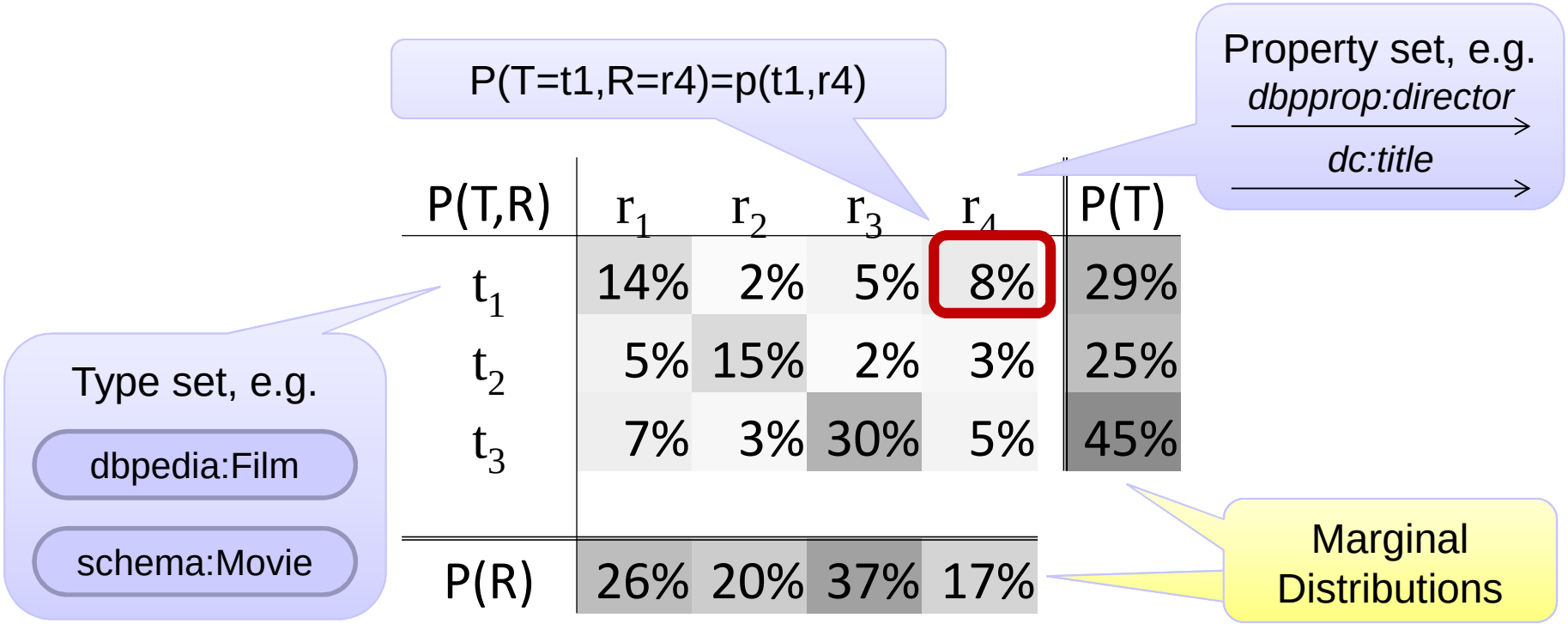
1. Information theoretic analysis over resources
2. Information theoretic analysis over PLDs
3. Schema structures over vocabularies in PLDs
4. Schema changes over time

Overview: Analysis of Schema Structures

1. **Information theoretic analysis over resources**
   a) Information encoded in type sets or property sets?
   b) Information still contained in the properties, once we know the types (and vice versa)?
   c) To which degree explain properties and types each other?
2. Information theoretic analysis over PLDs
3. Schema structures over vocabularies
4. Schema changes over time

Probabilistic Model of Schema Information

- Joint distribution of random variables T and R
- Over type sets (TS) and property sets (PS)
- Probability to observe type set t ∈ TS and property set r ∈ PS:
  $P(T = t, R = r) = p(t, r)$
- p(t,r) can be efficiently computed on large datasets [KGS+12]



P(T,R)	r_1	r_2	r_3	r_4	P(T)
t_1	14%	2%	5%	8%	29%
t_2	5%	15%	2%	3%	25%
t_3	7%	3%	30%	5%	45%
P(R)	26%	20%	37%	17%	

a) Information Encoded in TS / PS?

- Entropy of marginal distribution: probability that a resource has a specific property set (or type set) [Sha48]

$$H(R) = -\sum_{r \in PS} P(R = r) \cdot \log_2(P(R = r))$$

- Normalized marginal entropy: $H_0(R) = \frac{H(R)}{\log_2(|PS|)}$

Each PS is observed once!

	r_1	r_2	r_3	r_4	
P(R)	26%	20%	37%	17%	$H_0(R) = 0.967$
P(R)	1%	97%	1%	1%	$H_0(R) = 0.121$
P(R)	25%	25%	25%	25%	$H_0(R) = 1.000$

- Higher value indicates that R carries more information

a) Normalized Marginal Entropy

- Results for $H_0(R)$ and $H_0(T)$ on segments of BTC'12 data set

Data set	Triples	$H_0(T)$		$H_0(R)$
Rest	22.3M	0.252	<	0.366
Datahub	910.1M	0.263	>	0.250
Dbpedia	198.1M	0.093	<	0.324
Freebase	101.2M	0.127	<	0.166
Timbl	204.8M	0.214	<	0.276

- Tendencies
  - Entropy of property sets is higher, i.e. carry more information
  - No very high values
  - No values close to zero

b) Information in TS / PS once the other is known

- Conditional entropy: knowing type set t of a resource, what is the probability of showing a specific property set?

$$H(R \mid T = t) = -\sum_{r \in PS} P(r \mid T = t) \cdot \log_2(P(r \mid T = t))$$

- Expected conditional entropy: treating conditional entropy as random variable

$$H(R \mid T) = -\sum_{t \in TS} P(T = t) \cdot H(R \mid T = t)$$

Compute conditional entropy H(R|T=t) for all t

b) Expected conditional entropy

- Results for H(T|R) and H(R|T) on segments of BTC'12 data set

Data set	H(T|R)		H(R|T)		H(T)	H(R)
Rest	0.289	<	2.568		2.428	4.708
Datahub	1.319	>	0.876		3.904	3.460
Dbpedia	0.688	<	4.856	<	1.856	6.027
Freebase	0.286	<	1.117		2.037	2.868
Timbl	0.386	<	1.464		2.568	3.646

- Tendencies
  - Type set of a resource tells little about its properties, property sets tell more about type sets
  - Given information reduces the entropy

c) Redundancy in TS and PS?

- Mutual information: degree to which properties and types mutually explain each other [CT91]

$$I(T,R) = \sum_{t \in TS} \sum_{r \in PS} p(t,r) \cdot \log_2 \frac{p(t,r)}{P(T=t) \cdot P(R=r)}$$

- Normalized mutual information

$$I_0(T,R) = \frac{I(T,R)}{\min(H(T), H(R))}$$

c) Normalized Mutual Information

Ordered by decreasing redundancy

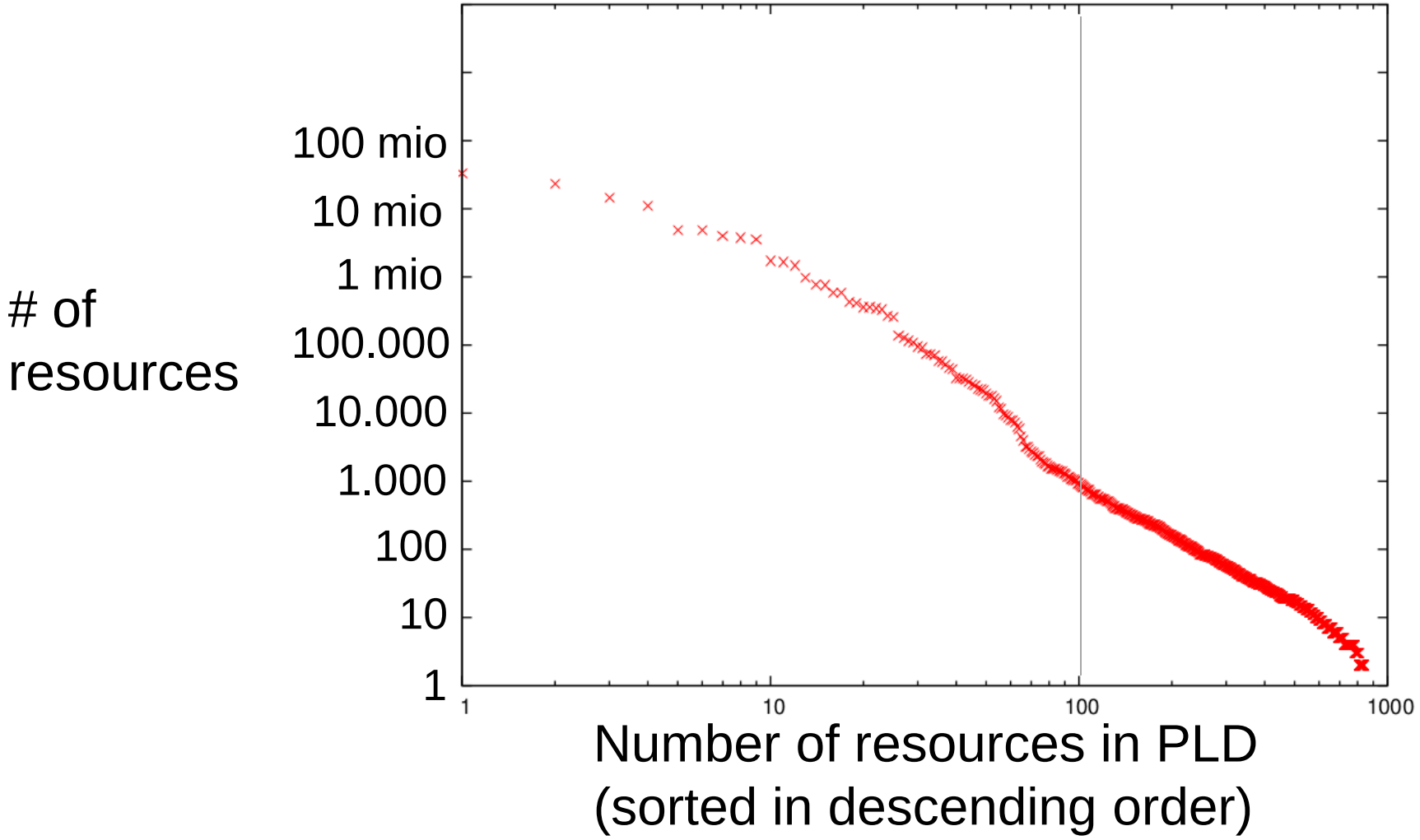
Data set	$I_0(T,R)$
Rest	0.881
Freebase	0.860
Timbl	0.850
Datahub	0.747
Dbpedia	0.635

- Tendencies
  - Relatively high redundancy (between 64% to 88%)
  - Freebase: pre-defined schema, curated with high effort
  - Timbl: narrow domain (FOAF profiles)
  - DBpedia: de-centralized use of type sets and property sets

Overview: Analysis of Schema Structures

1. Information theoretic analysis over resources
2. **Information theoretic analysis over PLDs**
   a) How is the information distributed over PLDs?
   b) What are typical values of entropy and redundancy in PLDs?
   c) For which PLD do we observe a outlier behaviour?
3. Schema structures over vocabularies in PLDs
4. Schema changes over time

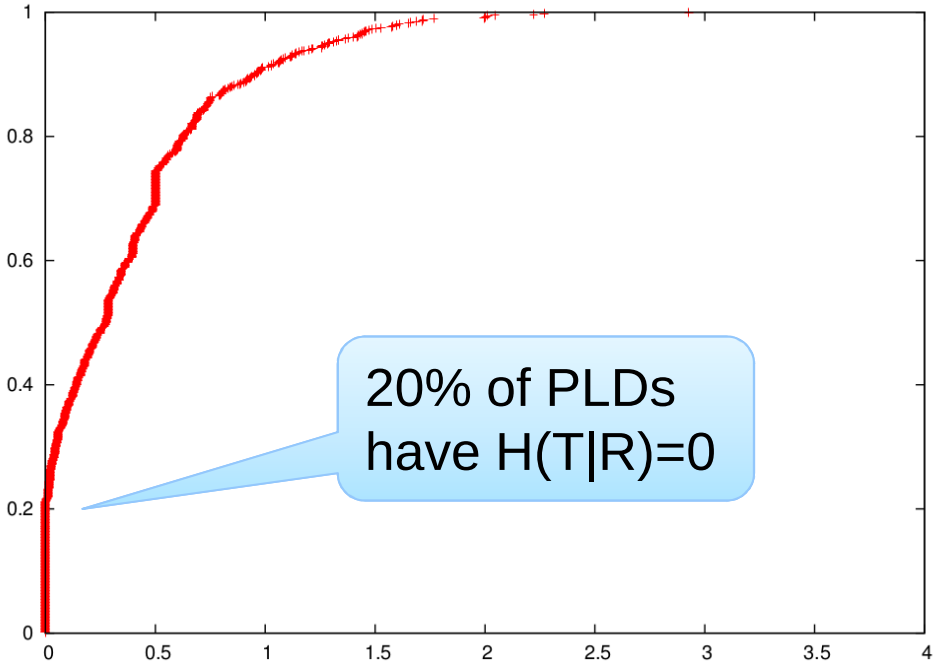
Distribution of 840 PLDs in BTC‘12



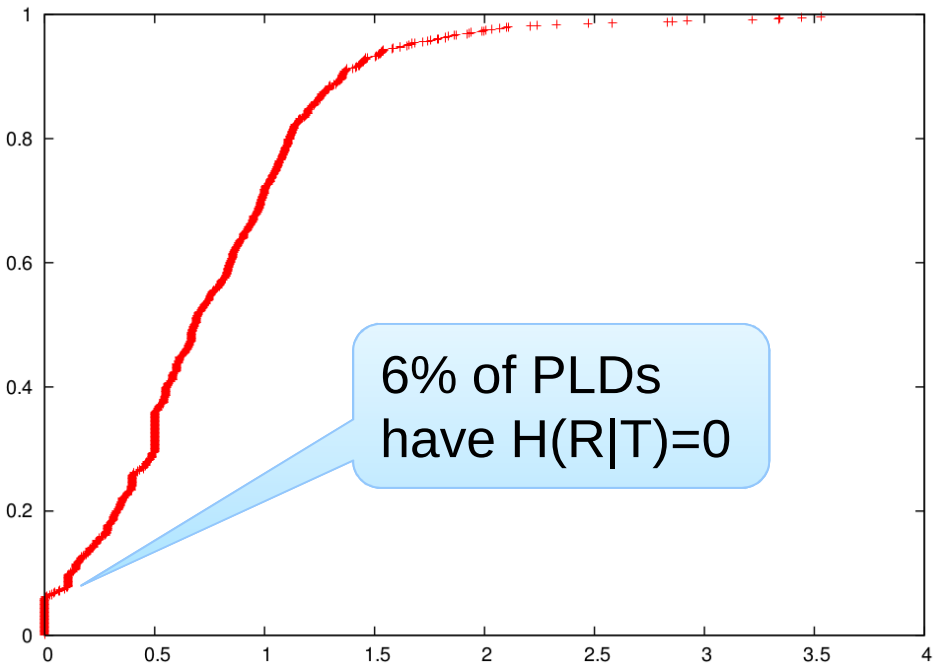
- Top 100 PLDs make about 99,84% of all resources
- Idea: model previous measures as distribution over 840 PLDs

Cumulative Distribution of Expected Conditional Entropy H(T|R) and H(R|T)

H(T|R=r)

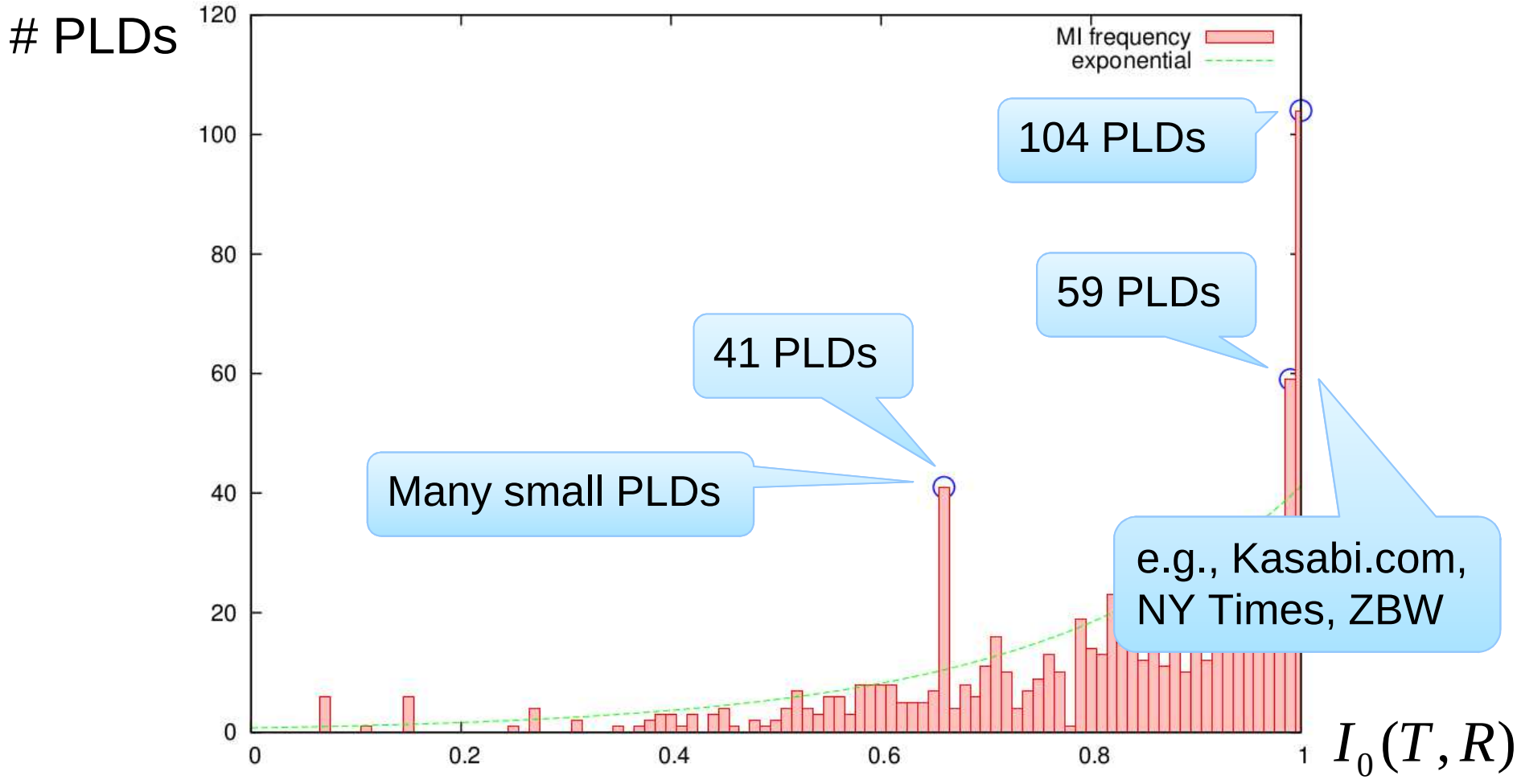


H(R|T=t)



- Increase slighter for H(R|T) → reflects global trend that types are less important
- Observation is not an artifact of small PLDs (not shown)

Normalized Mutual Information over PLDs



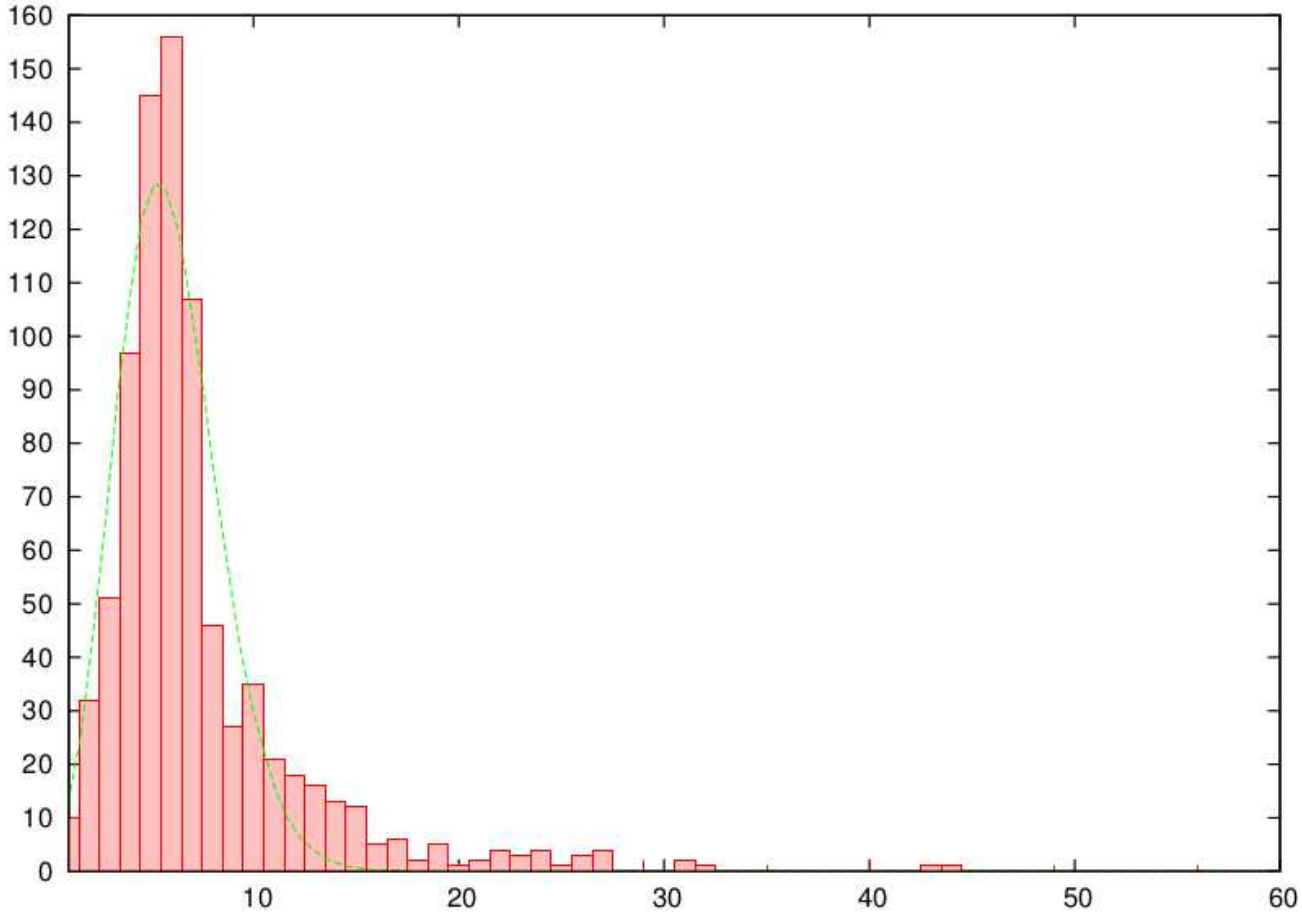
- Majority of PLDs have I_0 of 0.5 or higher
- Top 100 PLDs are equally distributed (not shown)

Overview: Analysis of Schema Structures

1. Information theoretic analysis over resources
2. Information theoretic analysis over PLDs
3. **Schema structures over vocabularies in PLDs**
   a) Strength of vocabularies in specific PLDs?
   b) Patterns one can observe in use of vocabularies?
4. Schema changes over time

of Vocabularies used in the PLDs

of PLDs using exactly that many vocabs



vocabs

- Most PLDs use between 3 to 8 vocabularies (M=6)
- Aligns with LOD best practice using a *good mix* of vocabularies

Strength of Vocabulary in PLD

- Simple metric to determine importance of vocabulary
- Percentage of resources in PLD described by vocabulary V (either via a RDF type or property)

$$strength(V, PLD) = \frac{|\{res \in PLD \mid res \text{ is described using } V\}|}{|PLD|}$$

- Example
  - 7 out of 10 resources use FOAF makes a strength of 0.7

Example: Vocabulary Distribution

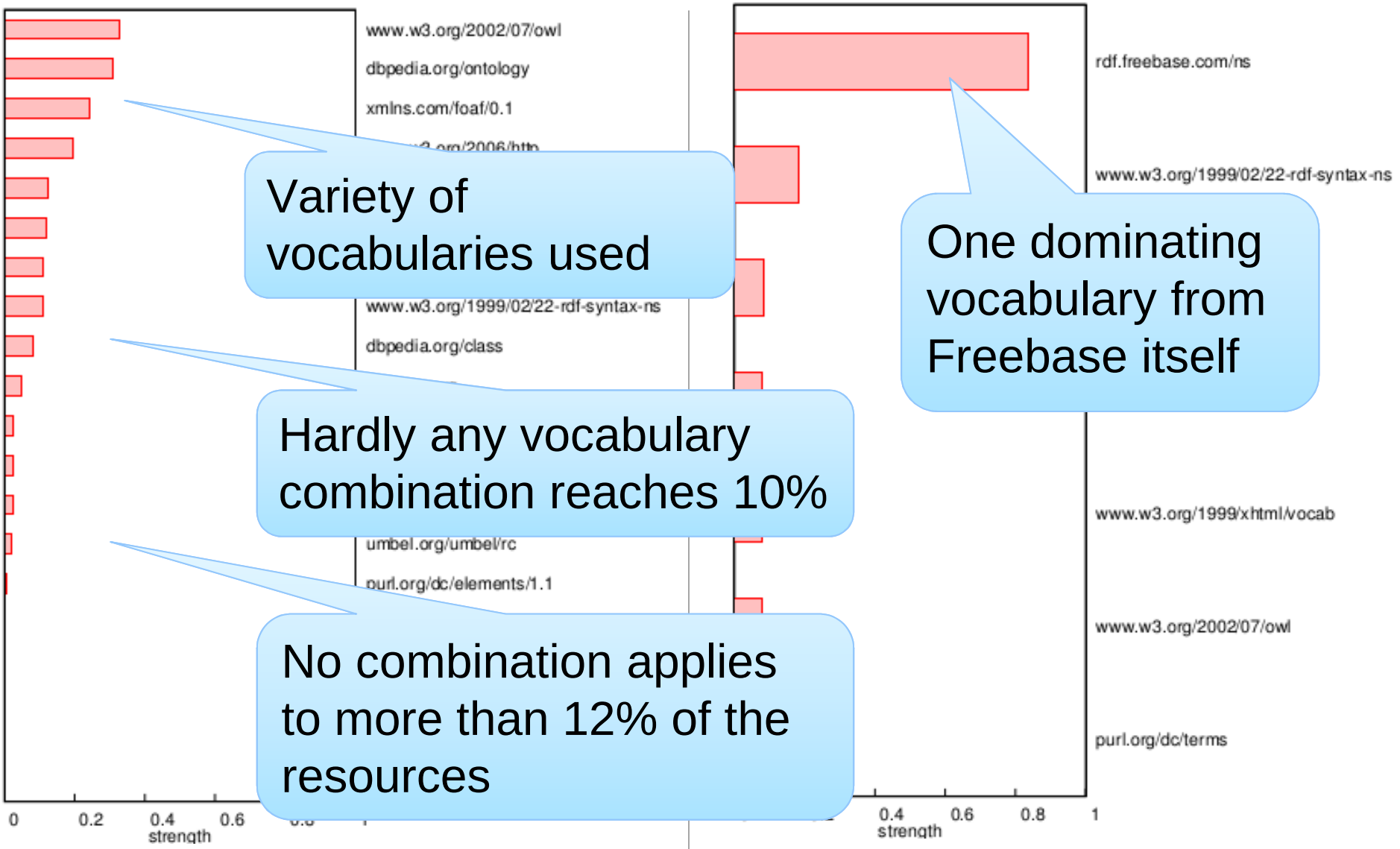
66%: RDF, DC Terms, and FOAF

32%: use RDF, RDFS, DC Terms, FOAF, and domain specific vocabulary

Two levels of vocabulary strengths

Kasabi.com $I_0 = 0.99$

Example: Vocabulary Distribution



Variety of vocabularies used

Hardly any vocabulary combination reaches 10%

No combination applies to more than 12% of the resources

One dominating vocabulary from Freebase itself

DBPedia.org $I_0 = 0.62$

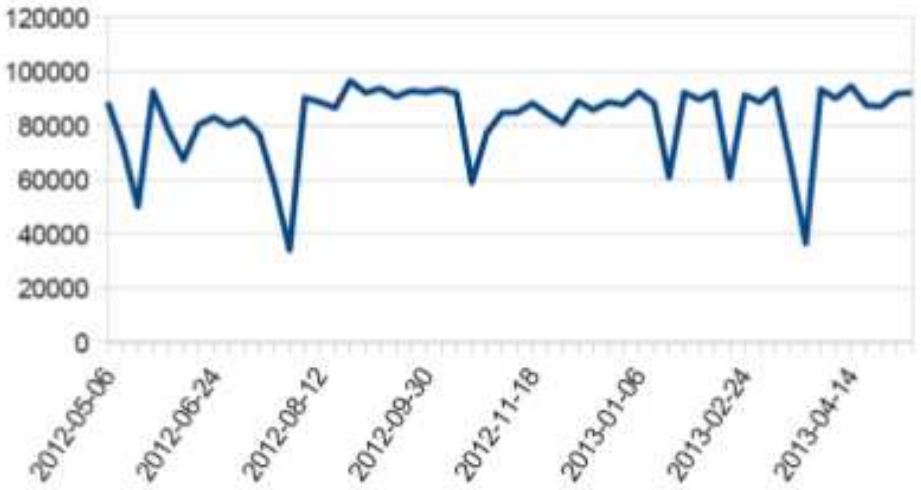
Freebase.com $I_0 = 0.85$

Overview: Analysis of Schema Structures

1. Information theoretic analysis over resources
2. Information theoretic analysis over PLDs
3. Schema structures over vocabularies
4. **Schema changes over time in PLDs**
   a) Vocabularies are highly static [KAU+13], but use of vocabulary terms is highly dynamic [DSG+13]

Number of Vocabulary Combinations

- Extended Characteristic Set (ECS, cf. [NM11]): combined use of a type set and property set in the data
- Weekly snapshots of about ~100 Mio triples (5/'12-5'13) [KAU+13]



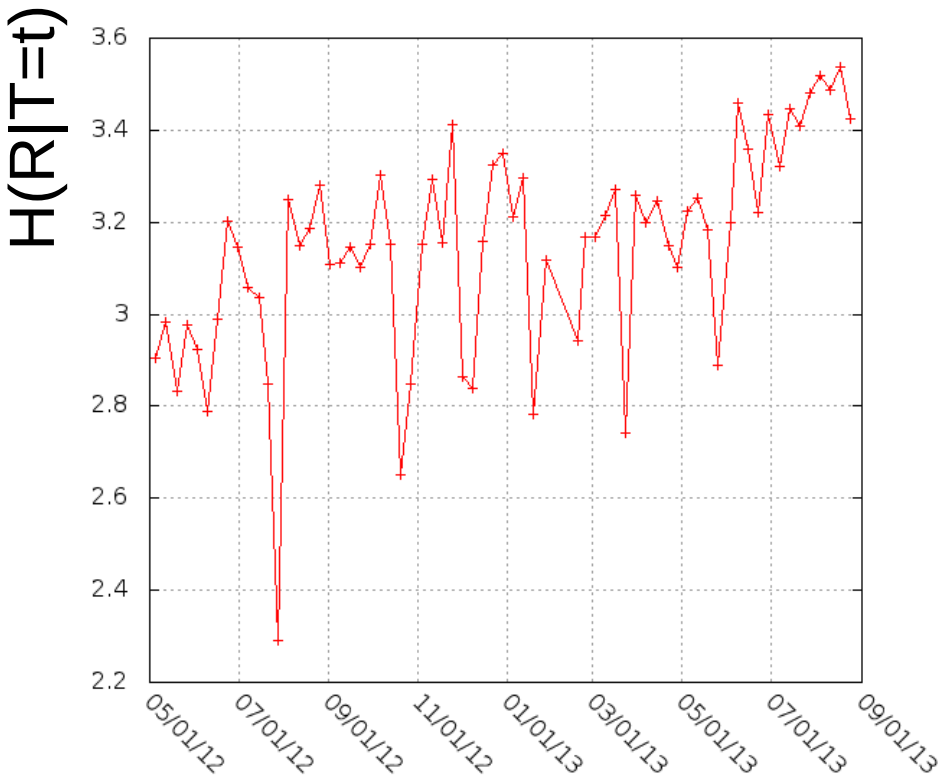
Total # ECS in weekly snapshot

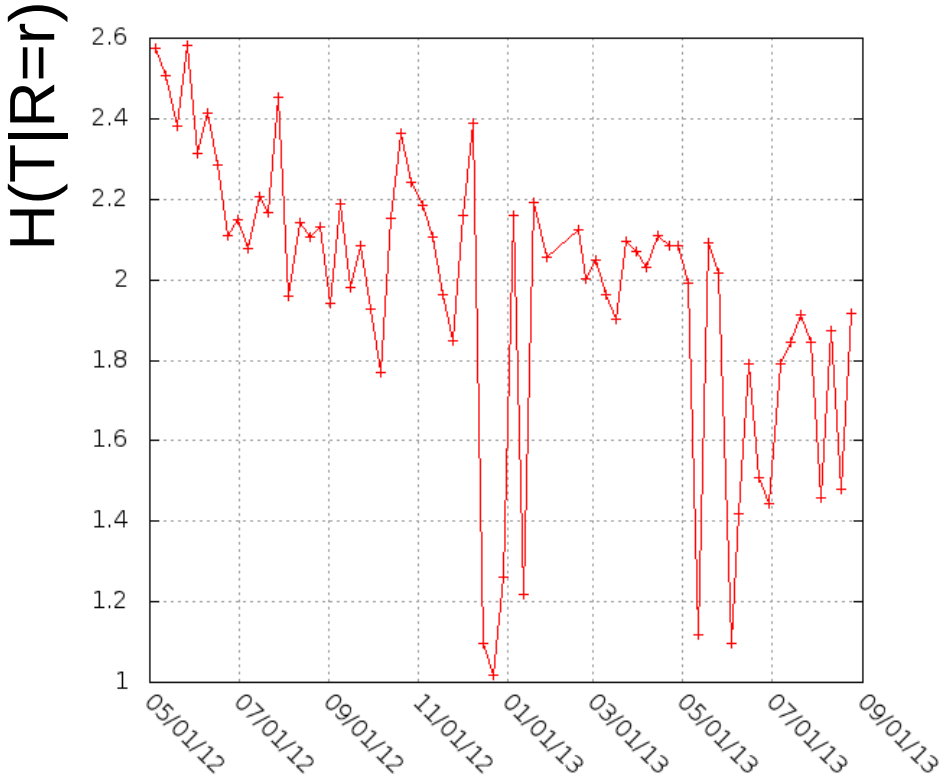
ECS unchanged to previous week (orange) / first week (blue)

- Total number of ECS relatively stable over time (M=83.898)
- 35% of ECS remain unchanged from first to last snapshot
- 73% unchanged from week to week (new: 27%, deleted: 29%)

Expected Cond. Entropy over Time

- So far: properties are more important than types
- Do types become less important over time?





- Further investigations w.r.t. changes and *dynamics* in schema

Take Home Messages

1. Information theoretic analysis over resources

Properties are more important than types

Redundancy 63-88%

2. Information theoretic analysis over PLDs

Types irrelevant for 20% of PLDs

3. Schema structures over vocabularies in PLDs

Two patterns of vocabulary uses

4. Schema changes over time

Use of vocabularies is indeed highly dynamic

References

[GKS14] Thomas Gottron, Malte Knauf, Ansgar Scherp: Analysis of Schema Structures in the Linked Open Data Graph based on Unique Subject URIs, Pay-level Domains, and Vocabulary Usage, Distributed and Parallel Databases, February 2014.

[DSG+13] Renata Queiroz Dividino, Ansgar Scherp, Gerd Gröner, Thomas Gottron: Change-a-LOD: Does the Schema on the Linked Data Cloud Change or Not? COLD 2013

[KGS+12] Mathias Konrath, Thomas Gottron, Steffen Staab, Ansgar Scherp: SchemEX - Efficient construction of a data catalogue by stream-based indexing of linked data. J. Web Sem. 16: 52-58 (2012)

[Sha48] Claude Shannon: A mathematical theory of communication. Bell System Technical Journal 27, 379-423 and 623-656 (1948)

[CT91] T. M. Cover, J. A. Thomas: Elements of Information Theory. Wiley-Interscience (1991)

[KAU+13] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, Aidan Hogan: Observing Linked Data Dynamics. ESWC 2013: 213-227

[NM11] Thomas Neumann, Guido Moerkotte: Characteristic sets: Accurate cardinality estimation for RDF queries with ultiple joins. ICDE 2011: 984-994