Scalable Semantic Labeling of Subject Columns in Tabular Data

Ahmad Alobaid, Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Oscar Corcho, Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Axel Ngonga-Ngomo, DICE Group, University of Paderborn,Germany

aalobaid@fi.upm.es
@oeg_upm

5-9-2019
UPM, Montegancedo

How many tabular datasets are there on the web?

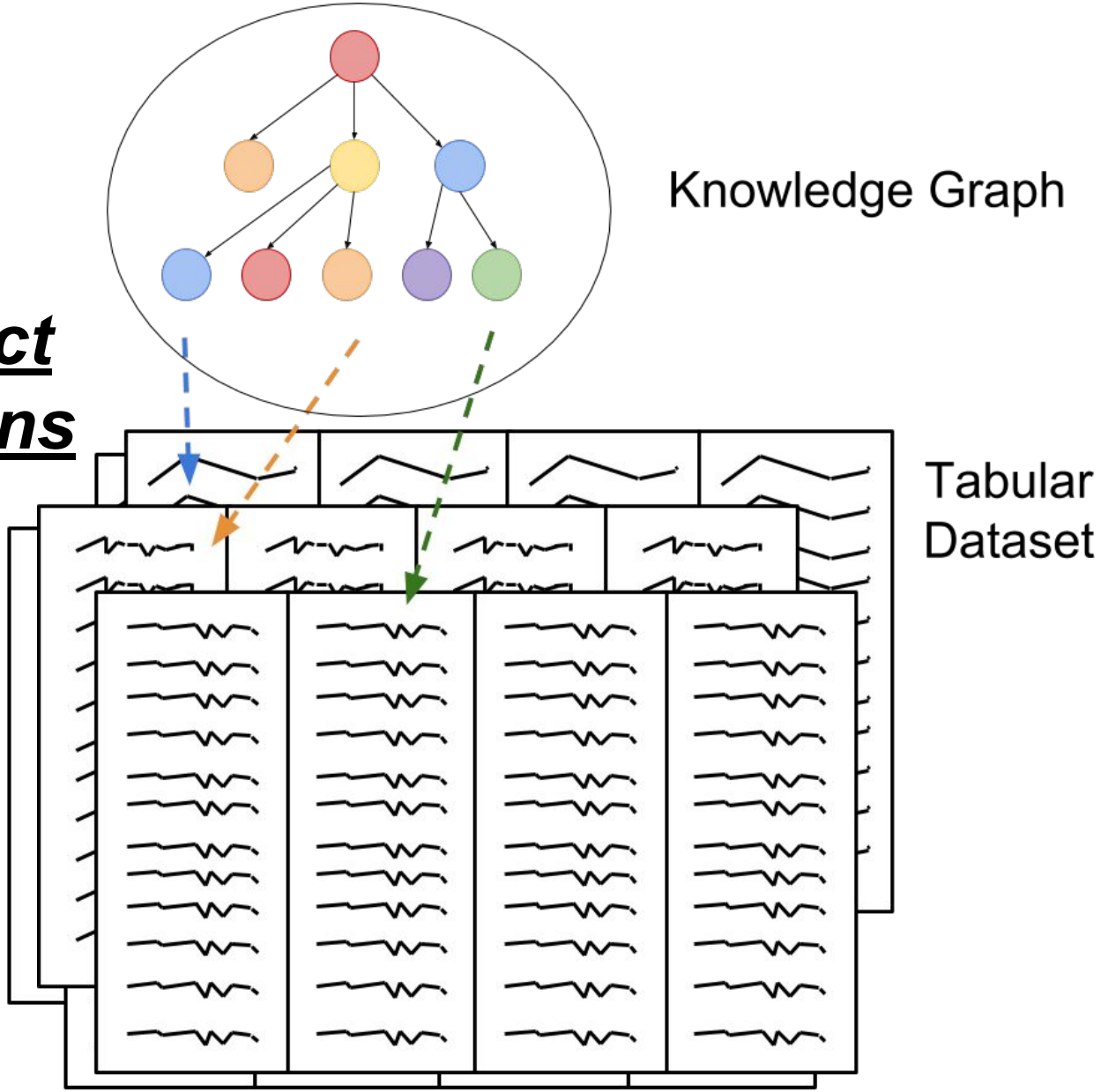
Stacked Webtables

> 150 Million
Web Tables

Burj Khalifa

Cafarella, M.J., Halevy, A., Wang, D.Z., Wu, E., Zhang, Y.: Webtables: exploring the power of tables on the web. Proceedings of the VLDB Endowment 1(1), 538– 549 (2008)

https://ru.pngtree.com/freepng/pile-of-books_1073705.html
https://tall.life/height-comparison-tool-celebrity-height-difference/


Knowledge Graph
Subject Columns
Tabular Dataset

سال	نام	جایزه
۱۹۰۴	خوزه اچه‌خارای	ادبیات
۱۹۰۶	سانتیاگو رامون کاخال	پزشکی
۱۹۲۲	خاسینتو بناونته	ادبیات
۱۹۵۶	خوآن رامون خیمنس	ادبیات
۱۹۵۹	سورو اوچوآ	پزشکی
۱۹۷۷	ویسنته آله‌ایخاندره	ادبیات
۱۹۸۹	کامیلو خوزه سلا	ادبیات

السنة	الصورة	الفائز	المجال
1904		خوسيه إتشيغاراي	الأدب
1906		سانتياغو رامون إي كاخال	الطب
1922		خاثينتو بيناڤينتي	الأدب
1956		خوان رامون خيمينيث	الأدب
1959		سيفيرو أوتشوا	الطب
1977		بيثنتي أليكساندري	الأدب
1989		كاميلو خوسيه ثيلا	الأدب

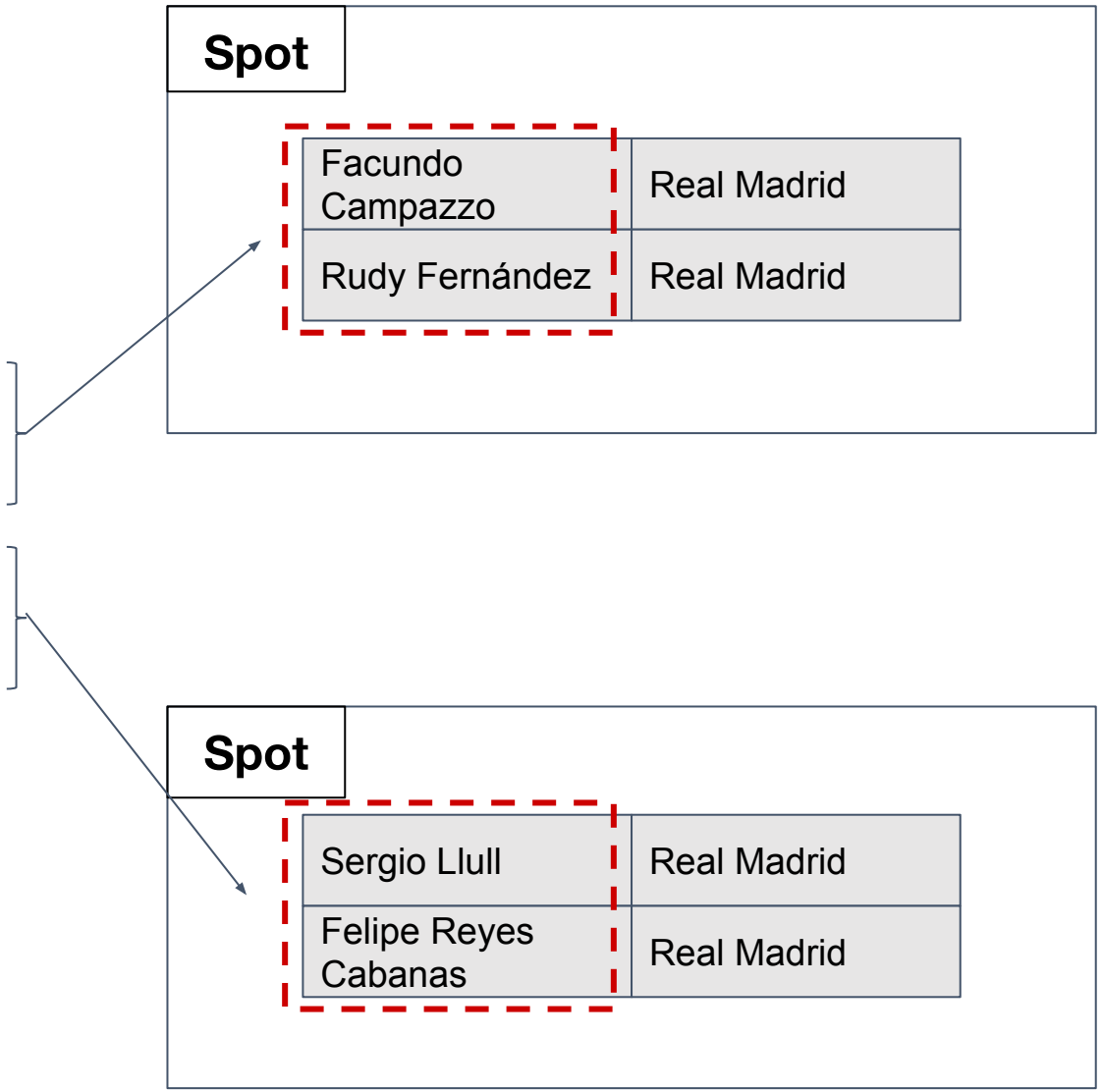
Subject column

Year	Winner	Field	Contribution
1904	José Echegaray	Literature	"in recognition of the numerous and brilliant compositions which, in an individual and original manner, have revived the great traditions of the Spanish drama"
1906	Santiago Ramón y Cajal	Medicine	"in recognition of their work on the structure of the nervous system"
1922	Jacinto Benavente	Literature	"for the happy manner in which he has continued the illustrious traditions of the Spanish drama"
1956	Juan Ramón Jiménez	Literature	"for his lyrical poetry, which in Spanish language constitutes an example of high spirit and artistical purity"
1959	Severo Ochoa	Medicine	"for their discovery of the mechanisms in the biological synthesis of ribonucleic acid and deoxyribonucleic acid"
1977	Vicente Aleixandre	Literature	"for a creative poetic writing which illuminates man's condition in the cosmos and in present-day society, at the same time representing the great renewal of the traditions of Spanish poetry between the wars"
1989	Camilo José Cela	Literature	"for a rich and intensive prose, which with restrained compassion forms a challenging vision of man's vulnerability"
2010	Mario Vargas Llosa	Literature	"for his cartography of structures of power and his trenchant images of the individual's resistance, revolt, and defeat"

Spot methods:

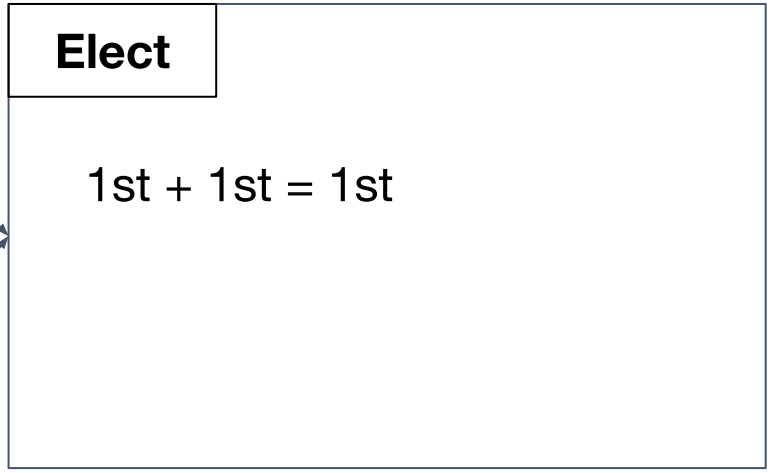
1. Left most
2. Non-numeric left most

Facundo Campazzo	Real Madrid
Rudy Fernández	Real Madrid
Sergio Llull	Real Madrid
Felipe Reyes Cabanas	Real Madrid

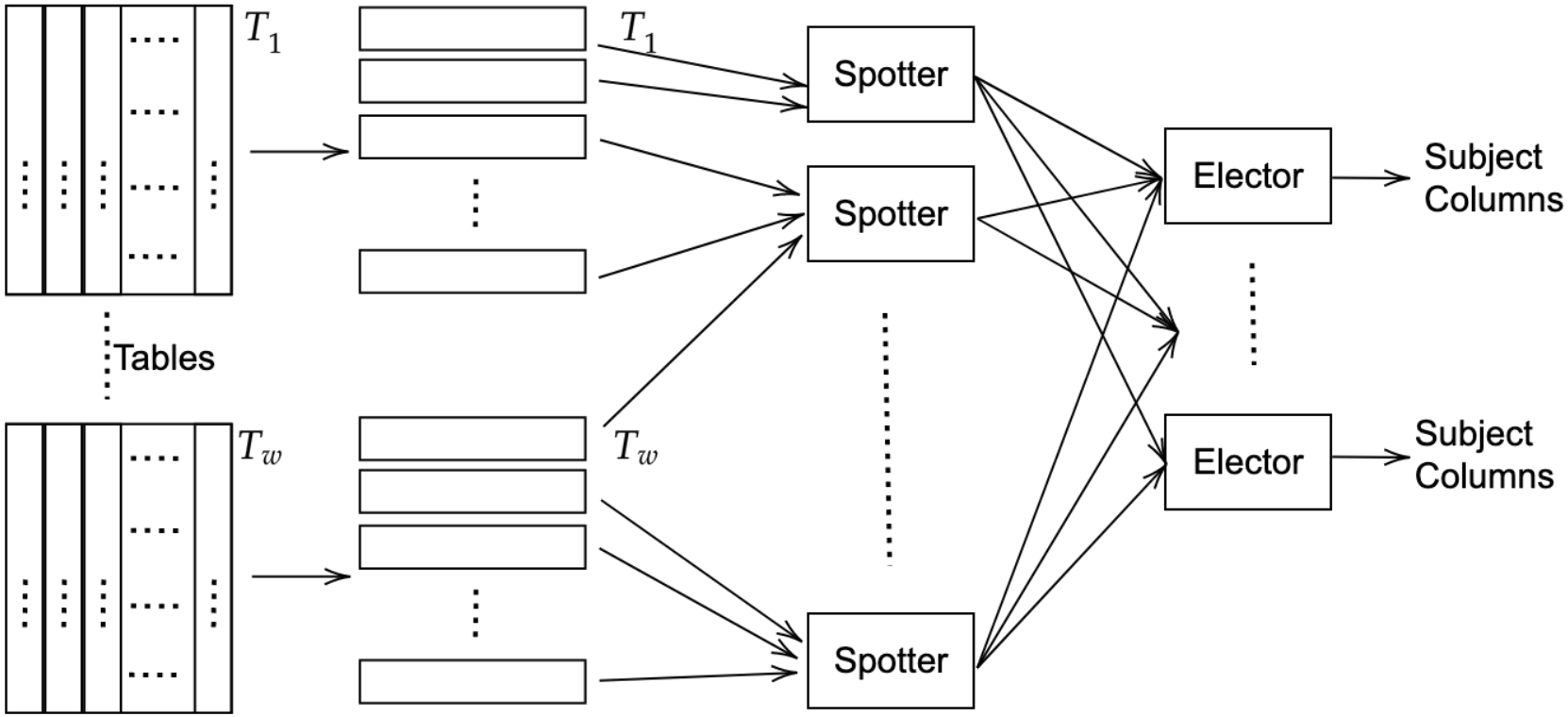


Elect methods:

1. Majority
2. Majority of found



Identifying Subject Columns



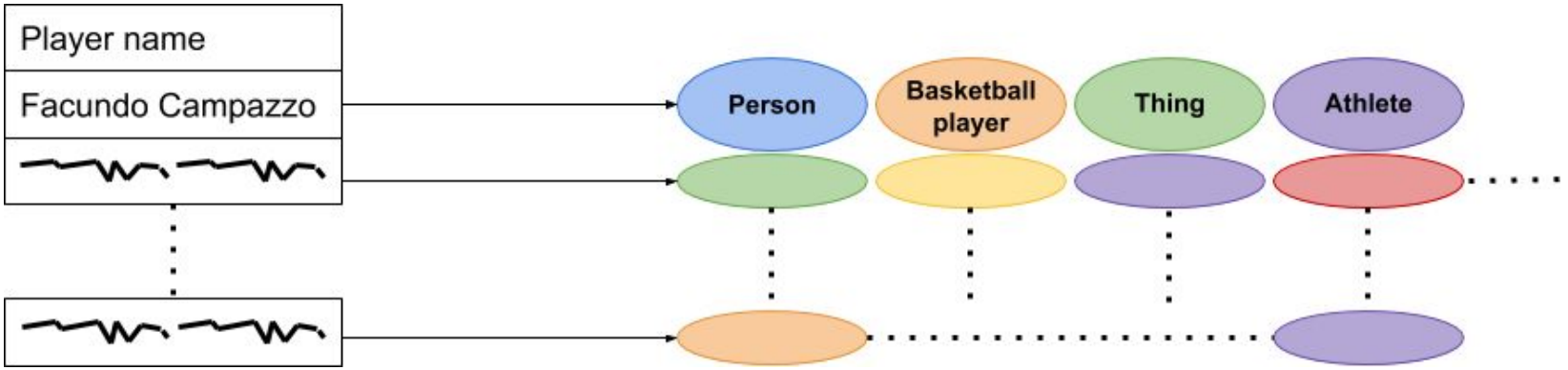
How? link cells to entities?

1. Get entities for each cell

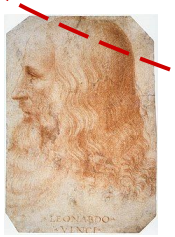
```
select distinct ?subject
where{select distinct ?subject
where{
?subject ?property "Facundo Campazzo"@en}
```

2. Get types for each entity

```
select distinct ?class where{
<http://dbpedia.org/resource/Facundo_Campazzo> a ?class}
```



Example

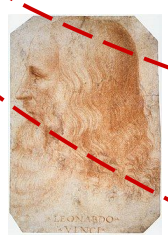


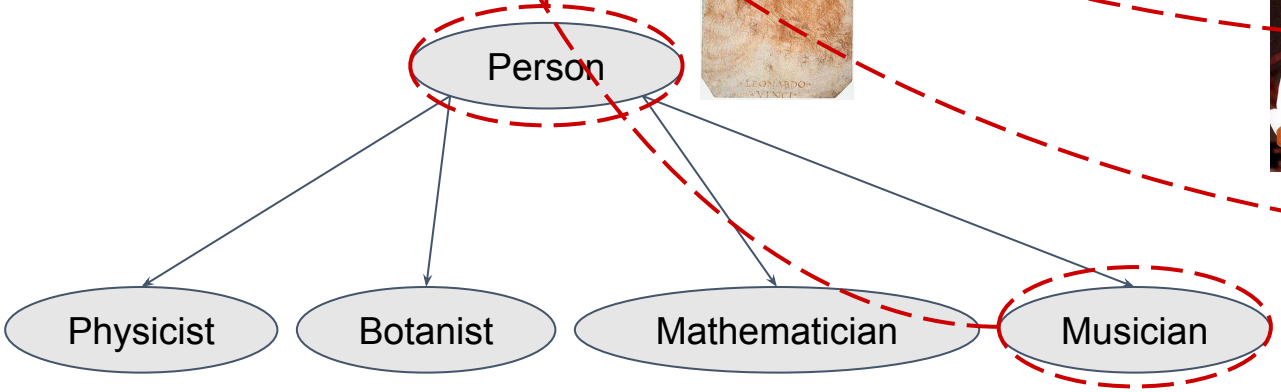
Person
Brian May
Leonardo da Vinci
Pharrell Williams
Bill Gates

http://cdn01.cdn.justjared.com/wp-content/uploads/2014/03/williams-shorts/pharrell-williams-wear-shorts-on-oscars-2014-red-carpet-03.jpg
https://pbs.twimg.com/profile_images/988775660163252226/XpgonN0X.jpg
https://www.queenie.cz/storage/temp/8f2b24db44d09bd8e3d563d0bb099fc2-400x800x1.jpg
https://www.thetimes.co.uk/imageserver/image/methode%2Ftimes%2Fprod%2Fweb%2Fbin%2F3f44abc8-2774-11e8-acc5-262aff1ca7a6.jpg?crop=988%2C556%2C910%2C214&resize=685
https://i1.wp.com/blog.eil.com/wp-content/uploads/2018/12/brian-may-astrophysicist-07212016.jpg?fit=970%2C545&ssl=1

Example

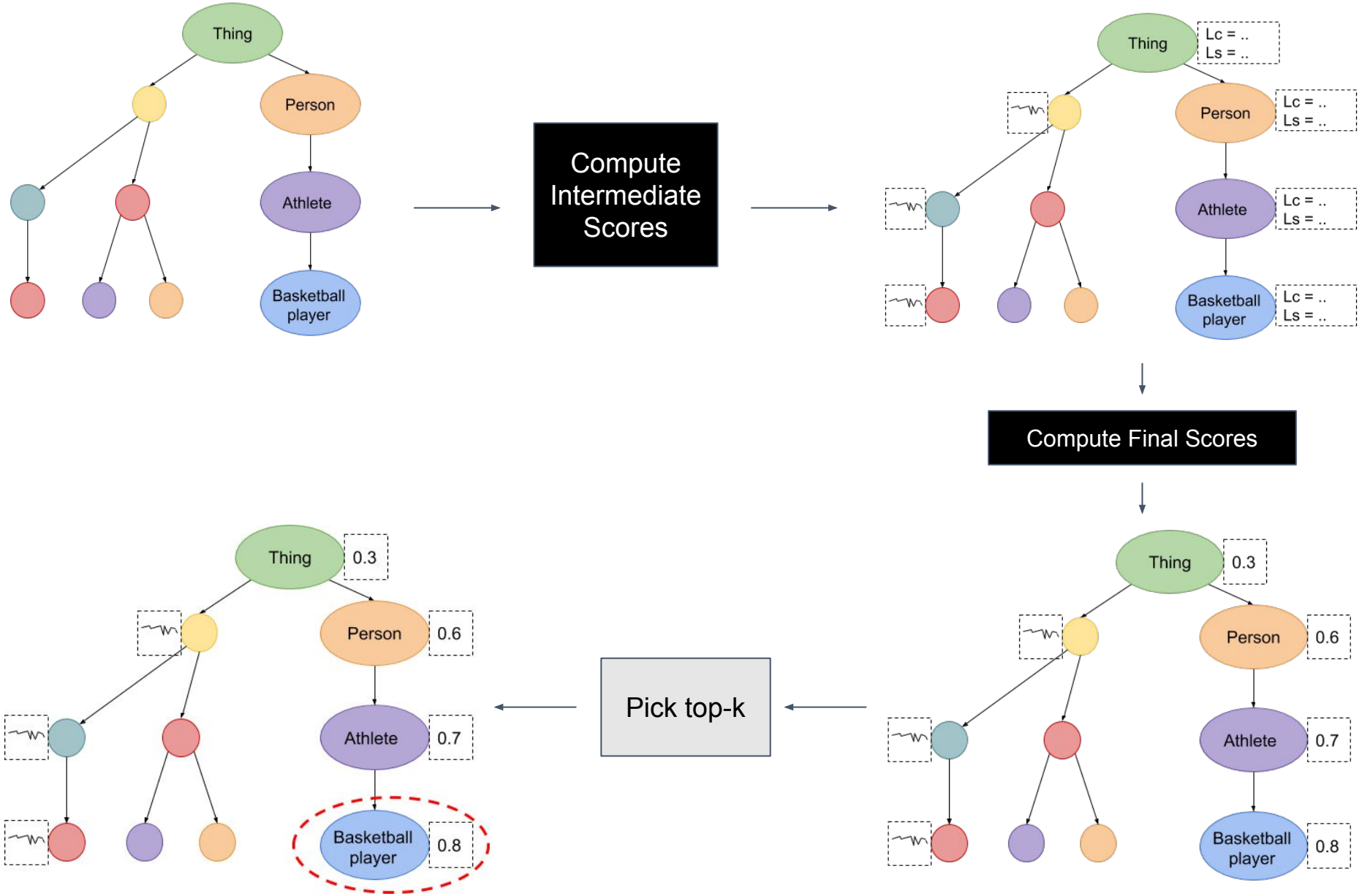
Person
Brian May
Leonardo da Vinci
Pharrell Williams



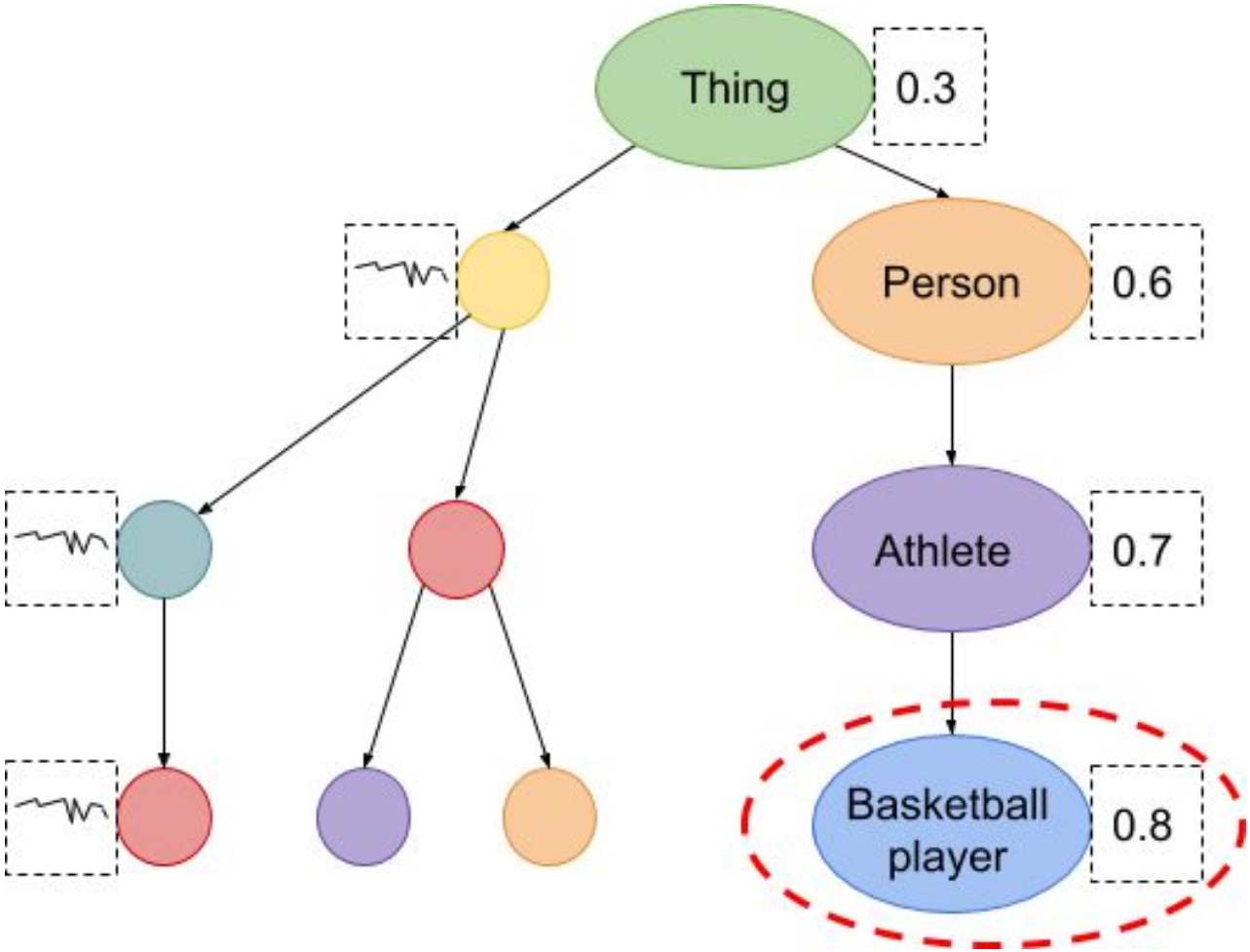


http://cdn01.cdn.justjared.com/wp-content/uploads/2014/03/williams-shorts/pharrell-williams-wear-shorts-on-oscars-2014-red-carpet-03.jpg
https://pbs.twimg.com/profile_images/988775660163252226/XpgonN0X.jpg
https://www.queenie.cz/storage/temp/8f2b24db44d09bd8e3d563d0bb099fc2-400x800x1.jpg
https://www.thetimes.co.uk/imageserver/image/methode%2Ftimes%2Fprod%2Fweb%2Fbin%2F3f44abc8-2774-11e8-acc5-262aff1ca7a6.jpg?crop=988%2C556%2C910%2C214&resize=685
https://i1.wp.com/blog.eil.com/wp-content/uploads/2018/12/brian-may-astrophysicist-07212016.jpg?fit=970%2C545&ssl=1

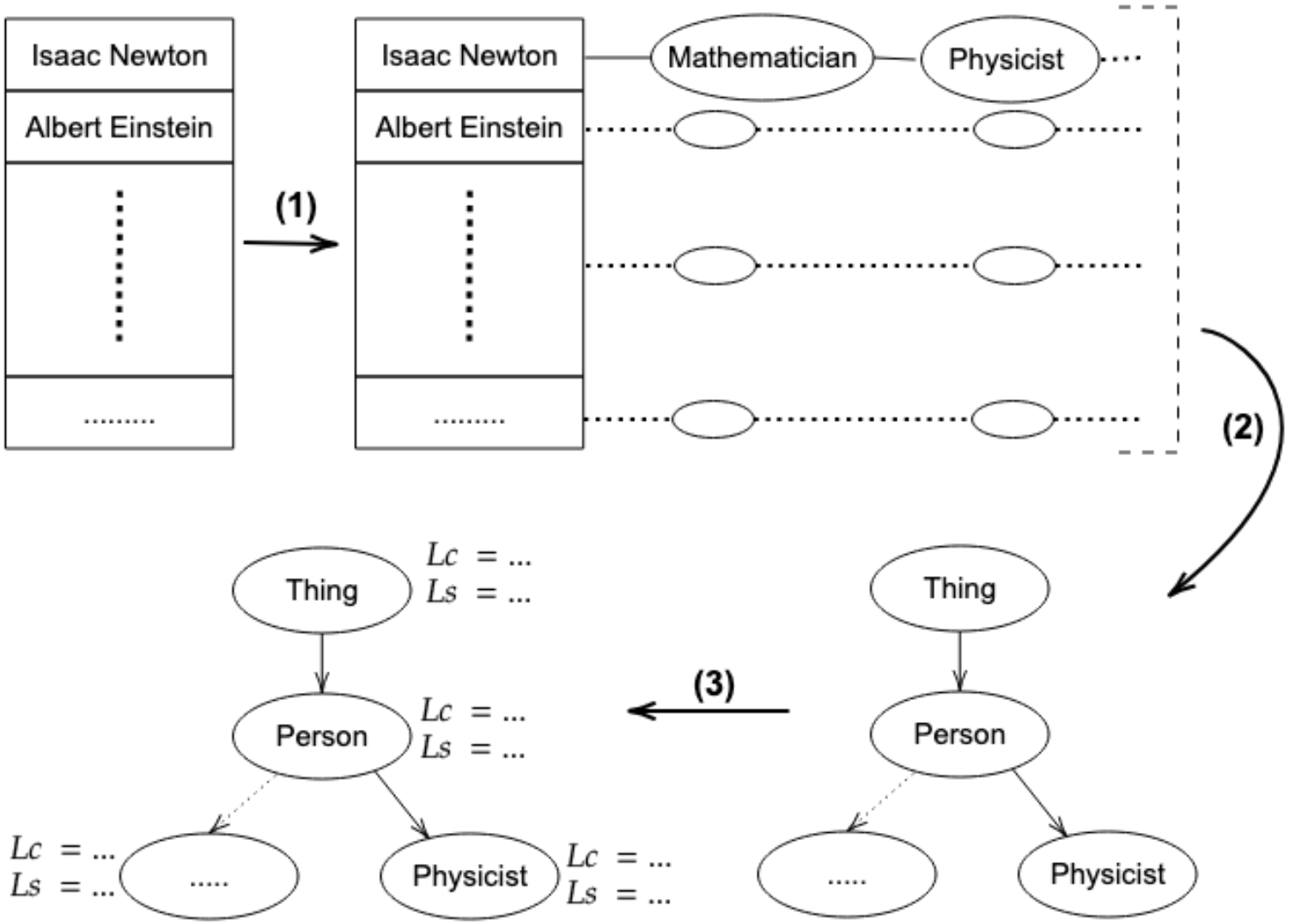
Singleton Workflow



Scored graph

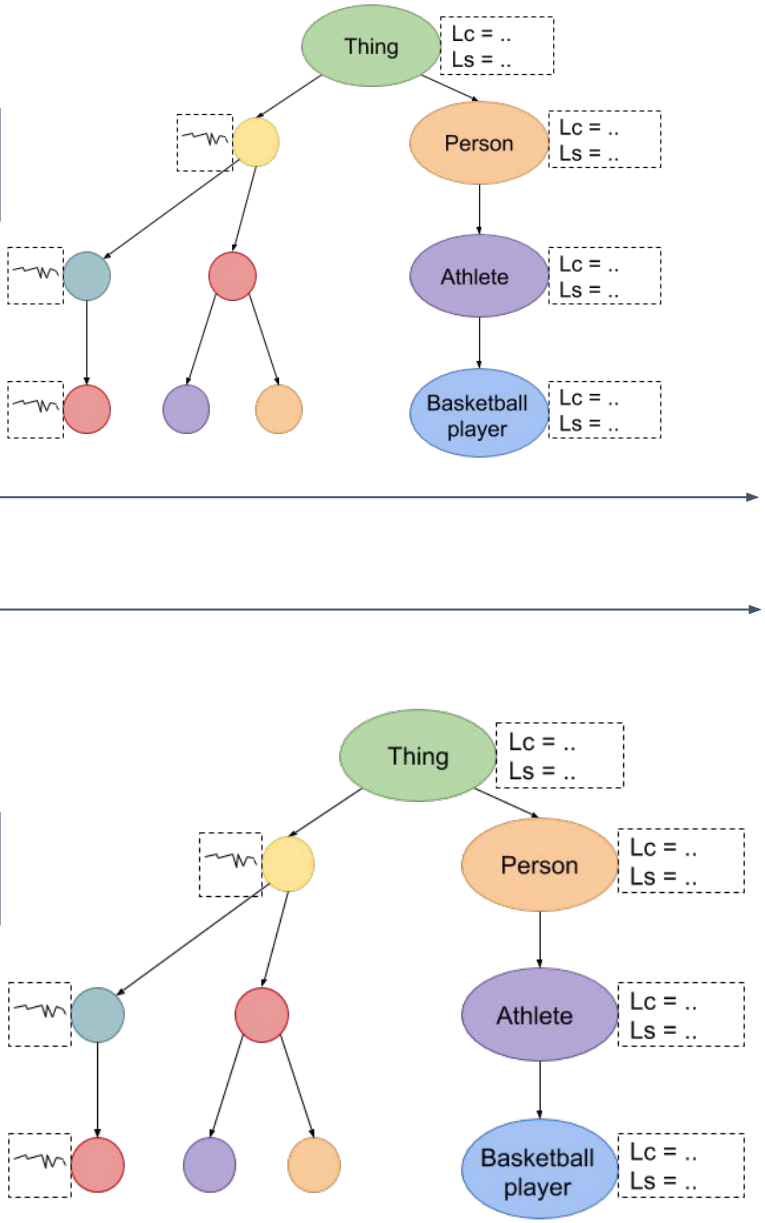


Workflow - singleton

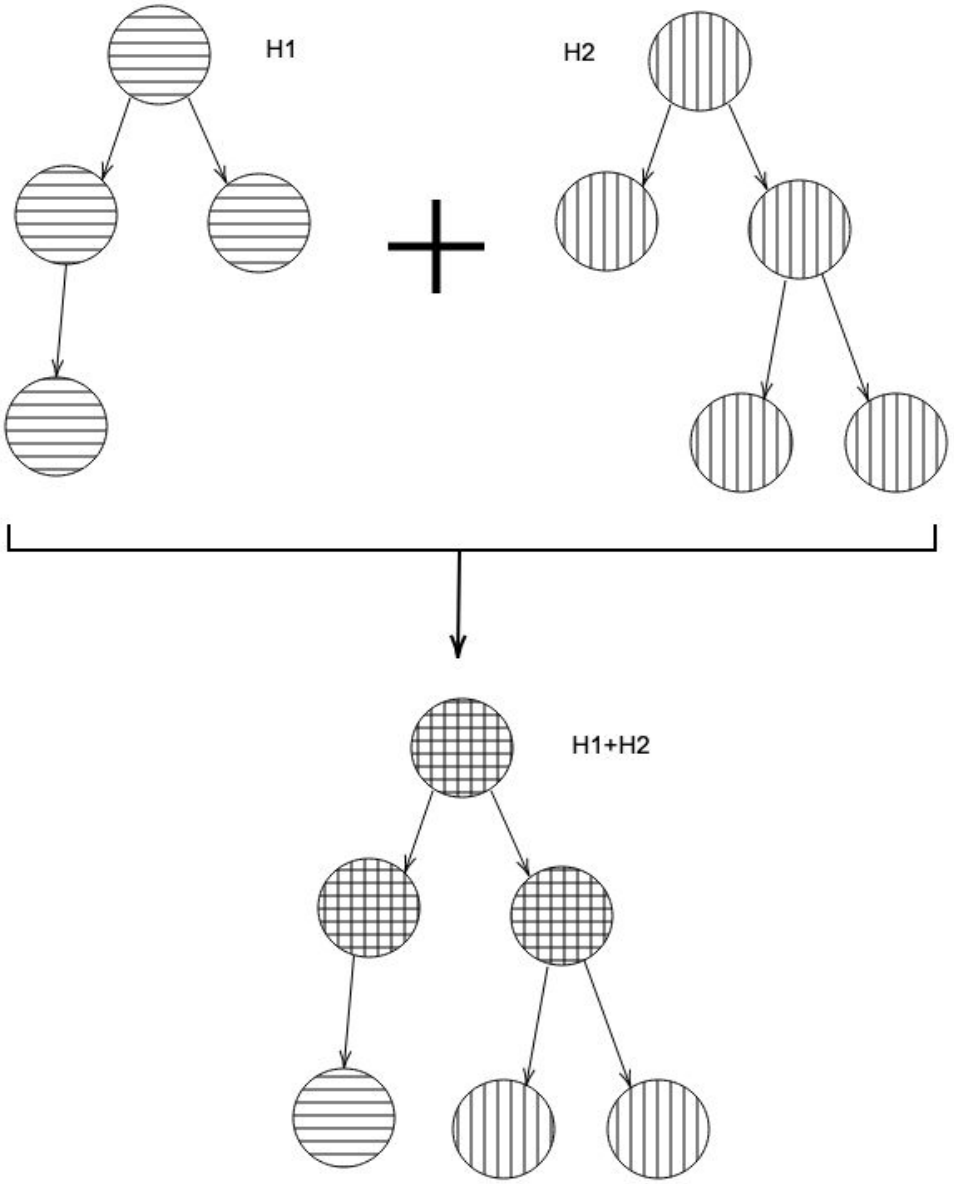


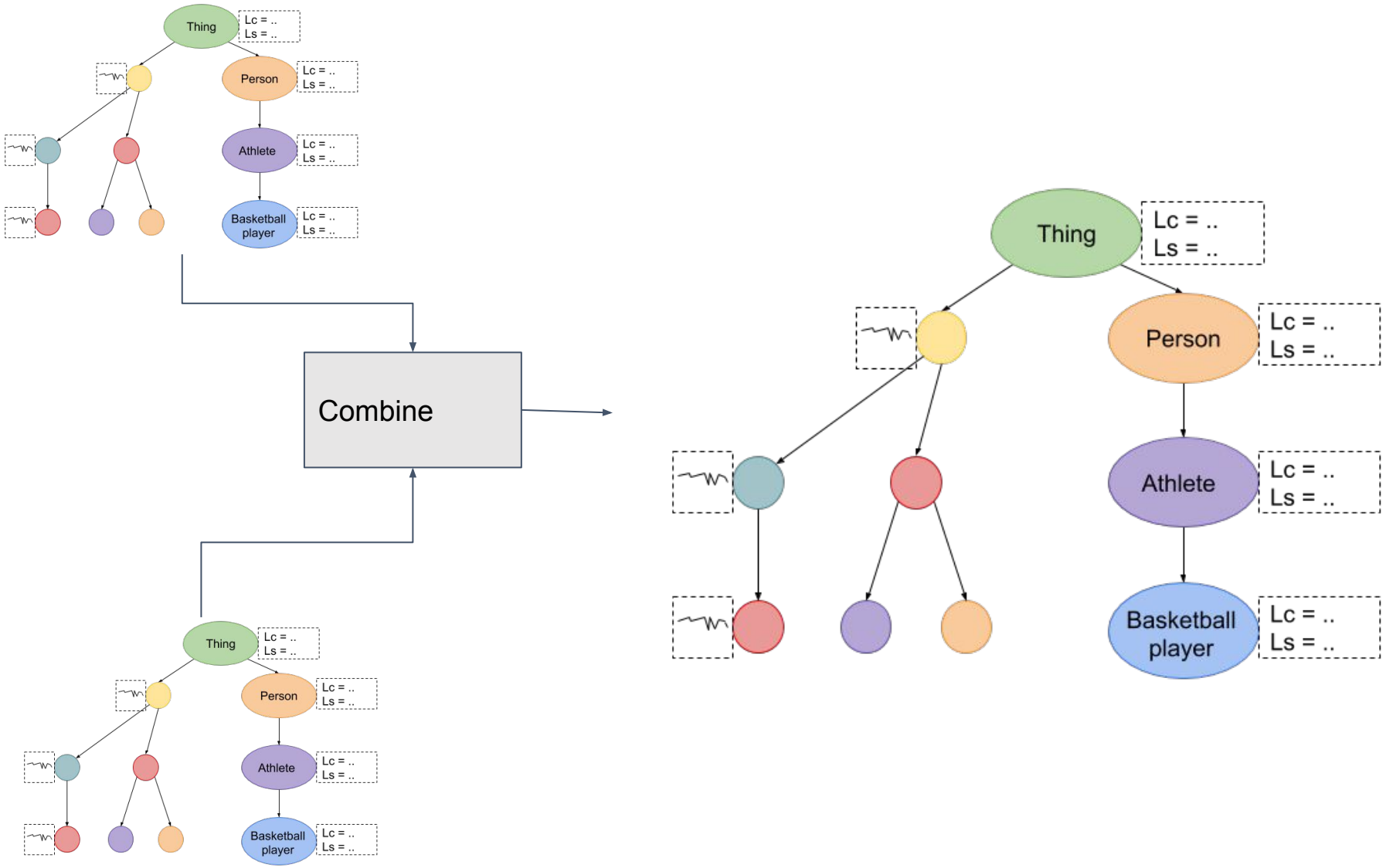
Can we make it scalable without losing accuracy?

Brian May
Leonardo da Vinci
Pharrell Williams
Bill Gates

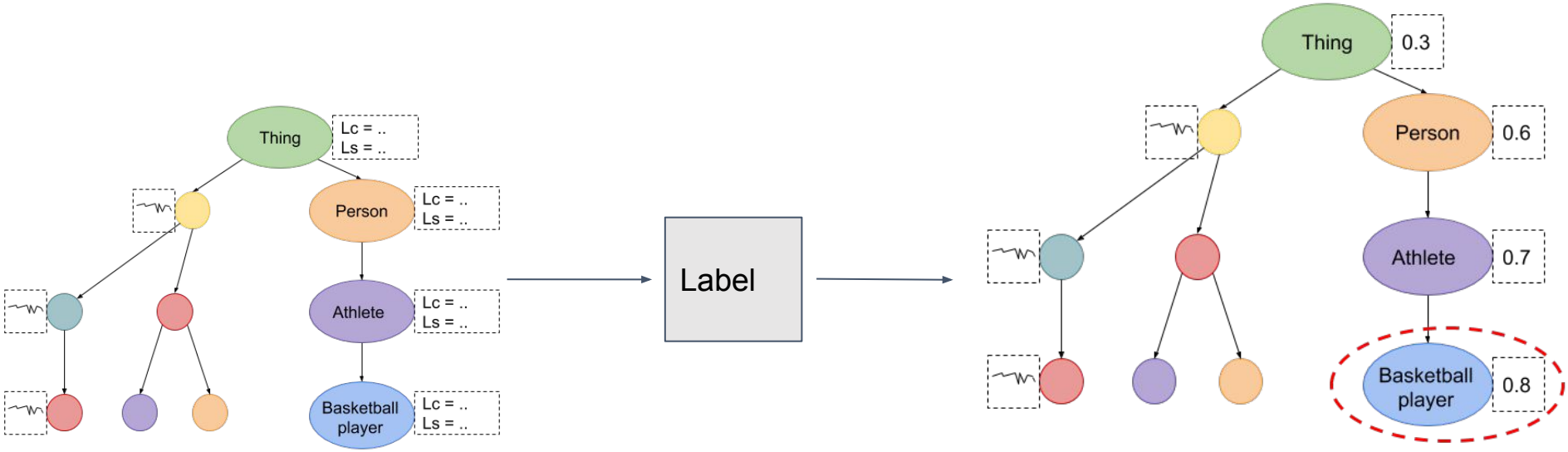


Combine

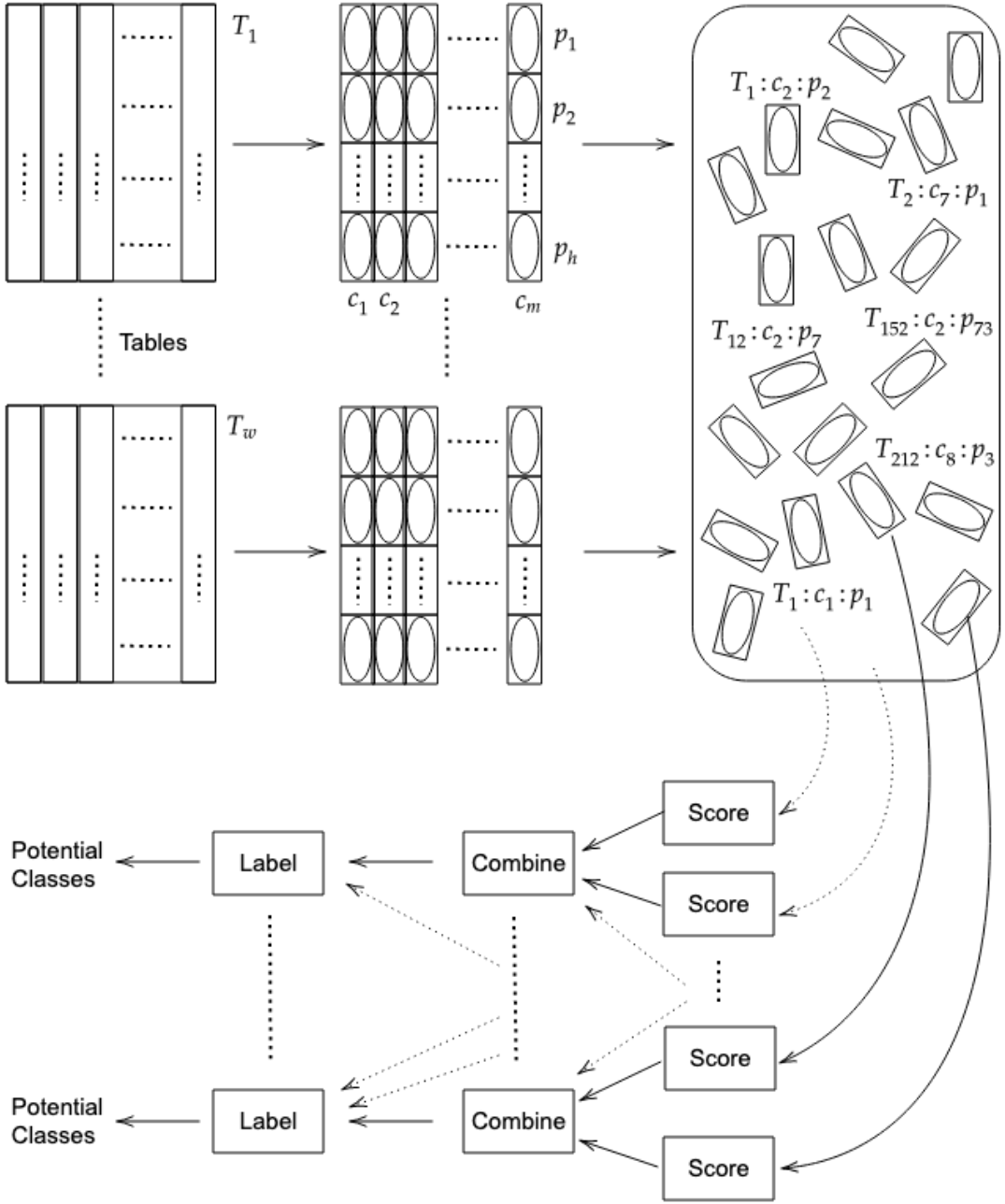



Thing
Lc = ..
Ls = ..
Person
Lc = ..
Ls = ..
Athlete
Lc = ..
Ls = ..
Basketball player
Lc = ..
Ls = ..
Combine
Thing
Lc = ..
Ls = ..
Person
Lc = ..
Ls = ..
Athlete
Lc = ..
Ls = ..
Basketball player
Lc = ..
Ls = ..
Thing
Lc = ..
Ls = ..
Person
Lc = ..
Ls = ..
Athlete
Lc = ..
Ls = ..
Basketball player
Lc = ..
Ls = ..

Score


Thing
Lc = ..
Ls = ..
Person
Lc = ..
Ls = ..
Athlete
Lc = ..
Ls = ..
Basketball player
Lc = ..
Ls = ..
Label
Thing
0.3
Person
0.6
Athlete
0.7
Basketball player
0.8

Workflow - distributed

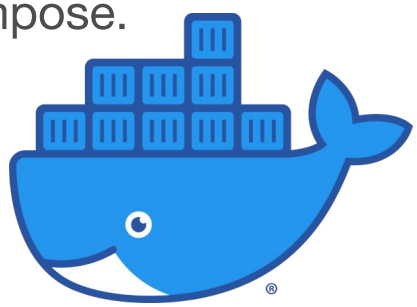


Setup:

Instances are docker containers created with docker-compose.

Machine: 8GB memory, 2.8 GHz Intel Core i7

Operating system: MacOS Mojave



Subject Column Identification:

- Spot: 6 instances
- Elect: 3 instances

Data:

- T2Dv2: Web Data Commons (annotated)
- T2DT: Web Data Commons (subset annotated by TAIPAN)

Runtime:

- ~15 minutes each run

T2Dv2:

Spotter	Elector	Precision	Recall	F1
left most	majority	0.54	1.0	0.7
left most	majority-of-found	0.54	1.0	0.7
non-numerical	majority	0.86	0.98	0.91
non-numerical	majority-of-found	0.86	1.0	0.92

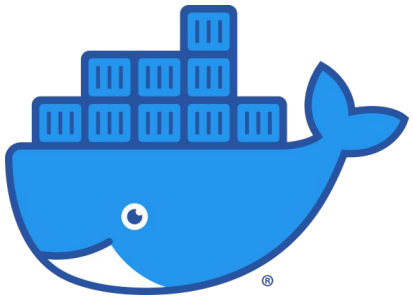
T2DT

Approach	Precision	Recall	F1
left most + majority	0.58	1.00	0.73
left most + majority-of-found	0.58	1.00	0.73
non-numerical + majority	0.85	0.92	0.88
non-numerical + majority-of-found	0.85	0.99	0.92
TAIPAN (Rule-based)	0.51	-	0.68*
TAIPAN (Support)	0.54	-	0.7*
TAIPAN (Connectivity)	0.36	-	0.53*
TAIPAN (Support+Connectivity)	0.56	-	0.72*

* F1 measure is computed with the assumption that recall=1.0

Subject Column Labeling:

- Score: 12 instances
- Combine: 4 instances



Data:

- T2Dv2: Web Data Commons

Runtime:

- ~28 hours

Approach	Precision	Recall	F1
T2K (Majority)	0.47	0.51	0.49
T2K (Majority + Frequency)	0.87	0.90	0.89
T2K (Page attributes)	0.97	0.37	0.53
T2K (Text)	0.75	0.34	0.46
T2K (Majority + Frequency + Page attributes + Text)	0.90	0.86	0.88
T2K Extended	0.93	0.91	**0.92**
Our approach	0.91	0.97	**0.94**

Conclusion and Future work

Conclusion: We were able to perform semantic labeling on a larger scale without any loss in precision or recall.

Future work:

- Experiment with sampling to label only a subset of the cells in subject columns.
- Perform semantic labeling on other columns as well.

Questions?