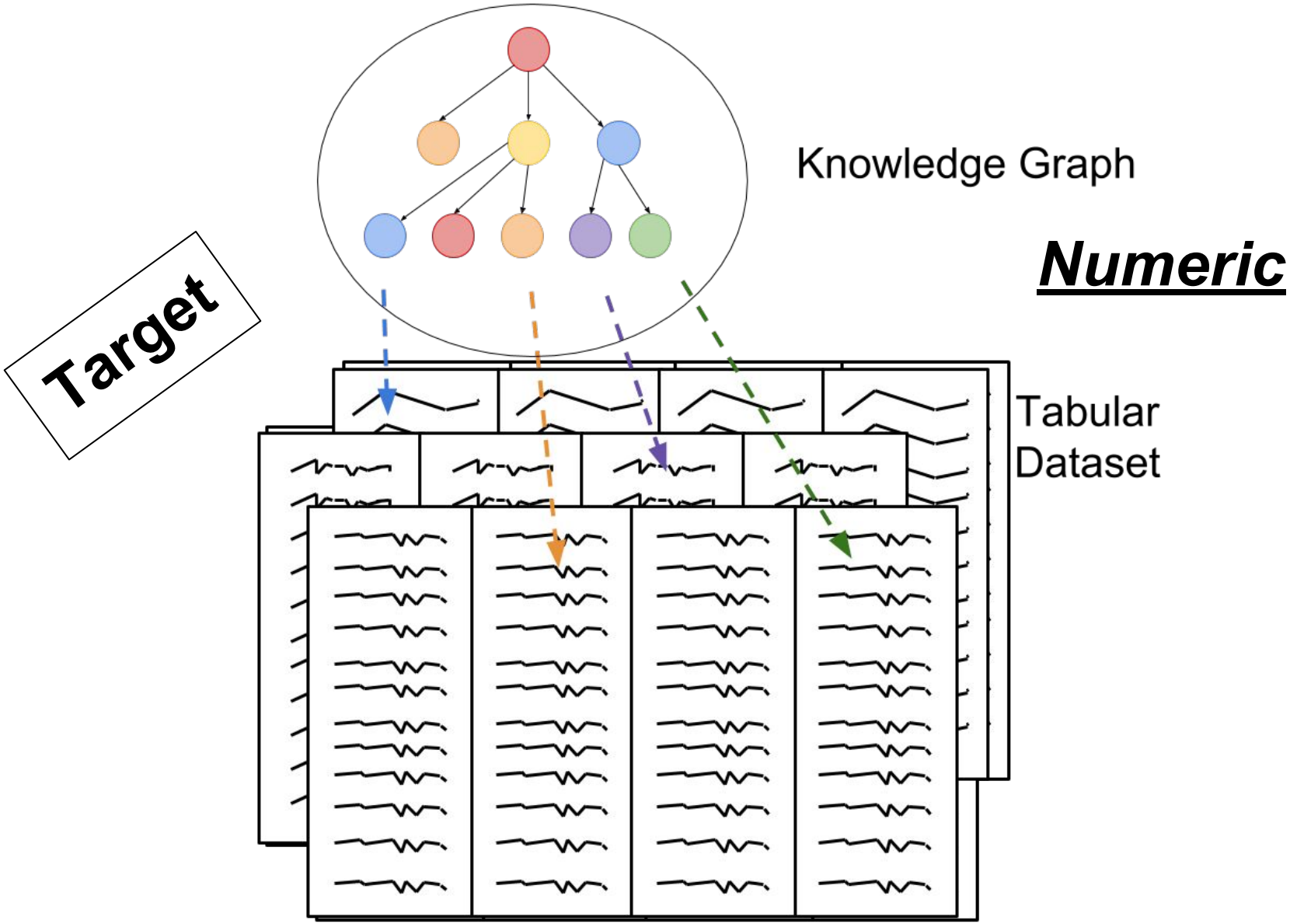
Typology-based Semantic Labeling of Numeric Tabular Data

Ahmad Alobaid, Emilia Kacprzak and Oscar Corcho
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

aalobaid@fi.upm.es
ahmad88csc

4-4-2019
1003, block 1, Montegancedo

CC BY SA


Knowledge Graph
Numeric
Target
Tabular
Dataset

Different Types of Numerical Data are treated the same

Mesopotamian

Egypt

Roman

Greek

China

0123456789

٠١٢٣٤٥٦٧٨٩

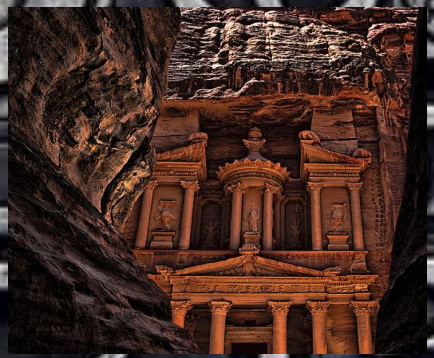
I II III IV V VI VII VIII IX X

০১২৩৪৫৬৭৮৯

๐๑๒๓๔๕๖๗๘๙

০১২৩৪৫৬৭৮৯

〇一二三四五六七八九



Arabs

Brahmagupta

https://www.britannica.com/art/Egyptian-art/images-videos/media/180644/94142
https://en.wikipedia.org/wiki/History_of_the_Arabs
https://learnfunfacts.com/2017/12/06/curious-number-patterns-5-9th-power-pattern-munchhausen-numbers-1729/
https://en.wikipedia.org/wiki/Ancient_Mesopotamian_units_of_measurement
https://www.storyofmathematics.com/indian_brahmagupta.html
https://en.wikipedia.org/wiki/Roman_Empire

Levels of Measurement

https://en.wikipedia.org/wiki/Cubit

Measurement error

Typology of Numbers

Stanley Smith Stevens

Nominal

Ordinal

Interval

http://braintour.harvard.edu/wp-content/uploads/2016/05/tile_TT_SmittyStevens.jpg
https://www.istockphoto.com/es/fotos/half-full-bottle-full-empty?sort=mostpopular&mediatype=photography&phrase=half%20full%20bottle%20full%20empty
https://en.wikipedia.org/wiki/Conversion_of_units_of_temperature#Comparison_of_temperature_scales
https://en.wikipedia.org/wiki/List_of_2014_Winter_Olympics_medal_winners
https://commons.wikimedia.org/wiki/File:TapeMeasure.png

Example of Interval Type

Comparison of temperature scales

Comment	Kelvin	Celsius	Fahrenheit	Rankine	Delisle	Newton	Réaumur	Rømer
Absolute zero	0.00	-273.15	−459.67	0.00	559.73	−90.14	−218.52	−135.90
Lowest recorded surface temperature on Earth[1]	184	−89.2[1]	−128.6[1]	331	284	−29	−71	−39
Fahrenheit's ice/salt mixture	255.37	−17.78	0.00	459.67	176.67	−5.87	−14.22	−1.83
Ice melts (at standard pressure)	273.15	0.00	32.00	491.67	150.00	0.00	0.00	7.50
Triple point of water	273.16	0.01	32.018	491.688	149.985	0.0033	0.008	7.50525
Average surface temperature on Earth	288	15	59	519	128	5	12	15
Average human body temperature*	310	37	98	558	95	12	29	27
Highest recorded surface temperature on Earth[2]	331	58[2]	136.4[2]	596	63	19	46	38
Water boils (at standard pressure)	373.1339	99.9839	211.97102[3]	671.64102[3]	0.00	33.00	80.00	60.00
Titanium melts	1941	1668	3034	3494	−2352	550	1334	883
The surface of the Sun	5800	5500	9900	10400	−8100	1800	4400	2900

Typology of Numbers

Stanley Smith Stevens

Nominal

Ordinal

Interval

Comment
Absolute zero
Lowest recorded surface temperature on Earth[1]
Fahrenheit's ice/salt mixture
Ice melts (at standard pressure)
Triple point of water
Average surface temperature on Earth
Average human body temperature*
Highest recorded surface temperature on Earth[2]
Water boils (at standard pressure)
Titanium melts
The surface of the Sun

Ratio

http://braintour.harvard.edu/wp-content/uploads/2016/05/tile_TT_SmittyStevens.jpg
https://www.istockphoto.com/es/fotos/half-full-bottle-full-empty?sort=mostpopular&mediatype=photography&phrase=half%20full%20bottle%20full%20empty
https://en.wikipedia.org/wiki/Conversion_of_units_of_temperature#Comparison_of_temperature_scales
https://en.wikipedia.org/wiki/List_of_2014_Winter_Olympics_medal_winners
https://commons.wikimedia.org/wiki/File:TapeMeasure.png

Typology of Numbers

Stanley Smith Stevens

Nominal

Sequential

http://braintour.harvard.edu/wp-content/uploads/2016/05/tile_TT_SmittyStevens.jpg
https://pbs.twimg.com/media/DICH9C9XsAYGaLG.jpg
https://s3.envato.com/files/186945900/008(basket04_4color_apples).jpg
https://www.ft.com/content/64d6dbc0-5275-11e6-9664-e0bdc13c3bef
https://www.guru99.com/images/MongoDB/112115_0607_Introductio11.png
https://upload.wikimedia.org/wikipedia/commons/0/03/PD_social_security_card.png

STARFLEET PERSONNEL FILE: SATO, HOSHI
SERIAL NUMBER: SA-037-0198-CL

Rank at retirement: Lieutenant Commander
Former Assignment: Communications and Protocol officer, Enterprise NX-01
Birthplace: Kyoto, Japan, Earth

Hoshi Sato served as translator, and protocol and communications officer on Starfleet's first warp five starship, Enterprise NX-01. Born in Kyoto, Japan on July 9th, 2129, she was the second child in a family of three. After leaving Starfleet in her late thirties, Sato created the linguacode translation matrix, which is still in use aboard Federation starships today.

PSYCHOLOGICAL PROFILE

Hoshi was a spirited, intelligent woman with an extraordinary gift for alien languages who also served as translator aboard the Starship Enterprise NX-01 — a vital role when making first contact. A "white-knuckle" space-farer, Hoshi reluctantly gave up her teaching job after being convinced by Captain Jonathan Archer to join Starfleet.

BIOGRAPHICAL OVERVIEW

Hoshi has also formed bonds with several of her crewmates, particularly Dr. Phlox, who she says has taken care of her on many occasions. Phlox is currently teaching her Denobulan — according to his wife Feezal, Hoshi's accent is very good. When Phlox was infected by the mysterious nanoprobes from hostile cybernetic beings, Hoshi offered to keep him company while he worked on a cure.

As Hoshi continued her tenure aboard Enterprise, she almost mirrored humankind in taking the initial steps into the intergalactic community: tentative and concerned at first, but more and more sure of herself as time goes on. While ready to take whatever action is necessary to help the crew and Starfleet, her accomplishments in communication also provide an example of how tense situations can be diffused through diplomatic means.

Tragically, Hoshi and her family were among the four thousand people who died on Tarsus Four in 2246 when a food shortage caused by an exotic fungus threatened the colony's population. Governor Kodos ordered the deaths of Sato and the others in order to save the rest of the colony. She was buried in Kyoto with her husband, Takashi Kimura.

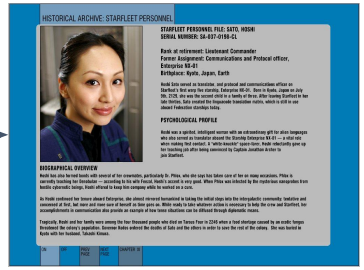
Typology of Numbers

Stanley Smith Stevens

Nominal

Sequential

Hierarchical



http://braintour.harvard.edu/wp-content/uploads/2016/05/tile_TT_SmittyStevens.jpg
https://pbs.twimg.com/media/DICH9C9XsAYGaLG.jpg
https://s3.envato.com/files/186945900/008(basket04_4color_apples).jpg
https://www.ft.com/content/64d6dbc0-5275-11e6-9664-e0bdc13c3bef
https://www.guru99.com/images/MongoDB/112115_0607_Introductio11.png
https://upload.wikimedia.org/wikipedia/commons/0/03/PD_social_security_card.png

SOCIAL SECURITY ACT

ACCOUNT NUMBER

721-07-4426

HAS BEEN ESTABLISHED FOR

Clinton Lester Rocker

4/27/43 Clinton Lester Rocker

DATE OF ISSUE EMPLOYEE'S SIGNATURE

Keep this card. It shows the account number used in keeping records of your Social Security Benefit rights under Federal and State Laws. Keep a record of this number as you might lose the card. Mention the number in all letters regarding your account.

Address inquiries concerning Unemployment Compensation (if there is a law in your State) to the State agency administering such law. Address inquiries concerning Federal Old-Age Retirement Benefits (not State Old-Age Assistance or Pensions) to the nearest office of the Social Security Board.

SIGN THIS CARD IMMEDIATELY AND REPORT THE NUMBER TO YOUR EMPLOYER.

Typology of Numbers

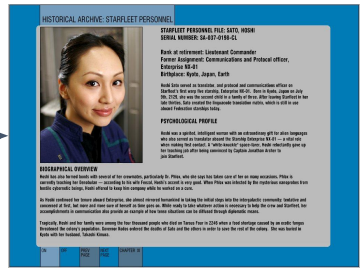
Stanley Smith Stevens

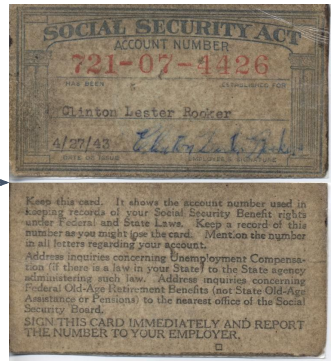
Nominal

Sequential

Hierarchical

Categorical





http://braintour.harvard.edu/wp-content/uploads/2016/05/tile_TT_SmittyStevens.jpg
https://pbs.twimg.com/media/DICH9C9XsAYGaLG.jpg
https://s3.envato.com/files/186945900/008(basket04_4color_apples).jpg
https://www.ft.com/content/64d6dbc0-5275-11e6-9664-e0bdc13c3bef
https://www.guru99.com/images/MongoDB/112115_0607_Introductio11.png
https://upload.wikimedia.org/wikipedia/commons/0/03/PD_social_security_card.png

	A	B	C	D	E	F	G	H	I	J
3	Age	Party	Gender	Income		Age	Party 1	Party 2	Gender 1	Income
4	20	Rep	Male	45000		20	1	0	1	45000
5	25	Dem	Male	39000		25	0	1	1	39000
6	45	Ind	Male	56000		45	0	0	1	56000
7	35	Rep	Female	49000		35	1	0	0	49000
8	50	Dem	Female	41000		50	0	1	0	41000
9	55	Ind	Female	42000		55	0	0	0	42000
10	39	Rep	Male	58000		39	1	0	1	58000
11	48	Dem	Male	55000		48	0	1	1	55000
12	30	Ind	Male	46000		30	0	0	1	46000
13	27	Rep	Female	42000		27	1	0	0	42000
14	47	Dem	Female	37000		47	0	1	0	37000
15	21	Ind	Female	25000		21	0	0	0	25000
16	48	Rep	Male	75000		48	1	0	1	75000
17	24	Ind	Male	43000		24	0	0	1	43000
18	28	Ind	Female	40000		28	0	0	0	40000
19	40	Dem	Female	31000		40	0	1	0	31000

Typology of Numbers

Stanley Smith Stevens

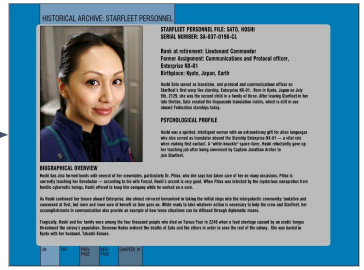
Nominal

Sequential

Hierarchical

Categorical

Random



HISTORICAL ARCHIVE: STARFLEET PERSONNEL

STARFLEET PERSONNEL FILE: SATO, HOSHI
SERIAL NUMBER: SA-027-0198-CL

Rank at retirement: Lieutenant Commander
Former Assignment: Communications and Protocol officer, Enterprise NX-01
Birthplace: Kyoto, Japan, Earth

Hoshi Sato served as translator, and protocol and communications officer on Enterprise NX-01 from the ship's first mission, through the Xindi War. Born in Kyoto on July 29, 2139, she was the second child in a family of three. After leaving Stanford in New Lake Shrine, Sato created the linguasonic translation matrix, which is still in use aboard Federation starships today.

PSYCHOLOGICAL PROFILE

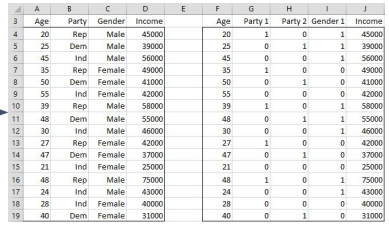
Hoshi was a spirited, intelligent woman with an extraordinary gift for alien languages who also served as translator aboard the Starship Enterprise NX-01 — a ship's ear, when making first contact. A 'whip-smarter' space-farer, Hoshi reluctantly gave up her teaching job after being convinced by Captain Jonathan Archer to join Starfleet.

BIOGRAPHICAL OVERVIEW

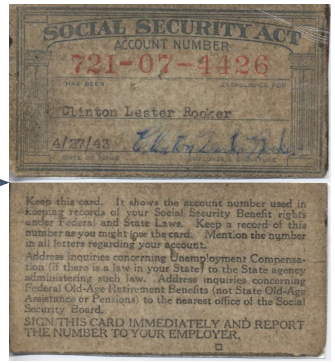
Hoshi has also learned bonds with several of her crewmates, particularly Dr. Phlox, who she says has taken care of her on many occasions. Phlox is currently teaching her Denobulan — according to his wife Feezal, Hoshi's accent is very good. When Phlox was infected by the mysterious xenoparasites from Hoshi's cybernetic implant, Hoshi offered to keep him company while he worked on a cure.

As Hoshi continued her tenure aboard Enterprise, she almost mirrored herself in taking the initial steps into the interpalatic community. Imitative and concerned at first, but more and more sure of herself as time goes on. Hoshi made to take whatever action is necessary to help the crew and Starfleet, her accomplishments in communication also provide an example of how tense situations can be diffused through diplomatic means.

Tragically, Hoshi and her family were among the four thousand people who died on Tarsus Four in 2245 when a food shortage caused by an erratic Tarsian threatened the colony's population. Governor Kodos ordered the deaths of Gaila and the others in order to save the rest of the colony. She was buried in Kyoto with her husband, Takashi Kimura.



	A	B	C	D	E	F	G	H	I	J
3	Age	Party	Gender	Income		Age	Party 1	Party 2	Gender 1	Income
4	20	Rep	Male	45000		20	1	0	1	45000
5	25	Dem	Male	39000		25	0	1	1	39000
6	45	Ind	Male	56000		45	0	0	1	56000
7	35	Rep	Female	49000		35	1	0	0	49000
8	50	Dem	Female	41000		50	0	1	0	41000
9	55	Ind	Female	42000		55	0	0	0	42000
10	39	Rep	Male	58000		39	1	0	1	58000
11	48	Dem	Male	55000		48	0	1	1	55000
12	30	Ind	Male	46000		30	0	0	1	46000
13	27	Rep	Female	42000		27	1	0	0	42000
14	47	Dem	Female	37000		47	0	1	0	37000
15	21	Ind	Female	25000		21	0	0	0	25000
16	48	Rep	Male	75000		48	1	0	1	75000
17	24	Ind	Male	43000		24	0	0	1	43000
18	28	Ind	Female	40000		28	0	0	0	40000
19	40	Dem	Female	31000		40	0	1	0	31000



SOCIAL SECURITY ACT
ACCOUNT NUMBER
721-07-4426
HAS BEEN ASSIGNED FOR
Clinton Lester Rooker
4/27/43
DATE OF ISSUE

Keep this card. It shows the account number used in keeping records of your Social Security Benefit rights under Federal and State Laws. Keep a record of this number as you might lose the card. Mention the number in all letters regarding your account.
Address inquiries concerning Unemployment Compensation (if there is a law in your State) to the State agency administering such law. Address inquiries concerning Federal Old-Age Retirement Benefits (not State Old-Age Assistance or Pensions) to the nearest office of the Social Security Board.
SIGN THIS CARD IMMEDIATELY AND REPORT THE NUMBER TO YOUR EMPLOYER.

http://braintour.harvard.edu/wp-content/uploads/2016/05/tile_TT_SmittyStevens.jpg
https://pbs.twimg.com/media/DICH9C9XsAYGaLG.jpg
https://s3.envato.com/files/186945900/008(basket04_4color_apples).jpg
https://www.ft.com/content/64d6dbc0-5275-11e6-9664-e0bdc13c3bef
https://www.guru99.com/images/MongoDB/112115_0607_Introductio11.png
https://upload.wikimedia.org/wikipedia/commons/0/03/PD_social_security_card.png

```
> db.Employee.find().forEach(printjson);
{
        "_id" : ObjectId("563479cc8a8a4246bd27d784"),
        "Employeeid" : 1,
        "EmployeeName" : "Smith"
}
{
        "_id" : ObjectId("563479d48a8a4246bd27d785"),
        "Employeeid" : 2,
        "EmployeeName" : "Mohan"
}
{
        "_id" : ObjectId("563479df8a8a4246bd27d786"),
        "Employeeid" : 3,
        "EmployeeName" : "Joe"
}
>
```

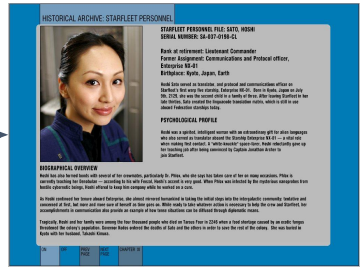
every row has a unique object id

Typology of Numbers

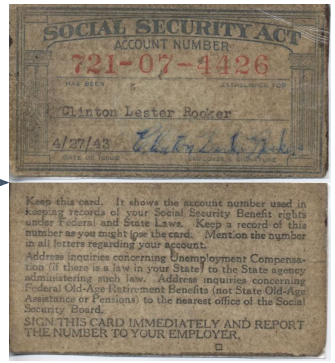
Stanley Smith Stevens

Nominal

Sequential



Hierarchical



Categorical

	A	B	C	D	E	F	G	H	I	J
3	Age	Party	Gender	Income		Age	Party 1	Party 2	Gender 1	Income
4	20	Rep	Male	45000		20	1	0	1	45000
5	25	Dem	Male	39000		25	0	1	1	39000
6	45	Ind	Male	56000		45	0	0	1	56000
7	35	Rep	Female	49000		35	1	0	0	49000
8	50	Dem	Female	41000		50	0	1	0	41000
9	55	Ind	Female	42000		55	0	0	0	42000
10	39	Rep	Male	58000		39	1	0	1	58000
11	48	Dem	Male	55000		48	0	1	1	55000
12	30	Ind	Male	46000		30	0	0	1	46000
13	27	Rep	Female	42000		27	1	0	0	42000
14	47	Dem	Female	37000		47	0	1	0	37000
15	21	Ind	Female	25000		21	0	0	0	25000
16	48	Rep	Male	75000		48	1	0	1	75000
17	24	Ind	Male	43000		24	0	0	1	43000
18	28	Ind	Female	40000		28	0	0	0	40000
19	40	Dem	Female	31000		40	0	1	0	31000

Random

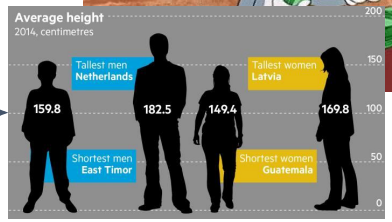
```
> db.Employee.find().forEach(printjson);
{
        "_id" : ObjectId("563479cc8a8a4246bd27d784"),
        "Employeeid" : 1,
        "EmployeeName" : "Smith"
}
{
        "_id" : ObjectId("563479d48a8a4246bd27d785"),
        "Employeeid" : 2,
        "EmployeeName" : "Mohan"
}
{
        "_id" : ObjectId("563479df8a8a4246bd27d786"),
        "Employeeid" : 3,
        "EmployeeName" : "Joe"
}
```

Ordinal

Interval + Ratio

(Simple) Count

Other



http://braintour.harvard.edu/wp-content/uploads/2016/05/tile_TT_SmittyStevens.jpg
https://pbs.twimg.com/media/DICH9C9XsAYGaLG.jpg
https://s3.envato.com/files/186945900/008(basket04_4color_apples).jpg
https://www.ft.com/content/64d6dbc0-5275-11e6-9664-e0bdc13c3bef
https://www.guru99.com/images/MongoDB/112115_0607_Introductio11.png
https://upload.wikimedia.org/wikipedia/commons/0/03/PD_social_security_card.png
https://thumbs.dreamstime.com/z/numbers-pedestal-sport-winners-golden-silver-bronze-marble-podium-first-second-third-place-isolated-white-49129110.jpg

Detection

Types Breakdown:

Nominal

-Sequential: 7000 to 9000

-Hierarchical: 2-88-12-3-1-1234

-Categorical: 1,1,1,2,2

-Random: 1239231,209,938423

Ordinal:

-Ordinal: 1,2,3,4

Interval-Ratio:

-Counts: 1,5,14,124

-Other: 169,173,181

Detection Order:

1-Ordinal: 1,2,3,4

2-Categorical: 1,1,1,2,2

3-Sequential: 7000 - 9000

4-Hierarchical: 2-88-12-3-1-1234

5-Counts: 1,5,14,124

6-Other: 169,173,181

?-Random: 1239231,209,938423

Detection Order:

1-Ordinal: 1,2,3,4

X = input data
Y = 1,2, … Max(X).

$$||X \cap Y|| > \sqrt{||Y||}$$

2-Categorical: 1,1,1,2,2

3-Sequential: 7000 - 9000

4-Hierarchical: 2-88-12-3-1-1234

5-Counts: 1,5,14,124

6-Other: 169,173,181

?-Random: 1239231,209,938423

Detection Order:

1-Ordinal: 1,2,3,4

2-Categorical: 1,1,1,2,2

X = input data
U = set(X)

$$1 < ||U|| < \sqrt{||X||}$$

3-Sequential: 7000 - 9000

4-Hierarchical: 2-88-12-3-1-1234

5-Counts: 1,5,14,124

6-Other: 169,173,181

?-Random: 1239231,209,938423

Detection Order:

1-Ordinal: 1,2,3,4

2-Categorical: 1,1,1,2,2

3-Sequential: 7000 - 9000

X = input data
Y = Min(X),Min(X)+c, …,Max(X)
c = constant

$$||X \cap Y|| > \sqrt{||Y||}$$

4-Hierarchical: 2-88-12-3-1-1234

5-Counts: 1,5,14,124

6-Other: 169,173,181

?-Random: 1239231,209,938423

Detection Order:

1-Ordinal: 1,2,3,4

2-Categorical: 1,1,1,2,2

3-Sequential: 7000 - 9000

4-Hierarchical: 2-88-12-3-1-1234

5-Counts: 1,5,14,124

6-Other: 169,173,181

?-Random: 1239231,209,938423

X = input data

$$num\_of\_digits(x_i) = num\_of\_digits(x_{i+1})$$
$$\forall x_i \in X$$

Detection Order:

1-Ordinal: 1,2,3,4

2-Categorical: 1,1,1,2,2

3-Sequential: 7000 - 9000

4-Hierarchical: 2-88-12-3-1-1234

5-Counts: 1,5,14,124 →

$$1.5 * (Q_3 - Q_1) + Q_3 \leqslant P_{95}$$

$$\frac{(P_{95} - Q_2)}{Q_2} \geqslant 2$$

P_a : ath percentile

Q_b : bth Quartile

6-Other: 169,173,181

?-Random: 1239231,209,938423

Detection Order:

1-Ordinal: 1,2,3,4

2-Categorical: 1,1,1,2,2

3-Sequential: 7000 - 9000

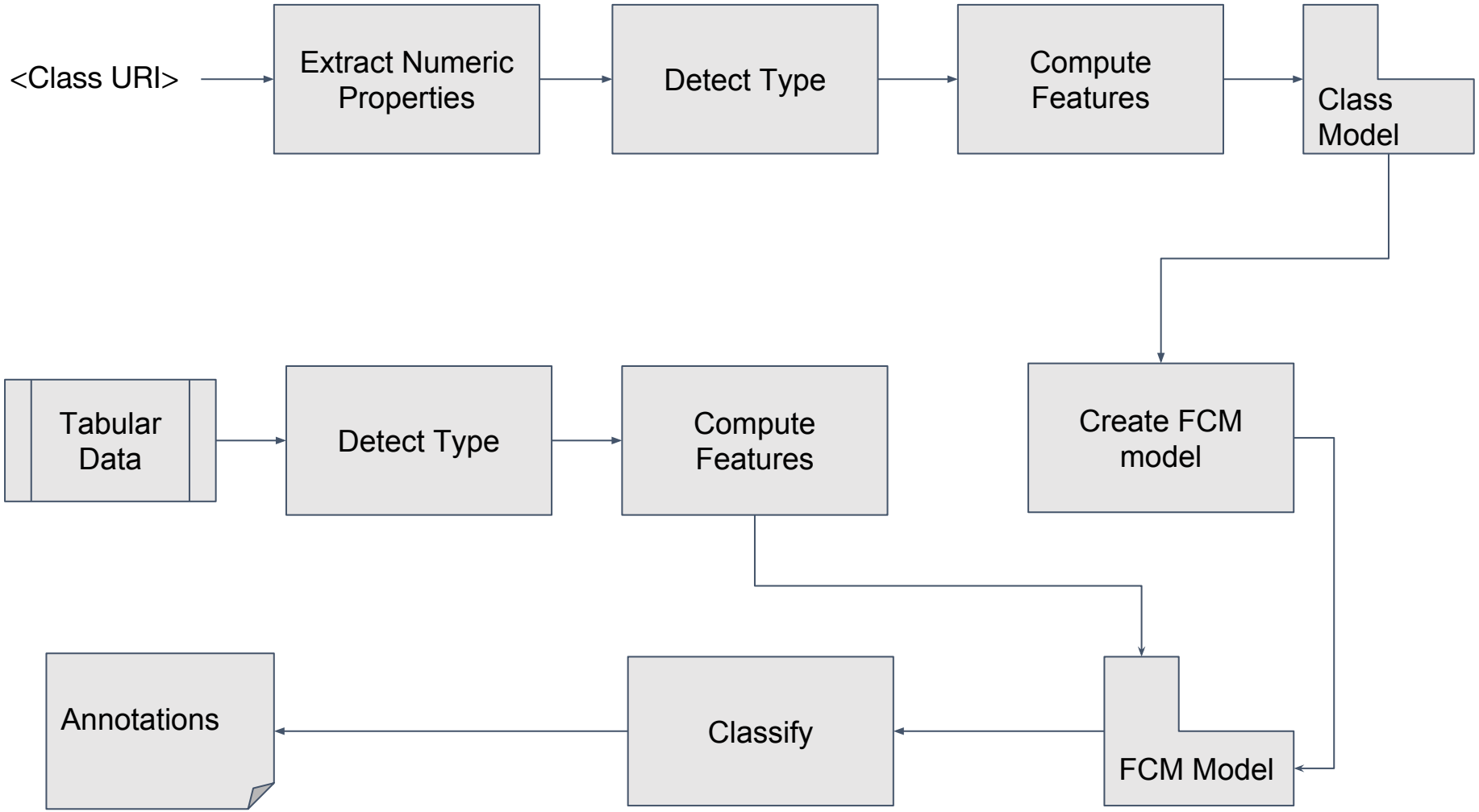
4-Hierarchical: 2-88-12-3-1-1234

5-Counts: 1,5,14,124

6-Other: 169,173,181 → Everything else

?-Random: 1239231,209,938423

Labeling Workflow



1. Get properties

```
SELECT distinct ?property WHERE {
?subject a <classURI>. ?subject ?property [].
} GROUP BY ?property
```

2. Filter numeric properties

If > 50% are numeric

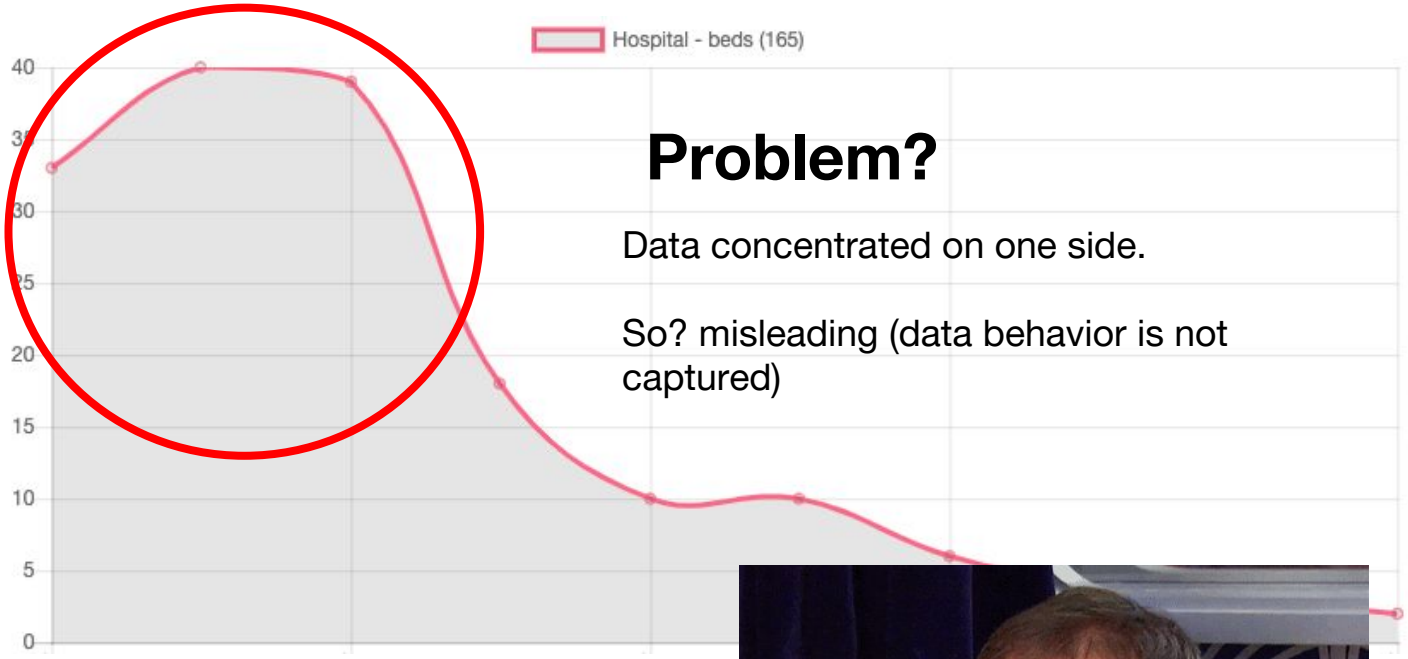
Features

(Sub-)Type	Features
Ordinal	tri-mean, tstd
Sequential	tri-mean, tstd
Categorical	num of categories, percentages of each (ordered)
Hierarchical	-
Counts	tri-mean, tstd (of re-expressed data)
Other	tri-mean, tstd

$$trimean = \frac{Q1 + 2 * Q2 + Q3}{4}$$

$$tstd = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - trimean)^2}$$

Long-tail problem



Problem?

Data concentrated on one side.

So? misleading (data behavior is not captured)

How to fix this?

Data re-expression

Long-tail problem

$$\sqrt{x_i}\ \forall x_i \in X$$

https://github.com/oeg-upm/property_cake

Features

(Sub-)Type	Features
Ordinal	tri-mean, tstd
Sequential	tri-mean, tstd
Categorical	num of categories, percentages of each (ordered)
Hierarchical	-
Counts	tri-mean, tstd (of re-expressed data)
Other	tri-mean, tstd

$$trimean = \frac{Q1 + 2 * Q2 + Q3}{4}$$

$$tstd = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - trimean)^2}$$

$$\sqrt{x_i}\ \forall x_i \in X$$

Model

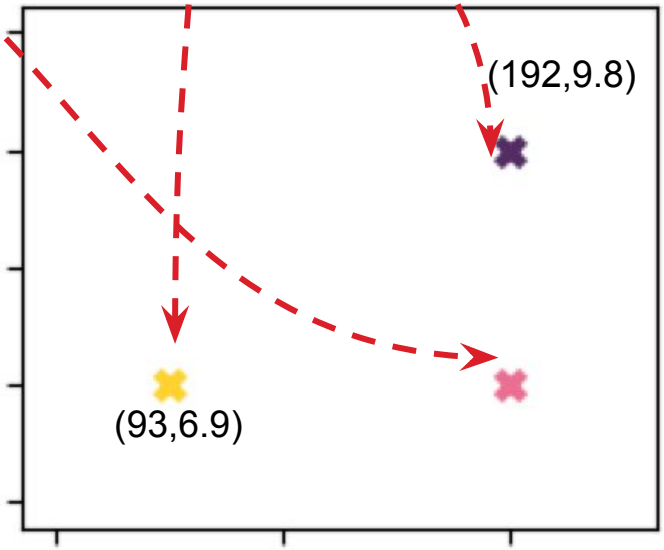
Property URI	Numeric Type/Sub-type	Features
.../military-service-number	sequential	95000, ...
.../height	other	192.4, ...
...	...	...

FCM Model

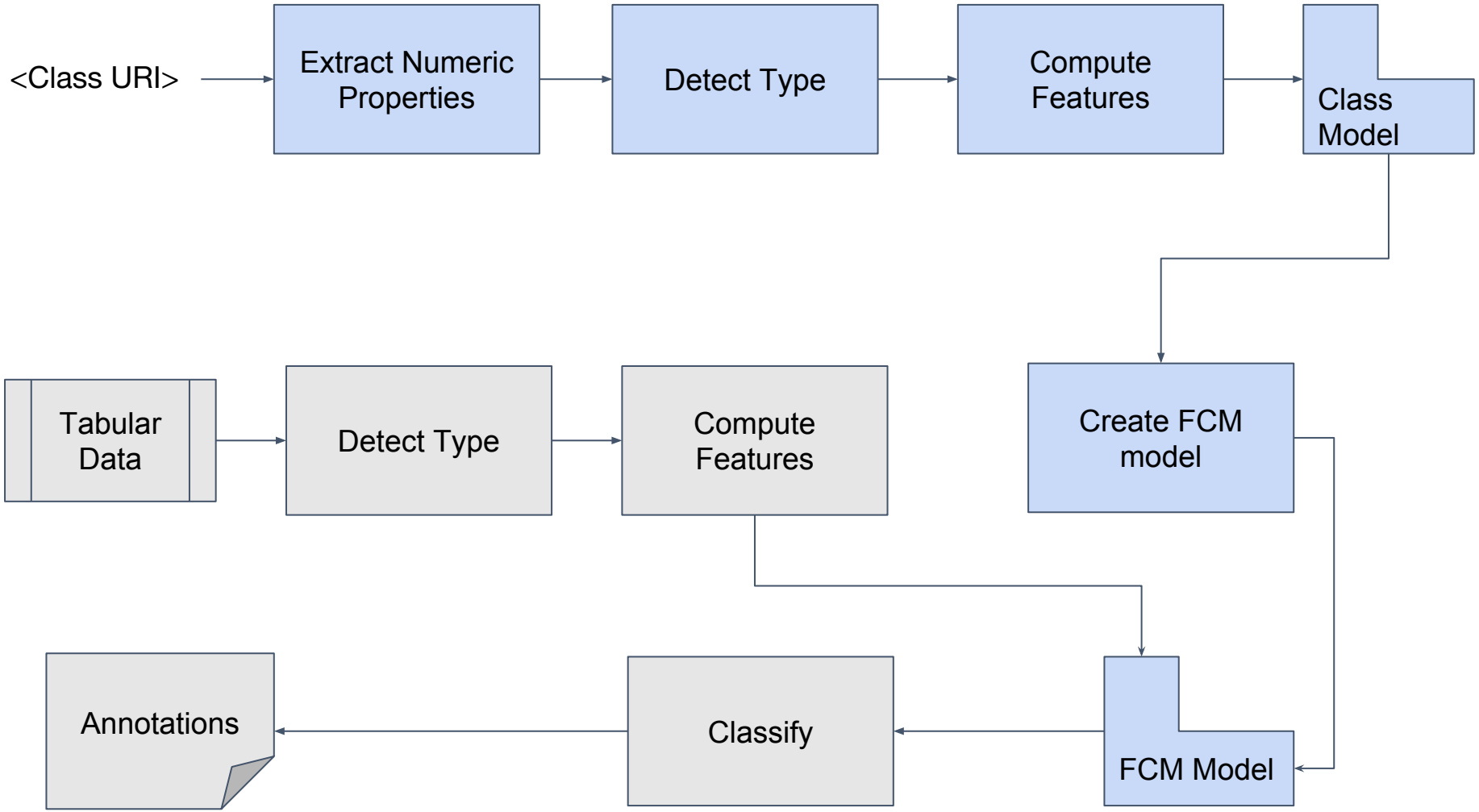
https://github.com/oeg-upm/fuzzy-c-means
https://pypi.org/project/fuzzycmeans/

FCM Model

Properties	Features
http://dbpedia.org/property/height	192,9.8
http://dbpedia.org/property/weight	93,6.9
...	...



Labeling Workflow



Hypothesis:

"Semantic labeling yields a higher precision score when taking the typology of the numeric values into account than using a general technique"

Data: T2Dv2 (We manually typed and annotated)

http://webdatacommons.org/webtables/goldstandardV2.html

Typology in T2Dv2 dataset

Numeric Type	Sub-type	Percentage
Nominal	Sequential	0.008
Nominal	Hierarchical	0.0
Nominal	Categorical	0.0
Nominal	Random	0.048
Nominal	combined	0.056
Ordinal	-	0.04
Ratio-Interval	Count	0.387
Ratio-Interval	Other	0.234
Ratio-Interval	combined	0.621
Year	-	0.282

Detection Score

Typology Detection Scores

Numeric Type	Sub-type	Precision	Recall	F1
Nominal	Sequential	0.0	0.0	N/A
Nominal	Hierarchical	N/A	N/A	N/A
Nominal	Categorical	N/A	N/A	N/A
Nominal	Random	N/A	N/A	N/A
Ordinal	-	0.8	1.0	0.889
Ratio-Interval	Count	0.792	0.809	0.8
Ratio-Interval	Other	0.552	0.516	0.533

Labeling Score

Compare Labeling Scores

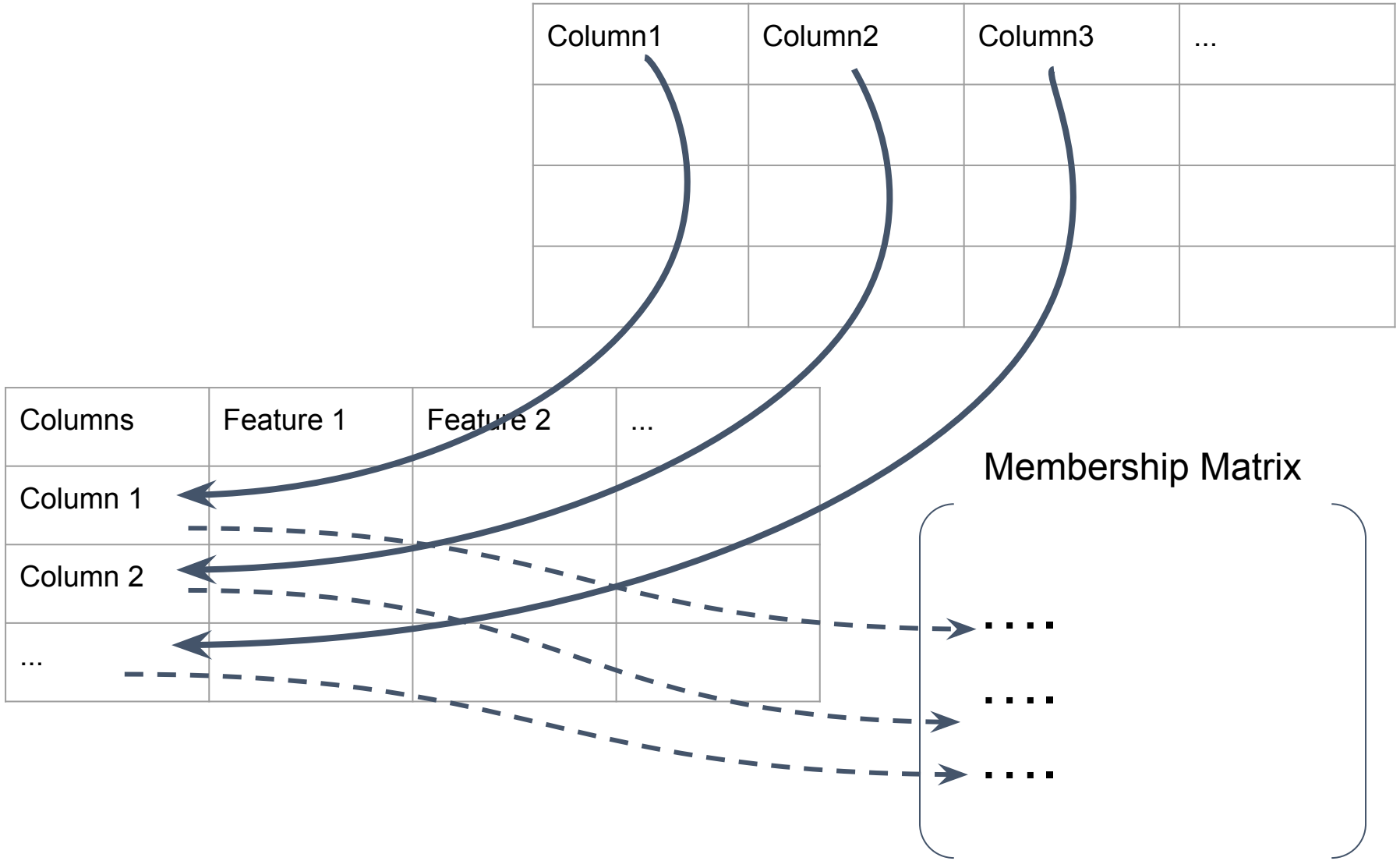
k	Approach	Precision	Recall	F1
1	TTLA	**0.687**	0.892	0.776
	FCM	0.34	-	-
	Random	0.0004	-	-
3	TTLA	**0.94**	0.892	0.915
	FCM	0.55	-	-
	Random	0.0012	-	-
5	TTLA	**0.976**	0.892	0.932
	FCM	0.83	-	-
	Random	0.002	-	-
10	TTLA	**1.0**	0.892	0.943
	FCM	0.91	-	-
	Random	0.004	-	-

Conclusion

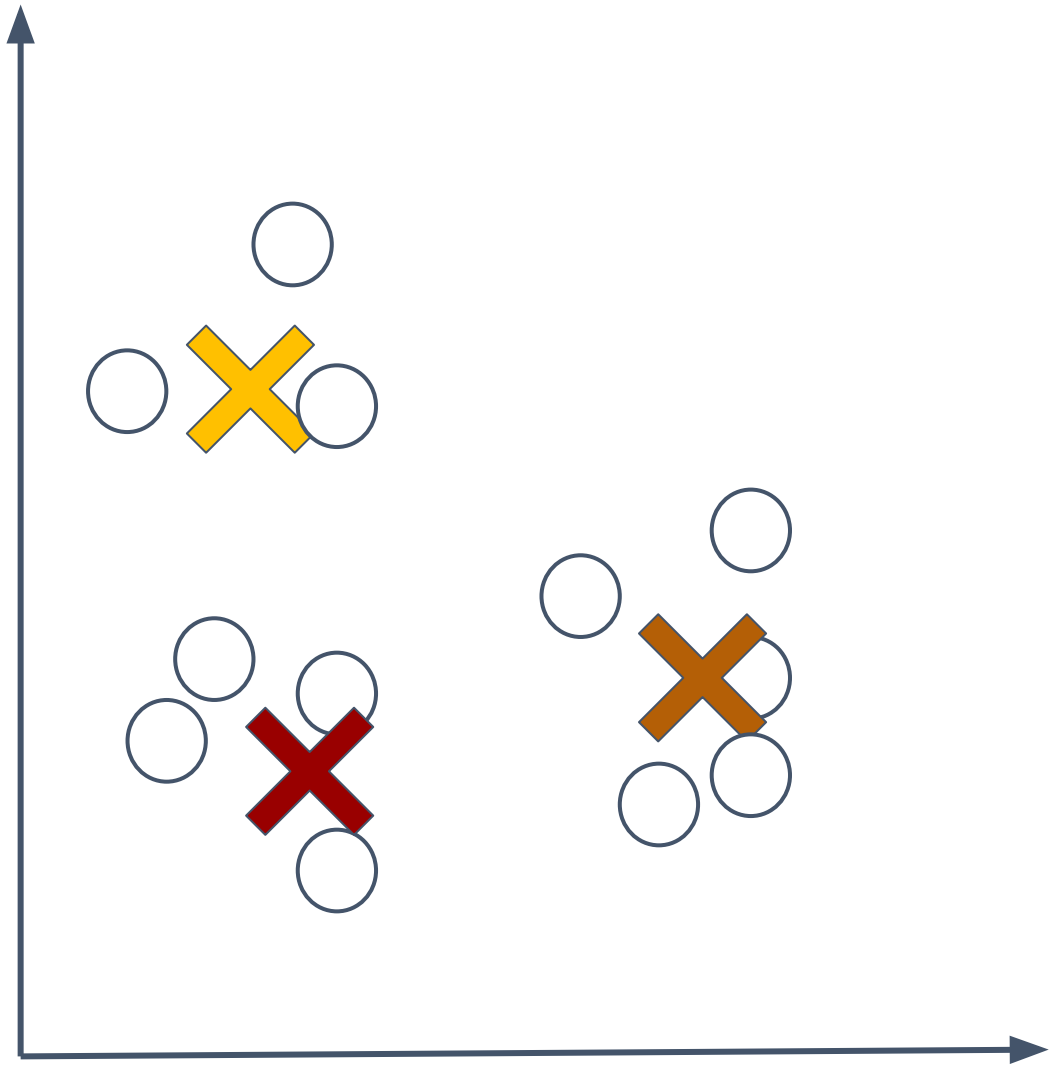
- Under-represented types in current benchmarks
- Taking into account typology yields a higher precision


BRICK
BRICK
QUESTIONS!

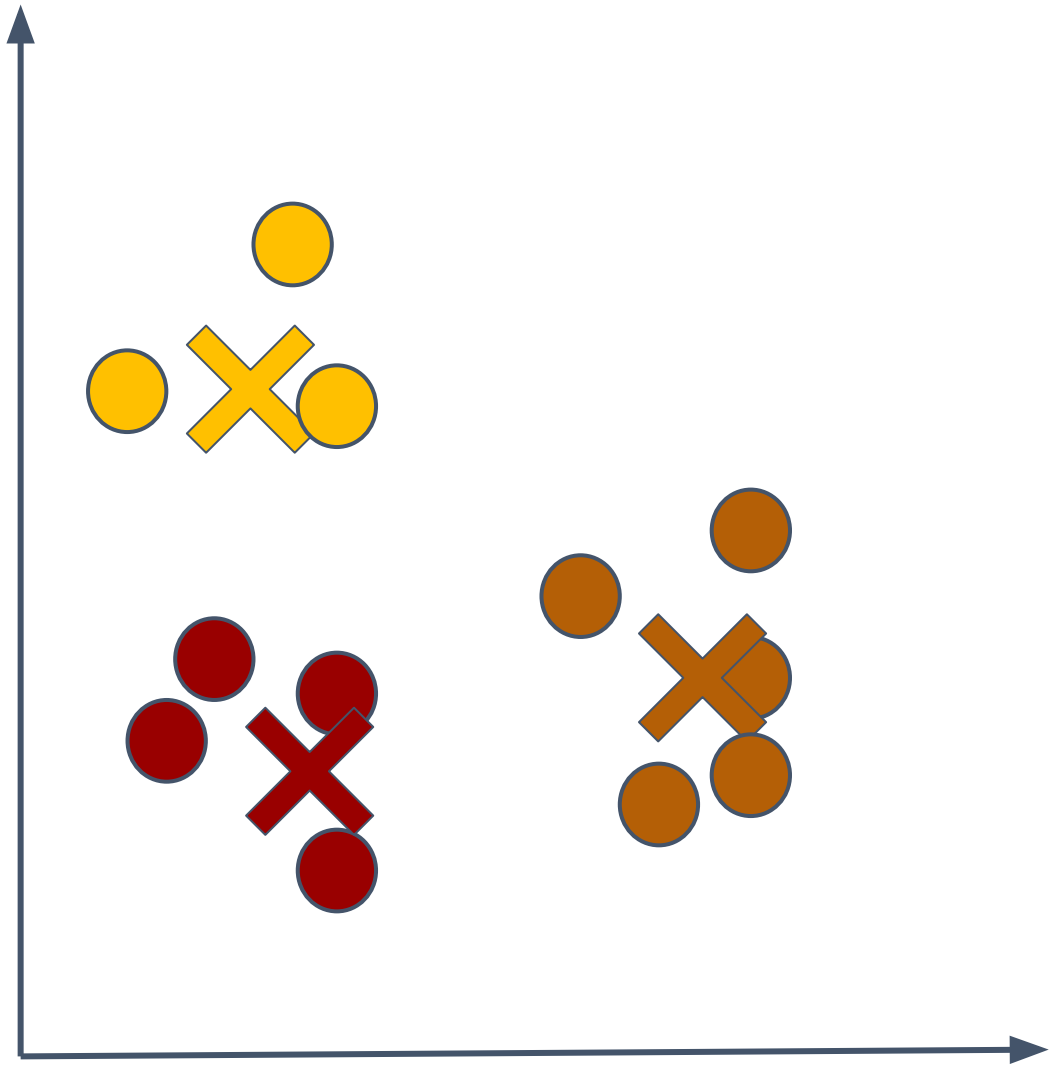
Computing membership



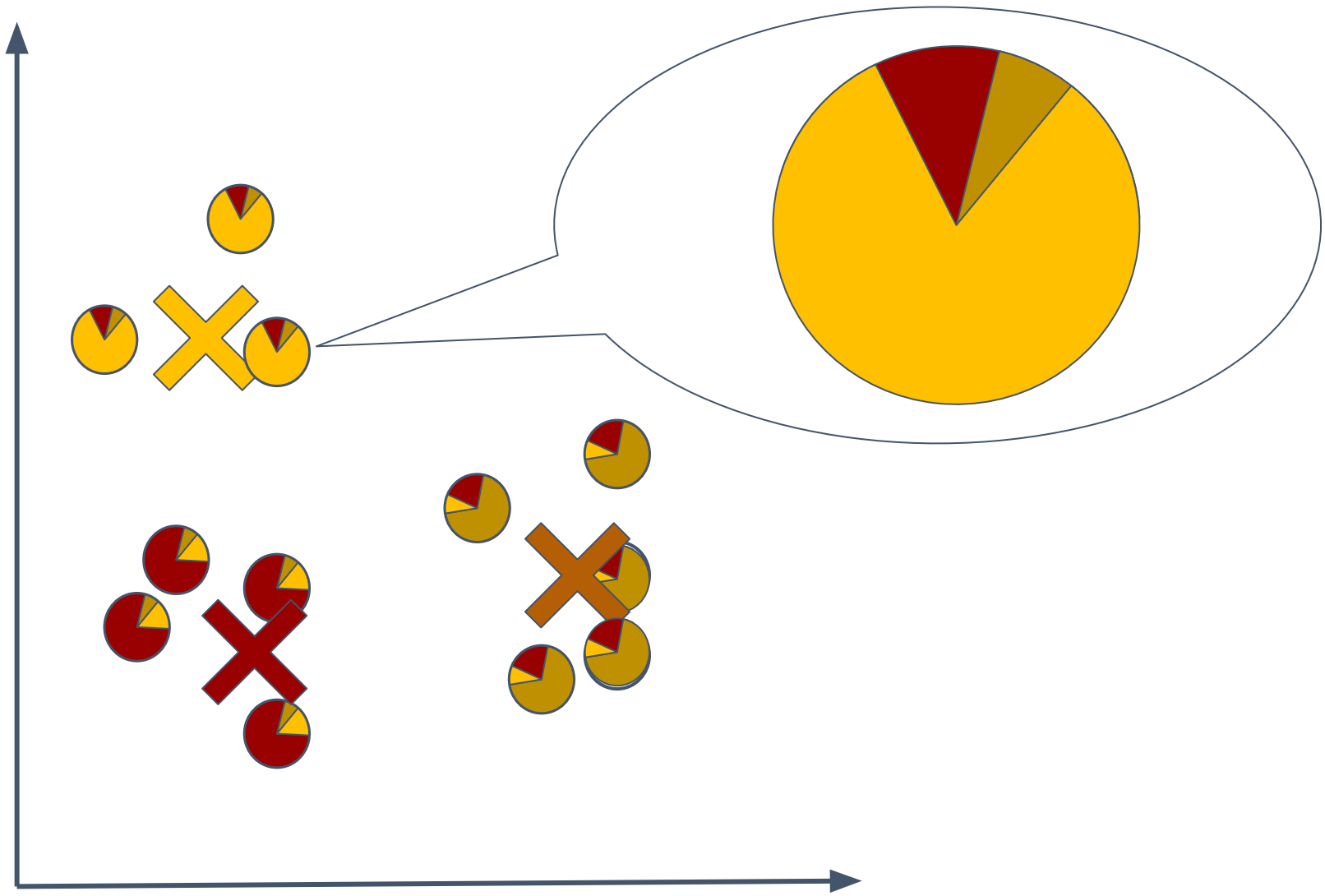
Clustering



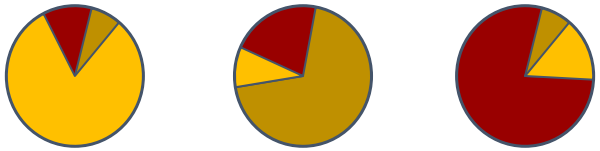
Hard Clustering



Membership



Membership Matrix



Clus 1	0.81	0.06	0.07	
Clus 2	0.03	0.92	0.89	
Clus 3	0.16	0.02	0.04	
	1.0	1.0	1.0	Sum