Non-Parametric **Class Completeness** Estimators for Collaborative **Knowledge Graphs**

The Case of Wikidata

Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux

ISWC 2019, Auckland

Agenda

- Motivation
- Species Richness Estimators
- Class Completeness Estimators
- Evaluation / Application

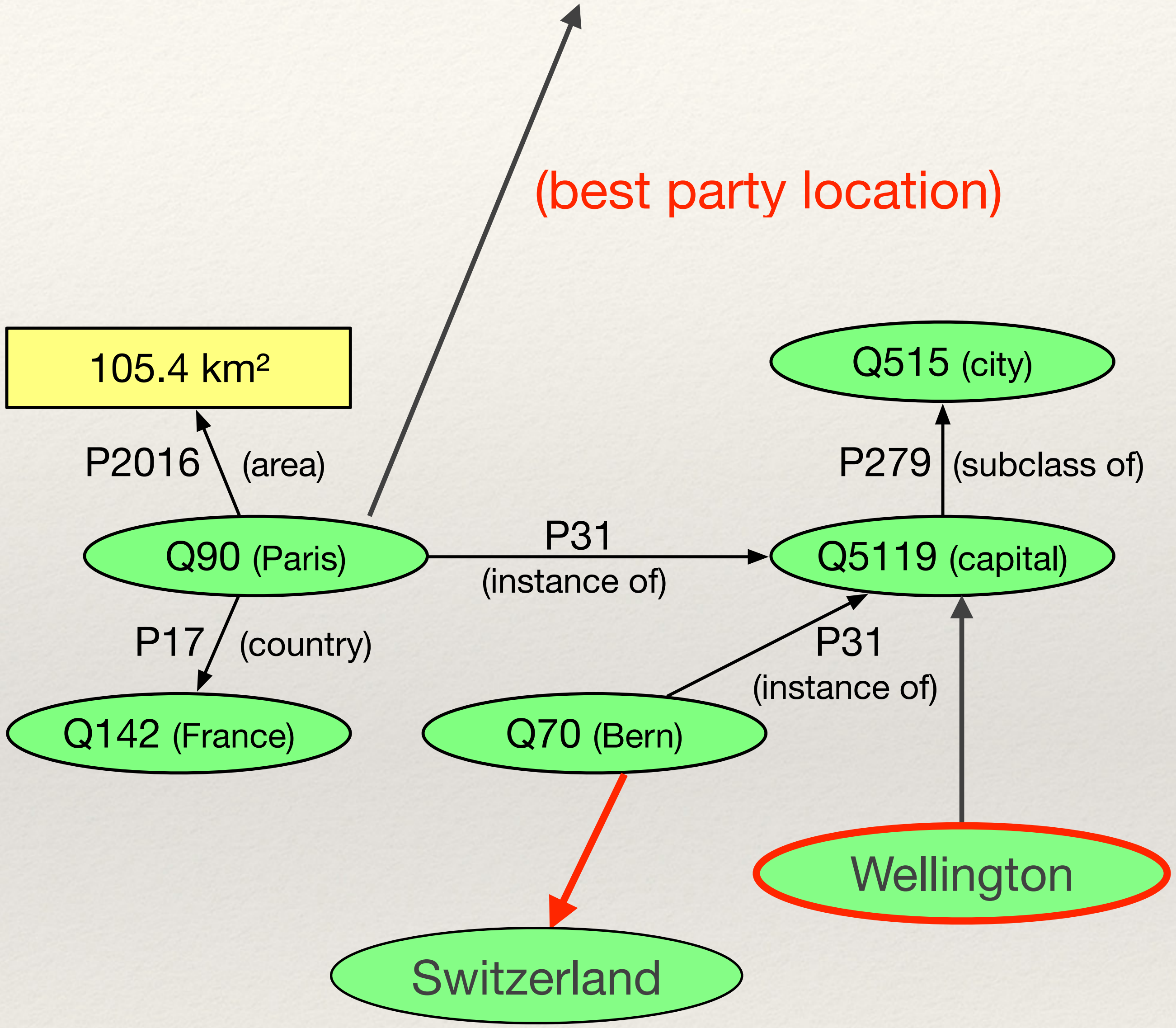
KG Completeness

Schema completeness

Property completeness

Interlinking completeness

Class completeness



After: Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: A survey. Semantic Web Journal

Errors through incomplete Classes

Missing entities can lead to **wrong conclusions:**

"There are no volcanos in New Zealand, so no need for an early warning system."

Missing entities can **bias statistics**:

"There are more Skyscrapers in Auckland, compared to NY, so Auckland is bigger."

The Question

How can we know if we have all real world entities of a class **C** in our Knowledge Base?

How many **Volcanos** are there?
How many **Hospitals** are there?
How many …….. are there?

$\mathbf{I}_C = \{\mathbf{I}_1 , \ldots, \mathbf{I}_N \}$

How many **I** has $\mathbf{I}_C$?

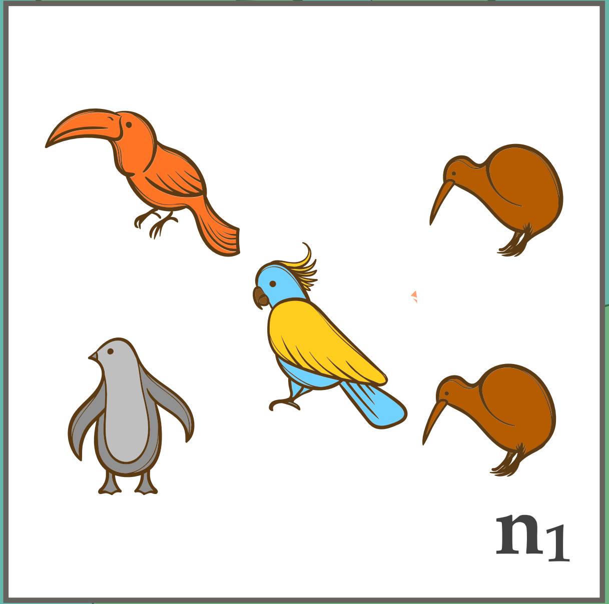
$N = |\mathbf{I}_C |$

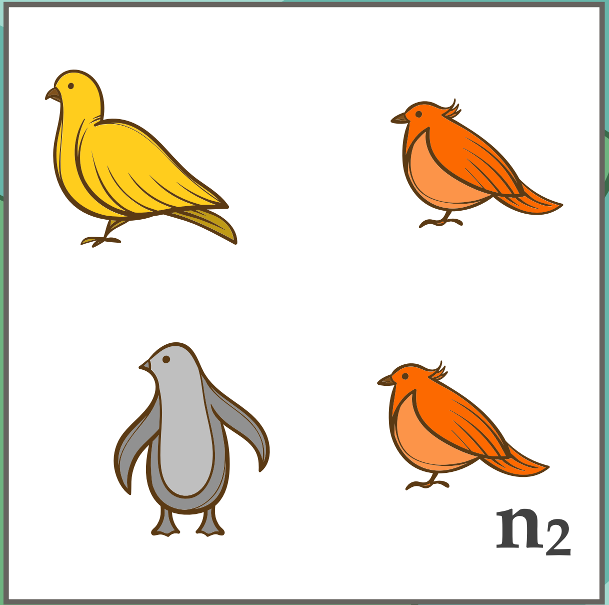
How many **Mountains** are there?

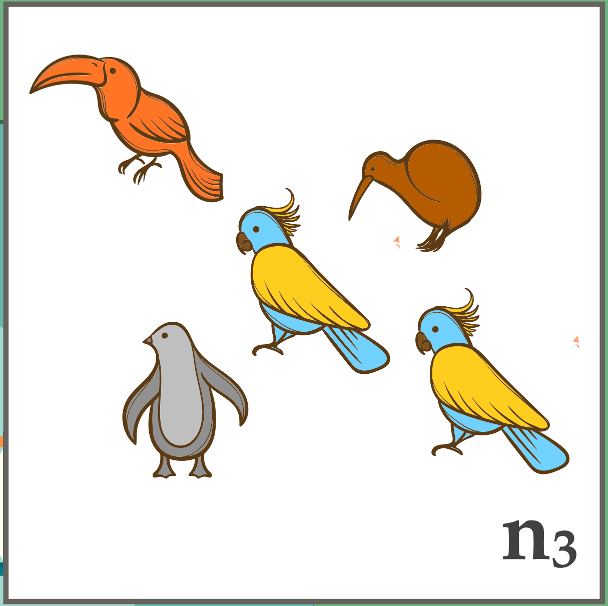
$$\left| I_{\text{Mountains}} \right|$$

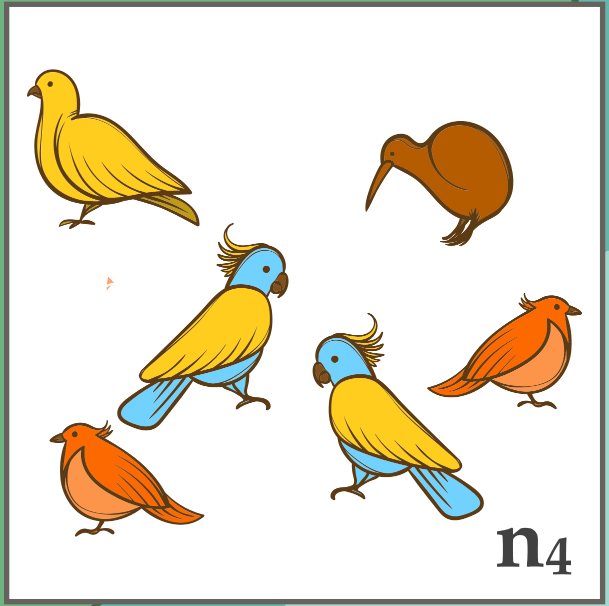
Species Richness Estimators

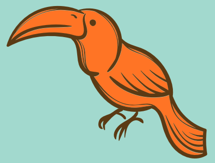
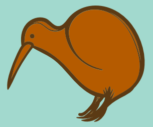
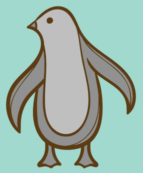
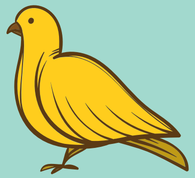
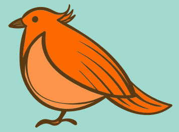
Species Richness Estimators





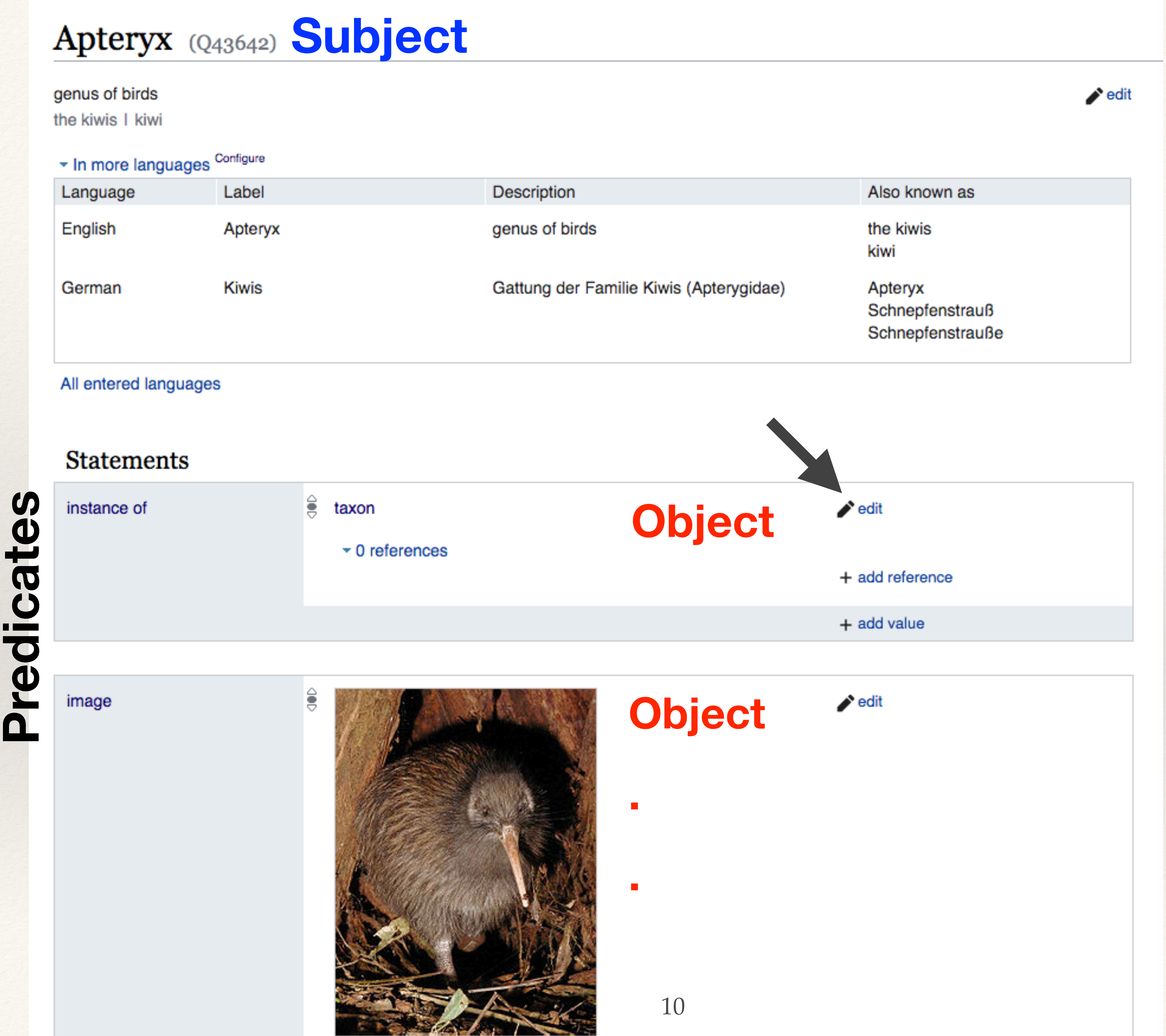




	n_1	n_2	n_3	n_4
	I		I	
	I		I	I
	I	I	I	
	I		I	I
		I		I
		I		I
f_1	4	5	2	0
f_2	0	1	3	3
f_3	0	0	2	3

Collaborative Knowledge Graphs

Wikidata



Apteryx (Q43642) **Subject**

genus of birds

the kiwis | kiwi

Language	Label	Description	Also known as
English	Apteryx	genus of birds	the kiwis kiwi
German	Kiwis	Gattung der Familie Kiwis (Apterygidae)	Apteryx Schnepfenstrauß Schnepfenstrauße

Statements

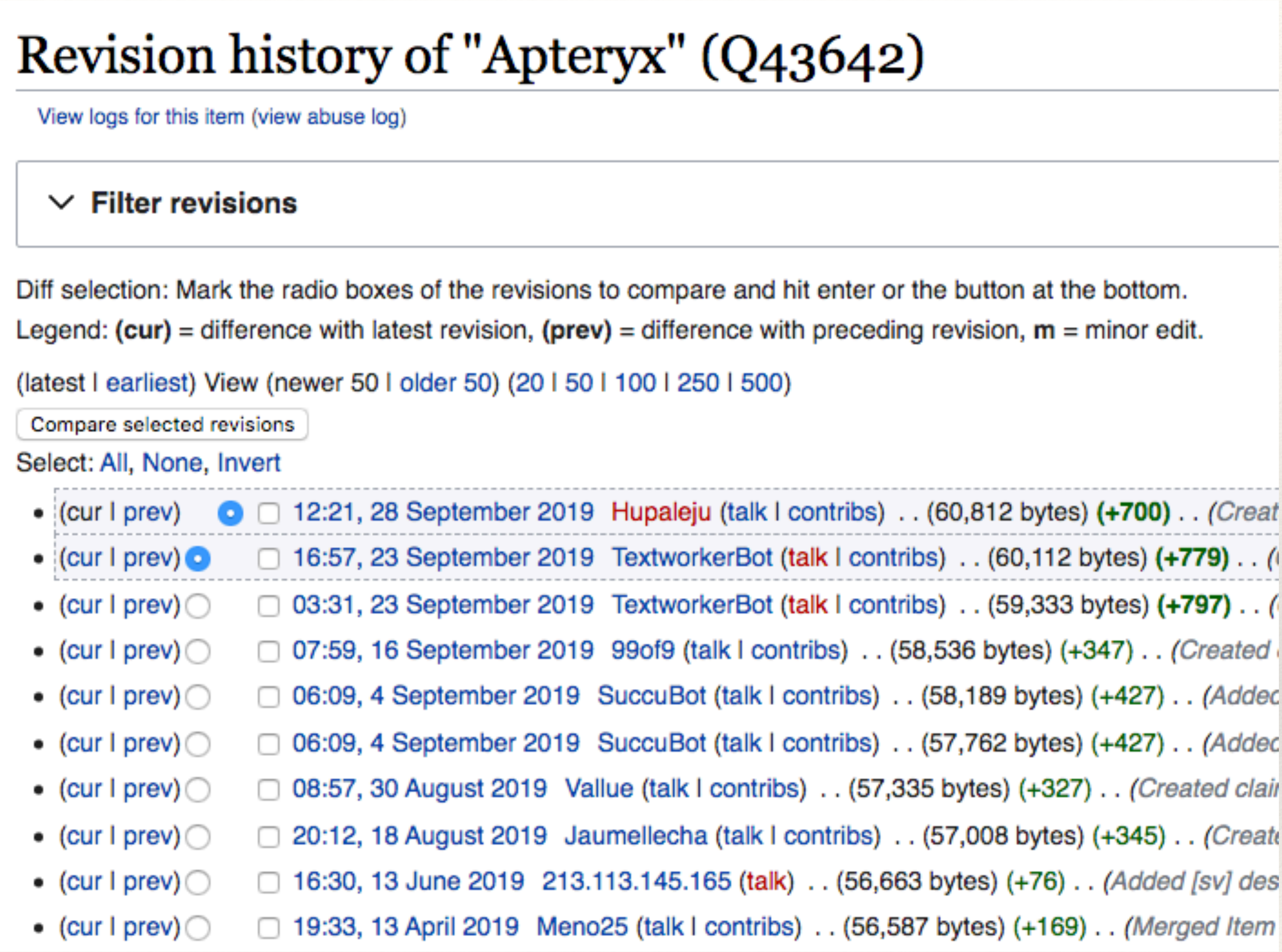
Predicates

instance of — taxon **Object**

image — **Object**

-
-

Wikidata Edits



Revision history of "Apteryx" (Q43642)

View logs for this item (view abuse log)

Filter revisions

Diff selection: Mark the radio boxes of the revisions to compare and hit enter or the button at the bottom.
Legend: **(cur)** = difference with latest revision, **(prev)** = difference with preceding revision, **m** = minor edit.

(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

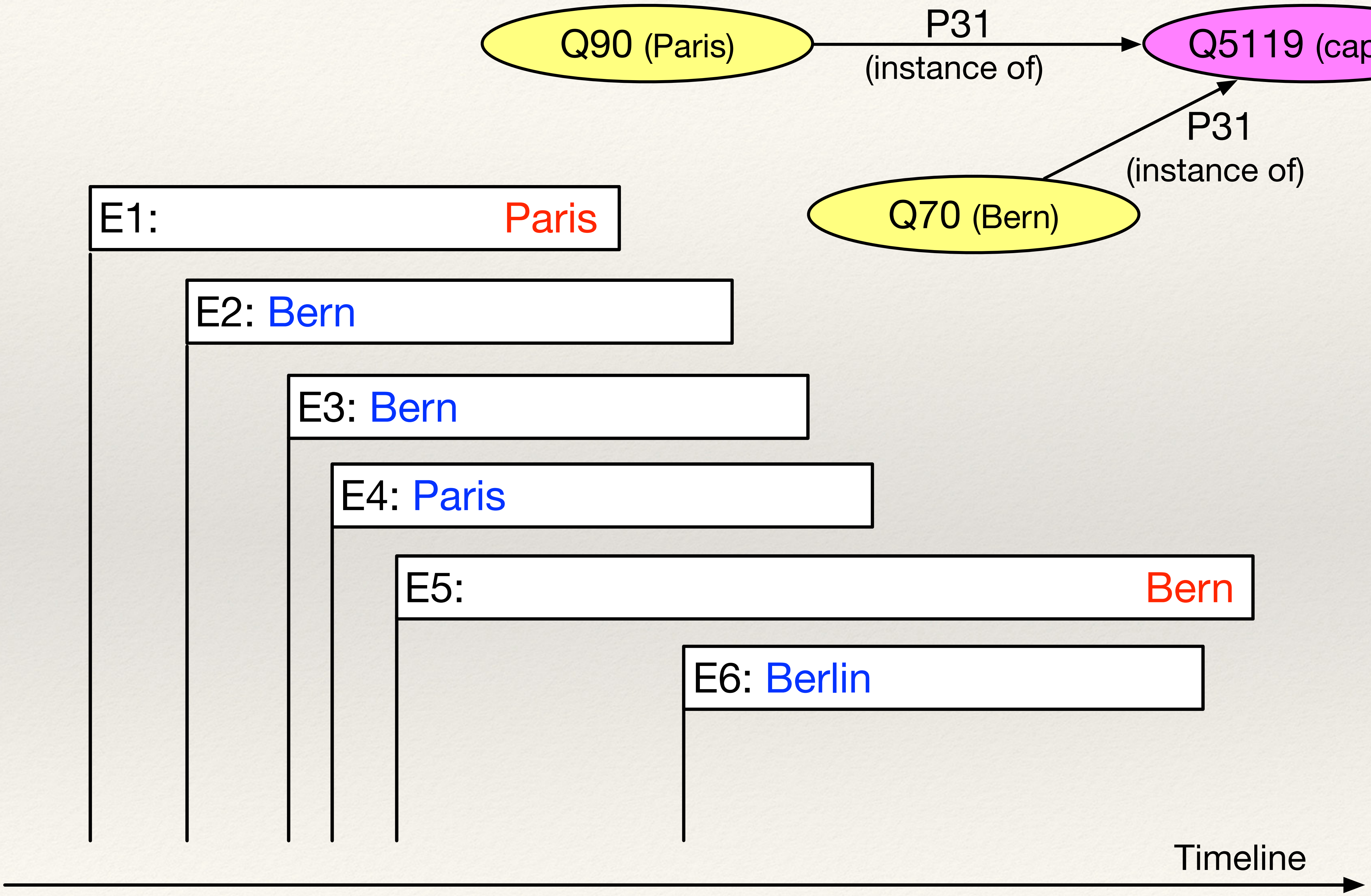
Select: All, None, Invert

- (cur | prev) ● □ 12:21, 28 September 2019 Hupaleju (talk | contribs) . . (60,812 bytes) **(+700)** . . *(Creat*
- (cur | prev) ● □ 16:57, 23 September 2019 TextworkerBot (talk | contribs) . . (60,112 bytes) **(+779)** . . (
- (cur | prev) ○ □ 03:31, 23 September 2019 TextworkerBot (talk | contribs) . . (59,333 bytes) **(+797)** . . (
- (cur | prev) ○ □ 07:59, 16 September 2019 99of9 (talk | contribs) . . (58,536 bytes) (+347) . . *(Created*
- (cur | prev) ○ □ 06:09, 4 September 2019 SuccuBot (talk | contribs) . . (58,189 bytes) (+427) . . *(Addec*
- (cur | prev) ○ □ 06:09, 4 September 2019 SuccuBot (talk | contribs) . . (57,762 bytes) (+427) . . *(Addec*
- (cur | prev) ○ □ 08:57, 30 August 2019 Vallue (talk | contribs) . . (57,335 bytes) (+327) . . *(Created clai*
- (cur | prev) ○ □ 20:12, 18 August 2019 Jaumellecha (talk | contribs) . . (57,008 bytes) (+345) . . *(Create*
- (cur | prev) ○ □ 16:30, 13 June 2019 213.113.145.165 (talk) . . (56,663 bytes) (+76) . . *(Added [sv] des*
- (cur | prev) ○ □ 19:33, 13 April 2019 Meno25 (talk | contribs) . . (56,587 bytes) (+169) . . *(Merged Item*

Edits: 161’445'153

Wikidata Edits

Classes: 54'698



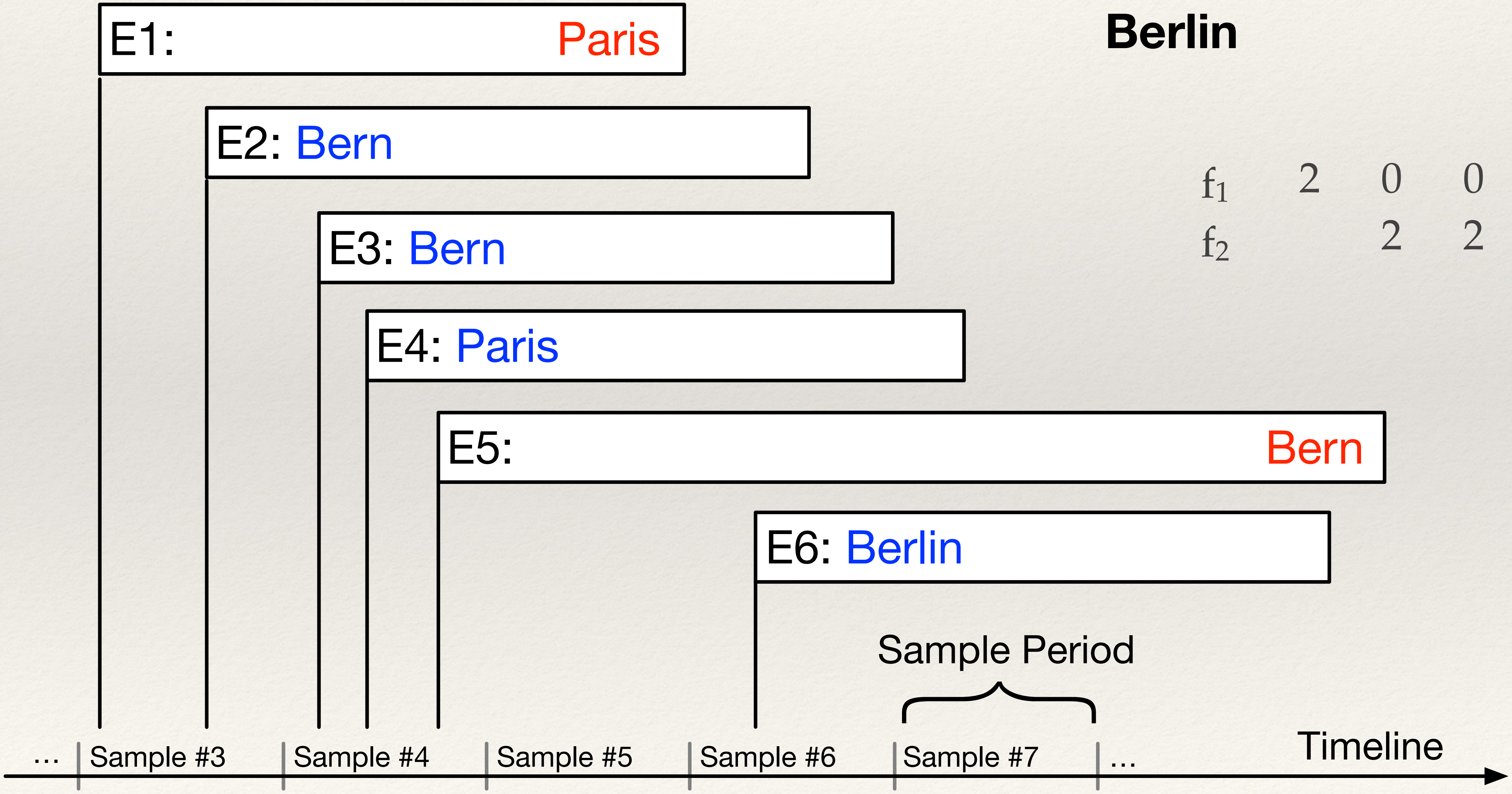
Observations

Edits: 161’445'153

Classes: 54'698

Observations: 370’250'842

	n_3	n_4	n_5	n_6
Paris	I	I		
Bern	I	I		
Berlin				I
f_1	2	0	0	1
f_2		2	2	2



Class Completeness Estimators

Class Completeness Estimators

Jack1	Jackknife Estimators
N1-UNIF	Sample Coverage and the Good-Turing Estimator
SOR	Singleton Outliers Reduction
Chao92	Abundance-based Coverage Estimator

Sample Coverage
and the Good-Turing Estimator

$$\hat{N}_{\text{N1-UNIF}} = \frac{D}{\hat{S}} = \frac{D}{1 - \frac{f_1}{n}}$$

D	Distinct Entities

$$S = \sum^{\cancel{N}} \cancel{p_i} \mathbb{1}[X_i > 0]$$

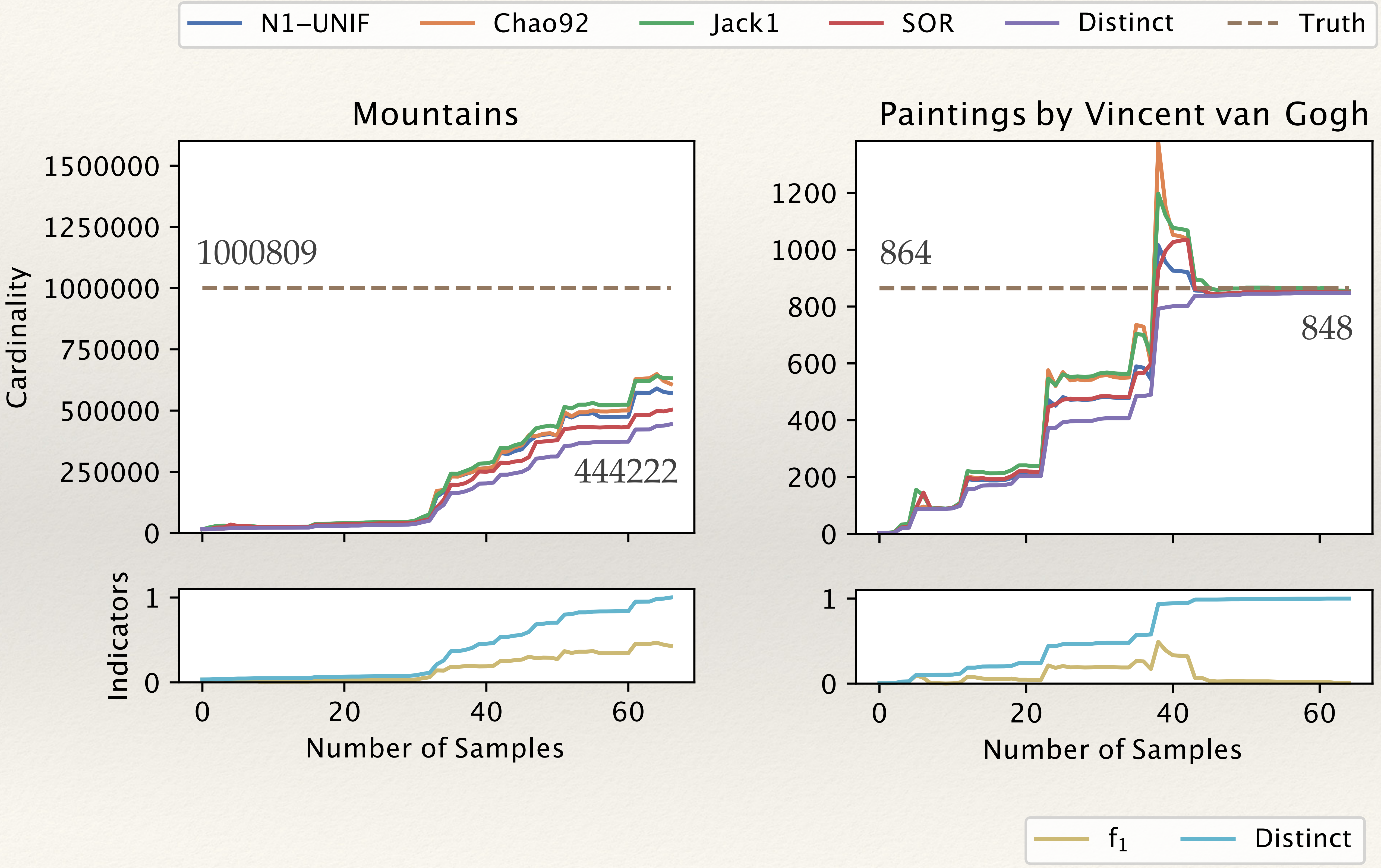
S	*Sample Coverage*
N	True class Size
pᵢ	Probability to Observe
Xᵢ	Frequency of Observation

$$\hat{S} = 1 - \frac{f_1}{n}$$

Ŝ	*Good-Turing Estimator*
f₁	Instances observed once
n	Number of Observations

Evaluation

Evaluation


N1–UNIF
Chao92
Jack1
SOR
Distinct
Truth
Mountains
Paintings by Vincent van Gogh
Cardinality
1500000
1250000
1000000
750000
500000
250000
0
1000809
444222
1200
1000
800
600
400
200
0
864
848
Indicators
1
0
20
40
60
Number of Samples
f_1
Distinct

Application

Convergence Metric

$$\rho = \frac{\sum_{i=k-w}^{k} \frac{|\hat{N}_i - D_i|}{D_i}}{w}$$

$\hat{\boldsymbol{N}}_i$ Entities Estimate per Period

$\boldsymbol{D}_i$ Distinct Entities per Period

$\boldsymbol{w}$ Window

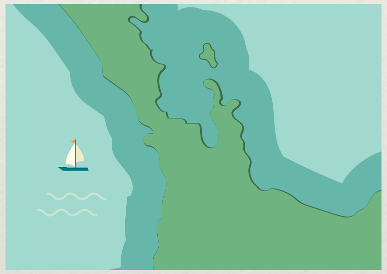
SOR	$\rho < 0.001$	Distinct
municipality of Japan	0.0000	739
Philippine TV series	0.0009	822
Landgemeinde of Austria	0.0000	1,116
district of China	0.0009	975
nuclear isomer	0.0002	1,322
international border	0.0000	529
commune of France	0.0001	34,937
village of Burkina Faso	0.0005	2,723
supernova	0.0005	5,906
township of Indiana	0.0002	999

SOR	$\rho > 0.1$	Distinct
urban beach	0.1759	683
hydroelectric power station	0.2975	2,936
aircraft model	0.1800	3,919
motorcycle manufacturer	0.1758	690
local museum	0.1760	1,150
waterfall	0.1942	5,322
race track	0.2783	946
film production company	0.2107	2,179
red telephone box	0.3469	2,716
mountain range	0.2390	21,390

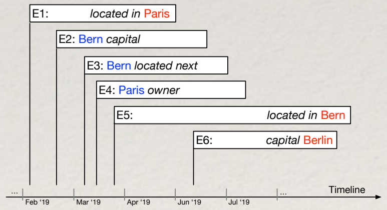
Wrap-Up

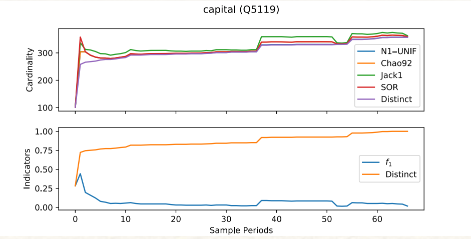
- The edit history of a KG can be used to inform statistical methods adapted from species estimators.

- We evaluated the effectiveness of statistical methods to estimate the class size on repeated sampling.



- With the convergence metric we are able to distinguish between complete and incomplete classes in a KG.





https://cardinal.exascale.info/