Predictive models for multidimensional data when the resolution context changes

José Hernández-Orallo[1], Nicolas Lachiche[2], and Adolfo Martínez-Usó[1]

[1]DSIC, Universitat Politècnica de València
{jorallo,admarus}@dsic.upv.es

[2]ICube, Université de Strasbourg
nicolas.lachiche@unistra.fr

LMCE 2014 – Nancy, 19th September 2014

Overview

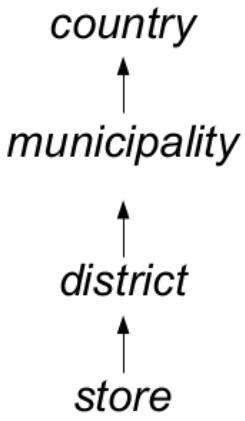
- Introduction
- Multidimensional contexts
- Experiments
  - Datamarts
  - Techniques
  - Context plots
  - Results
- Conclusions and Future work

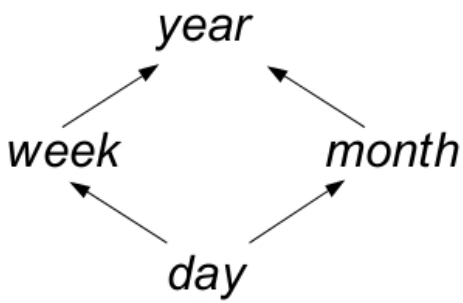
chist-era

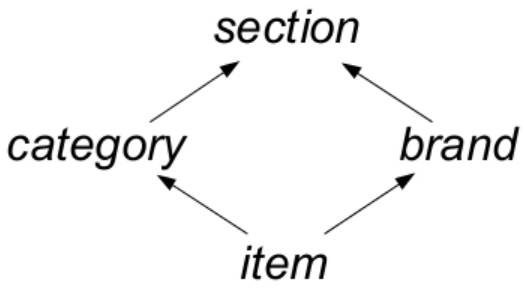
REFRAME

Multidimensional data

- Many applications use structured information in several dimensions





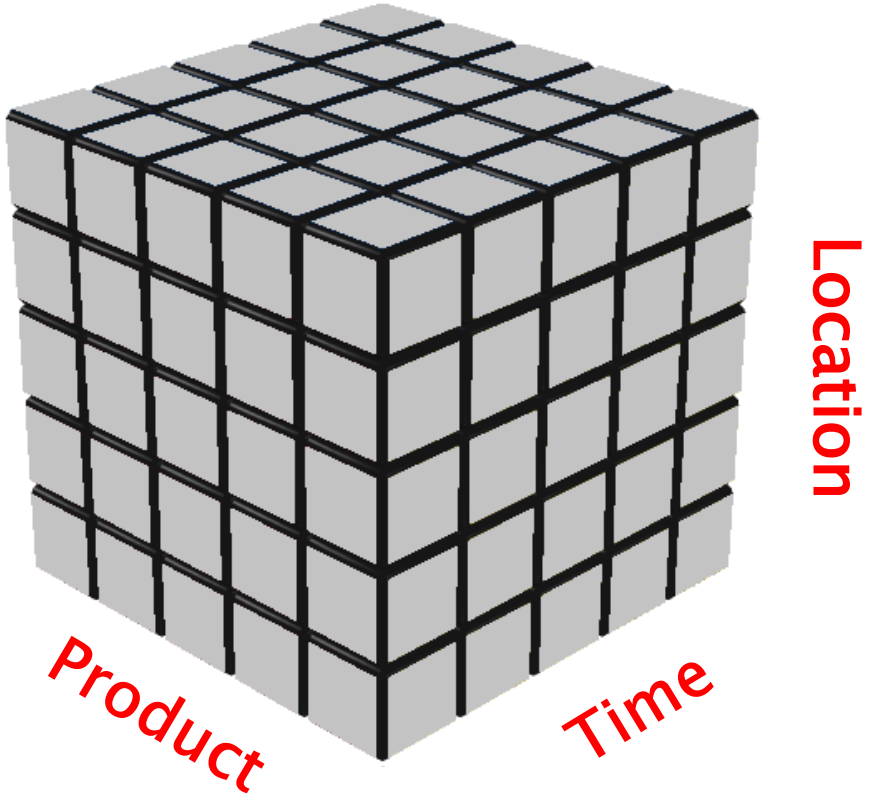


chist-era

REFRAME

Cubes represent contexts

- Data mining models are not designed to take hierarchical attributes



MD context: SL vs. LL (1/2)

- Two alternatives:

- One model for each operating context and apply it for that level of aggregation

SAME-LEVEL (SL)

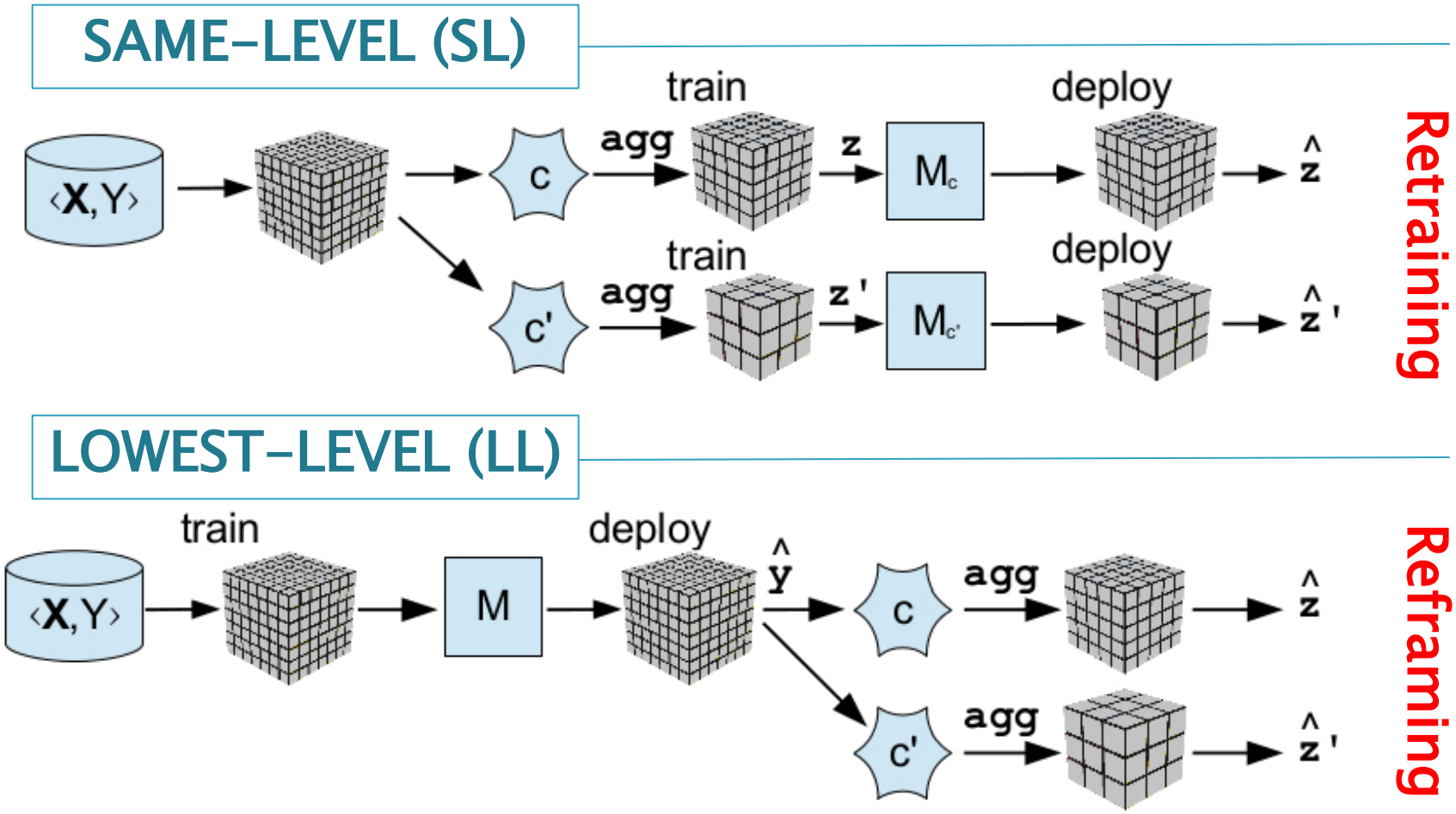
- One more versatile model at the lowest operating context and then aggregate its predictions

LOWEST-LEVEL (LL)

chist-era

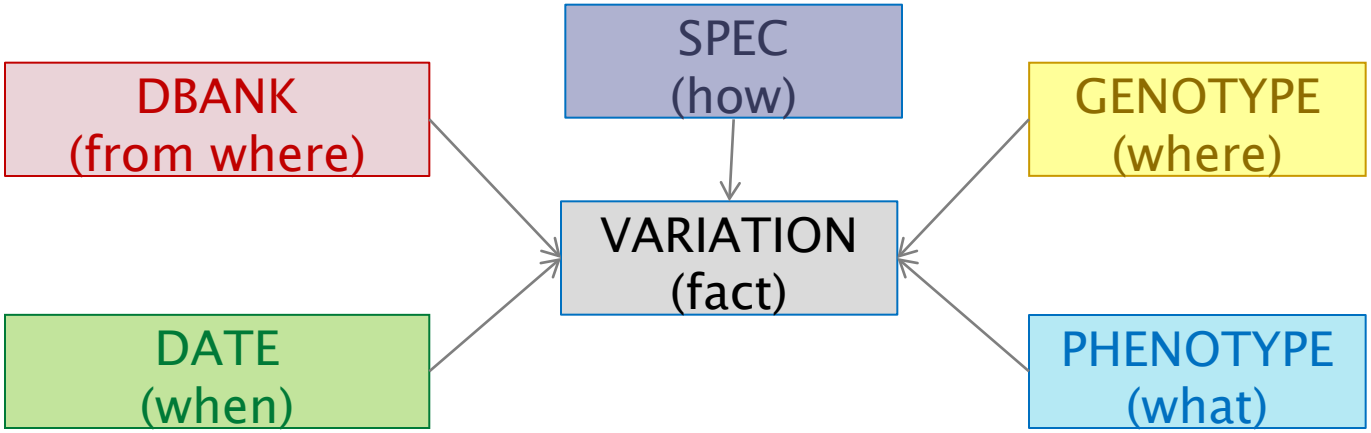
REFRAME

MD context: SL vs. LL (2/2)



Human genome (GENOMICS)

- Unified genomic variation repository to allow biologists to perform efficient recovery tasks about genomic mutations and their phenotype.
- Original schema:

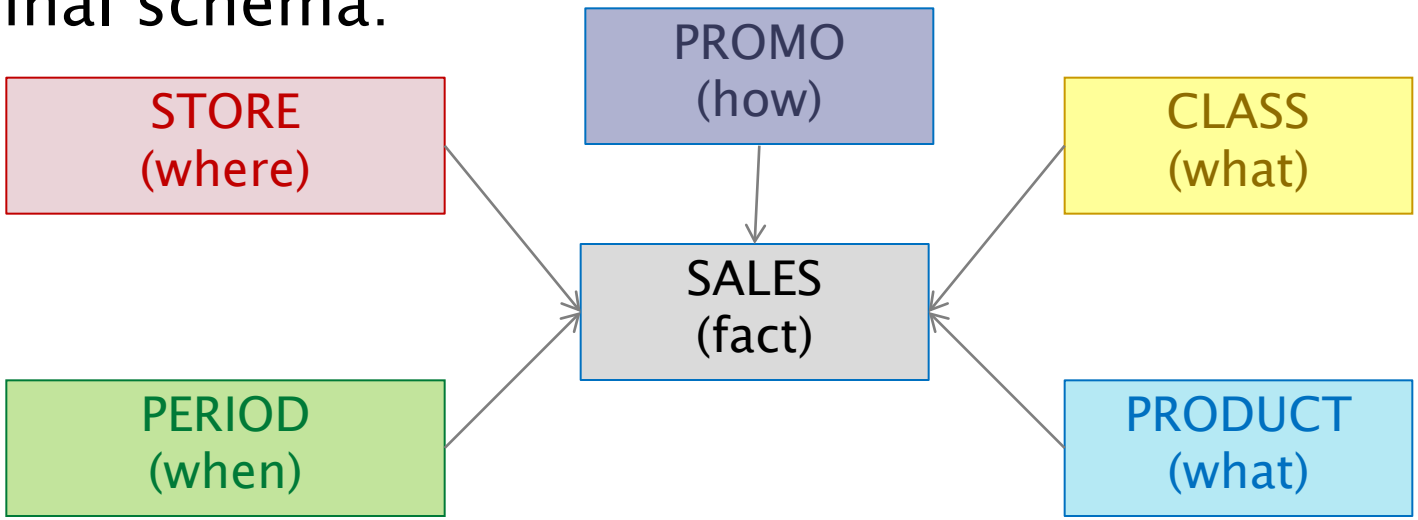


Fact:
There are 37 variations in chromosone 5 causing diseases of the category "cancer" with specialisation M discovered in 2012 and provided by any databank

- 5 dimensions and 48 possible multidimensional contexts (cubes).

IBM artificial (AROMA)

- This is an artificial dataset constructed from IBM sales information.
- Original schema:

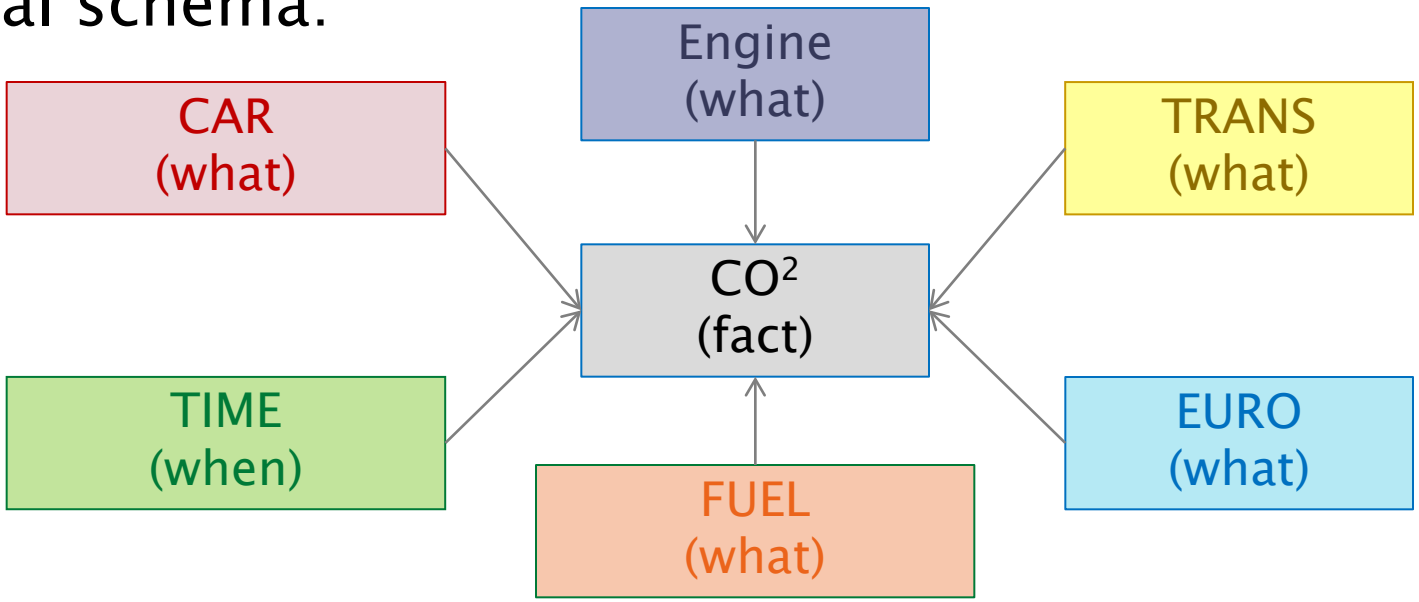


Fact:
The sale for class X of product "tomato" with promotion Y in september 2013 at Valencia store was 24,242 units (453,252 dollars)

- 5 dimensions and 84 possible multidimensional contexts (cubes).

CARS

- This dataset represents car fuel consumption and emissions.
- Original schema:



Fact:

CO^2 emission for car C within the euro standard with engine capacity X from year 2013 diesel and with automatic transmission had a concentration of 350

- 6 dimensions and 96 possible multidimensional contexts (cubes).

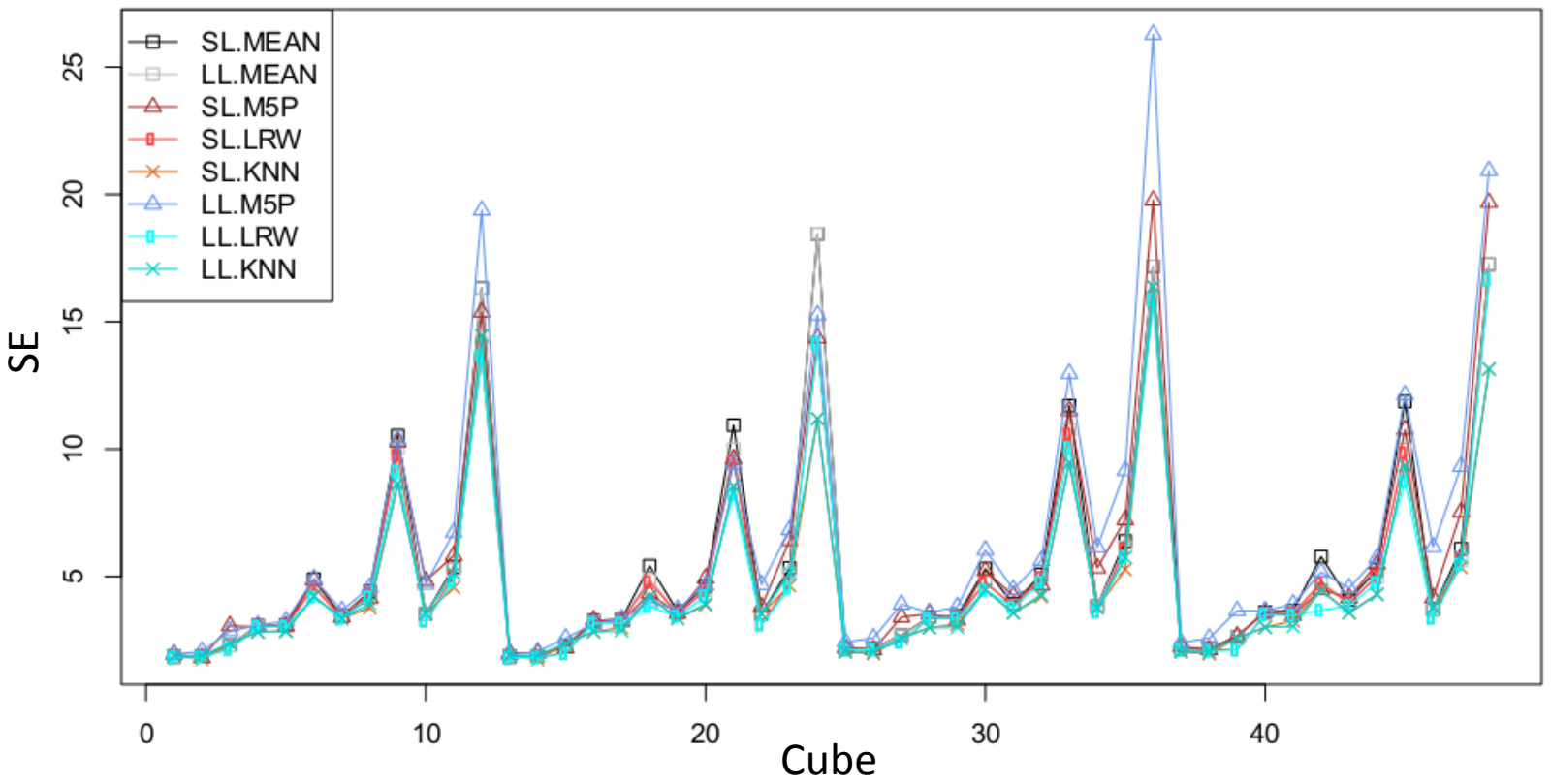
chist-era

REFRAME

Techniques

- We used four techniques:
  - MEAN
  - LRW (linear regression from Rweka)
  - M5P (regression tree from Rweka)
  - KNN (package kknn in R)

Multidimensional Context plots (1/4)



- **SE**: The higher the aggregation the higher the magnitudes but the number of rows decreases, so the magnitudes will be comparable.

Multidimensional Context plots (2/4)

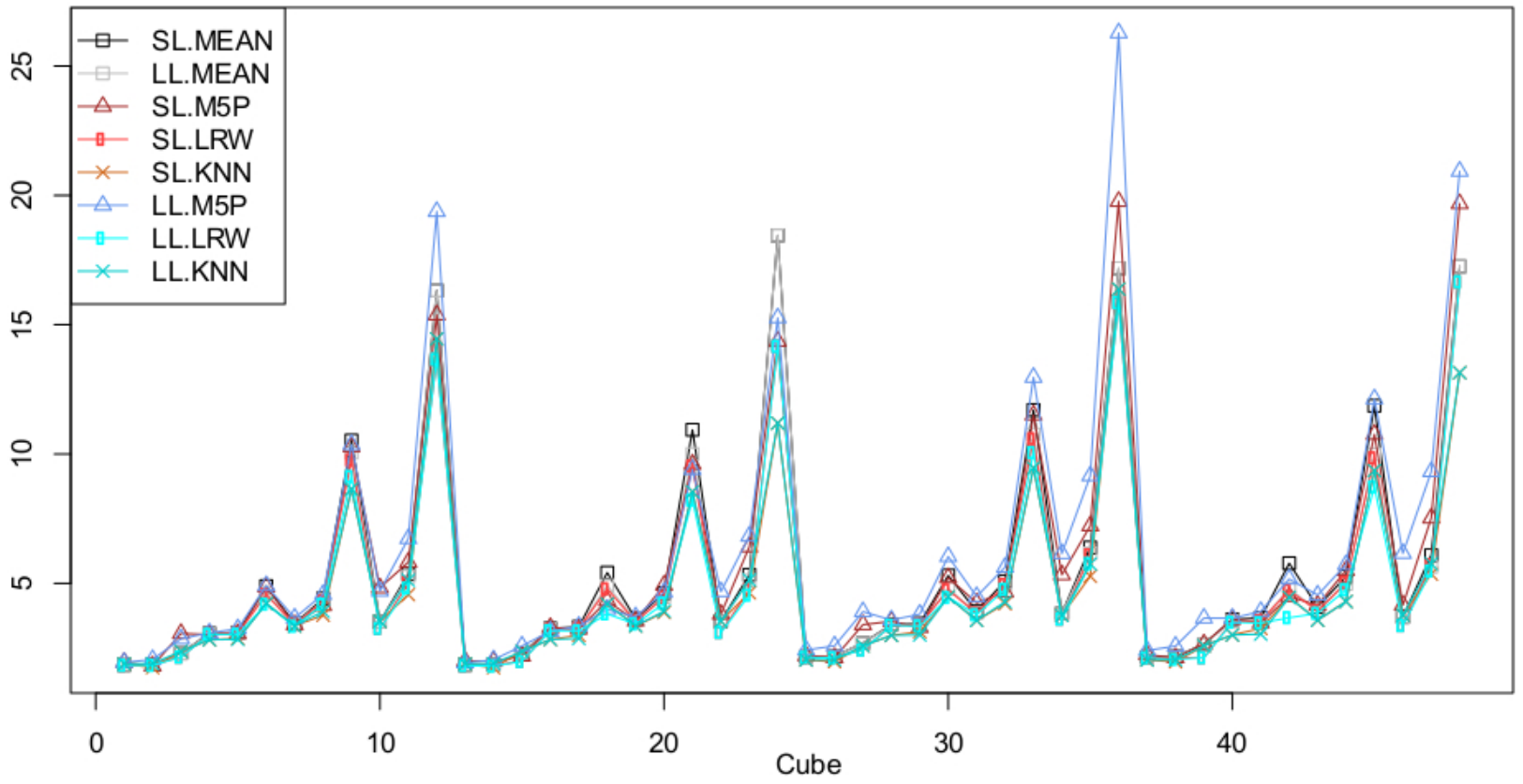
- Multidimensional context plots (MDC):
  - Normalised Squared Error (NSE) of a method (M)

$$NSE = \frac{MSE(\mathbf{M})}{MSE(\mathbf{SL.MEAN})}$$

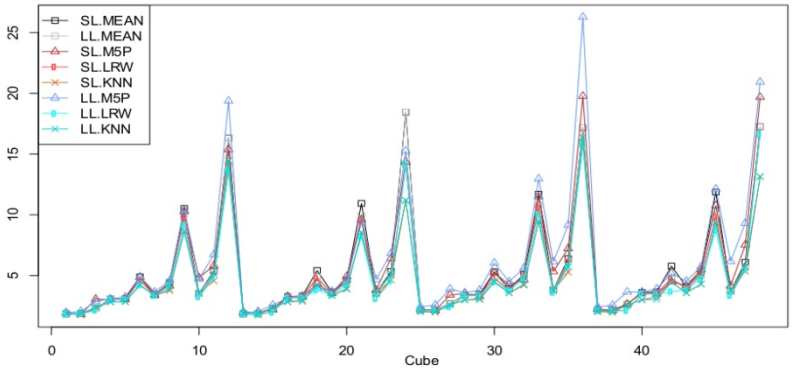
  - We use the SL.MEAN model since it will be constant for the multidimensional context during deployment.
  - We just see whether M is better or not than the mean model.

Multidimensional Context plots (3/4)

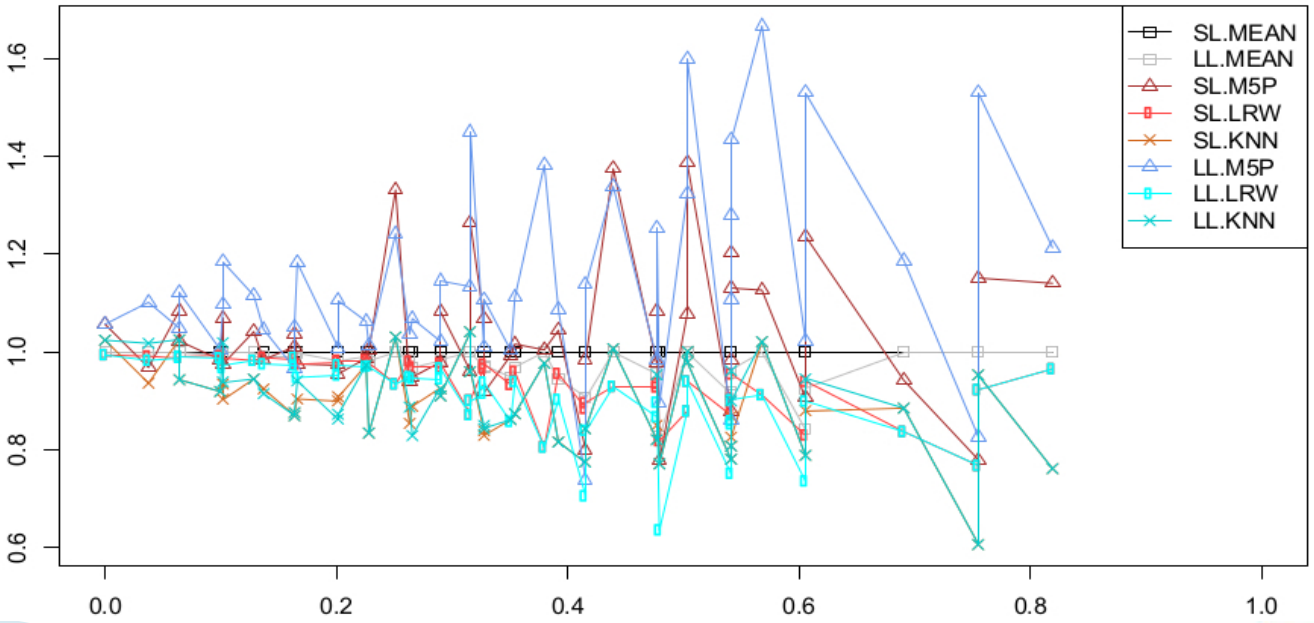
- Multidimensional context plots (MDC):



Multidimensional Context plots (4/4)

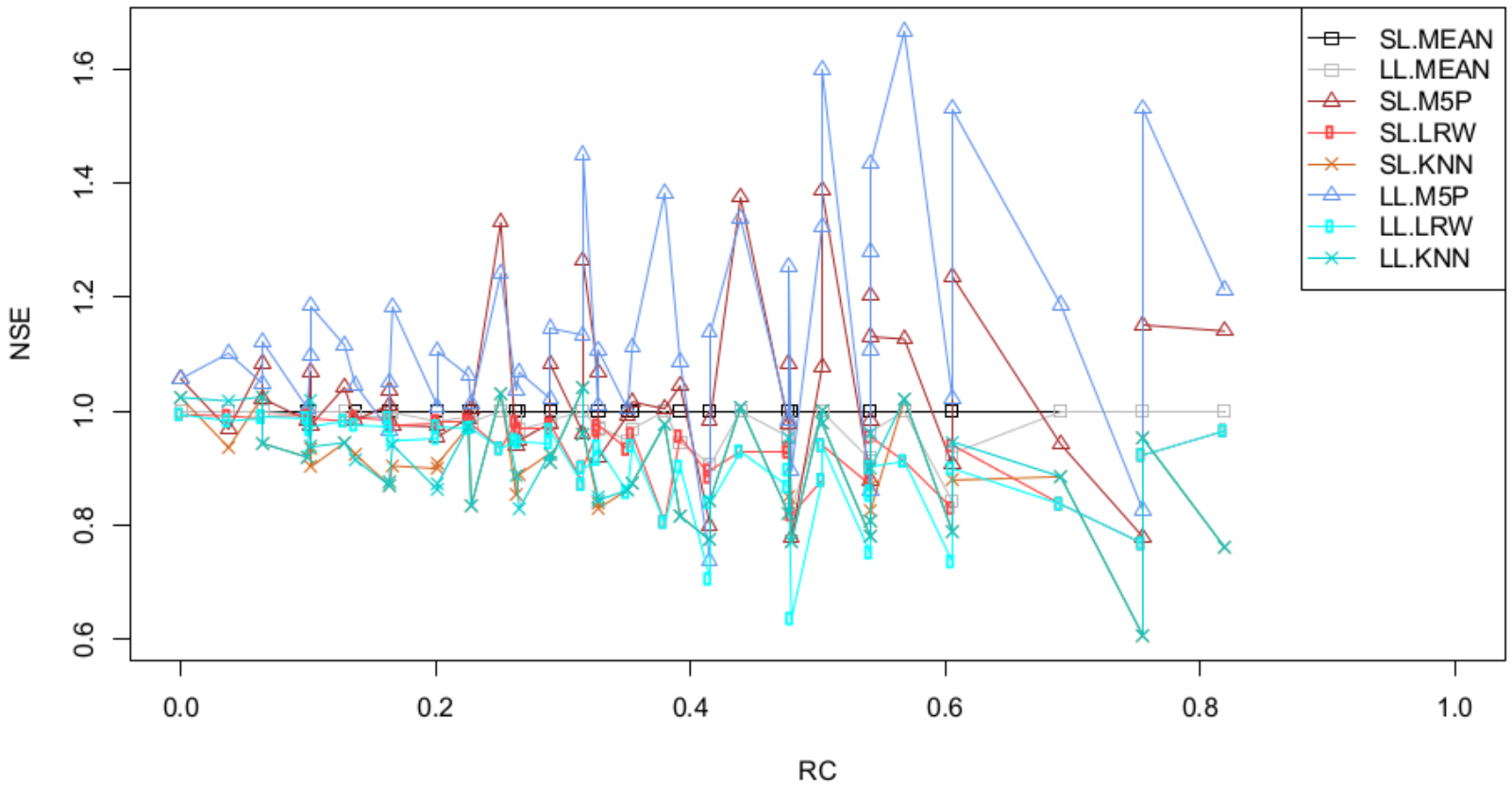


NSE

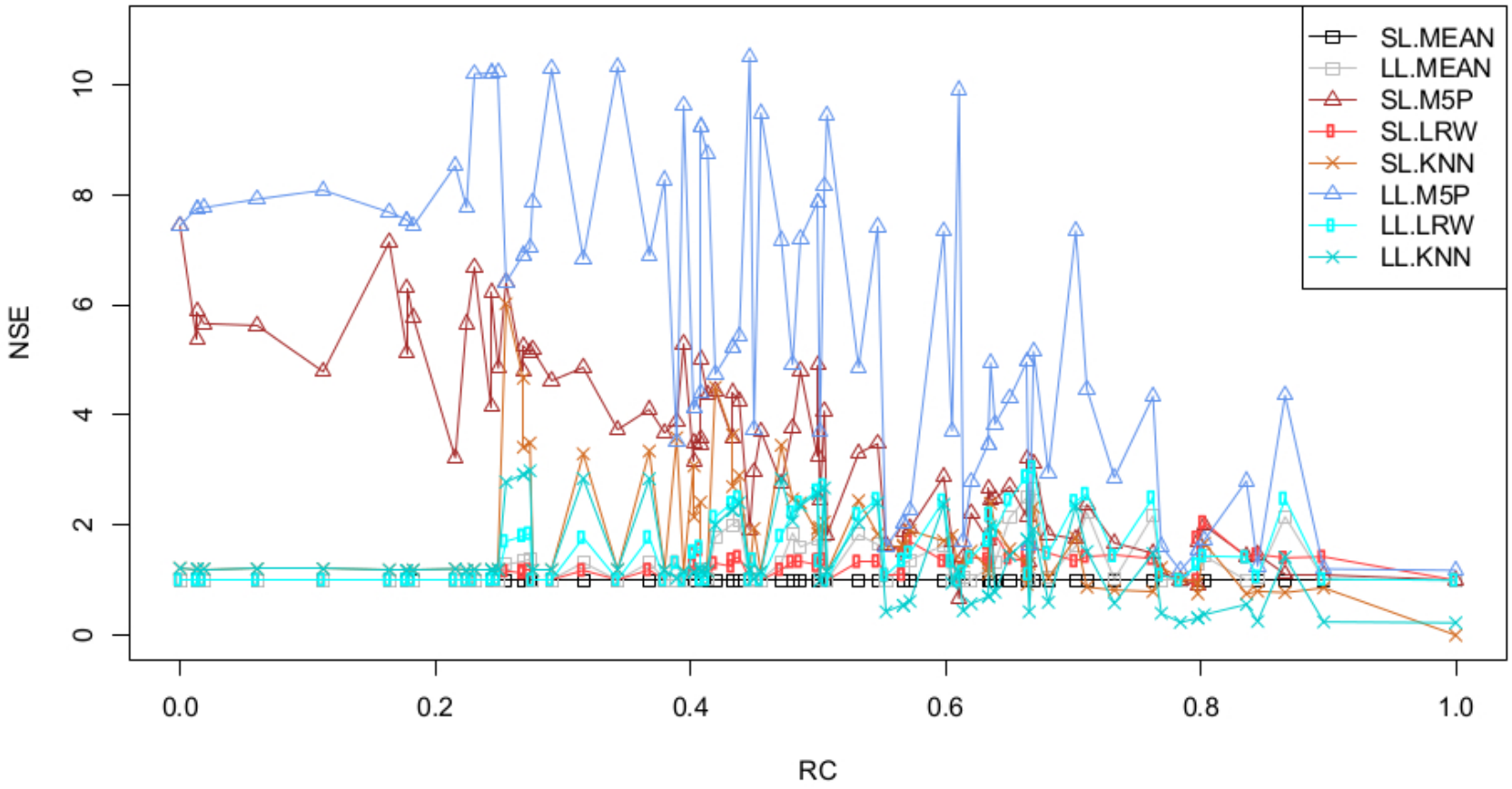


RC

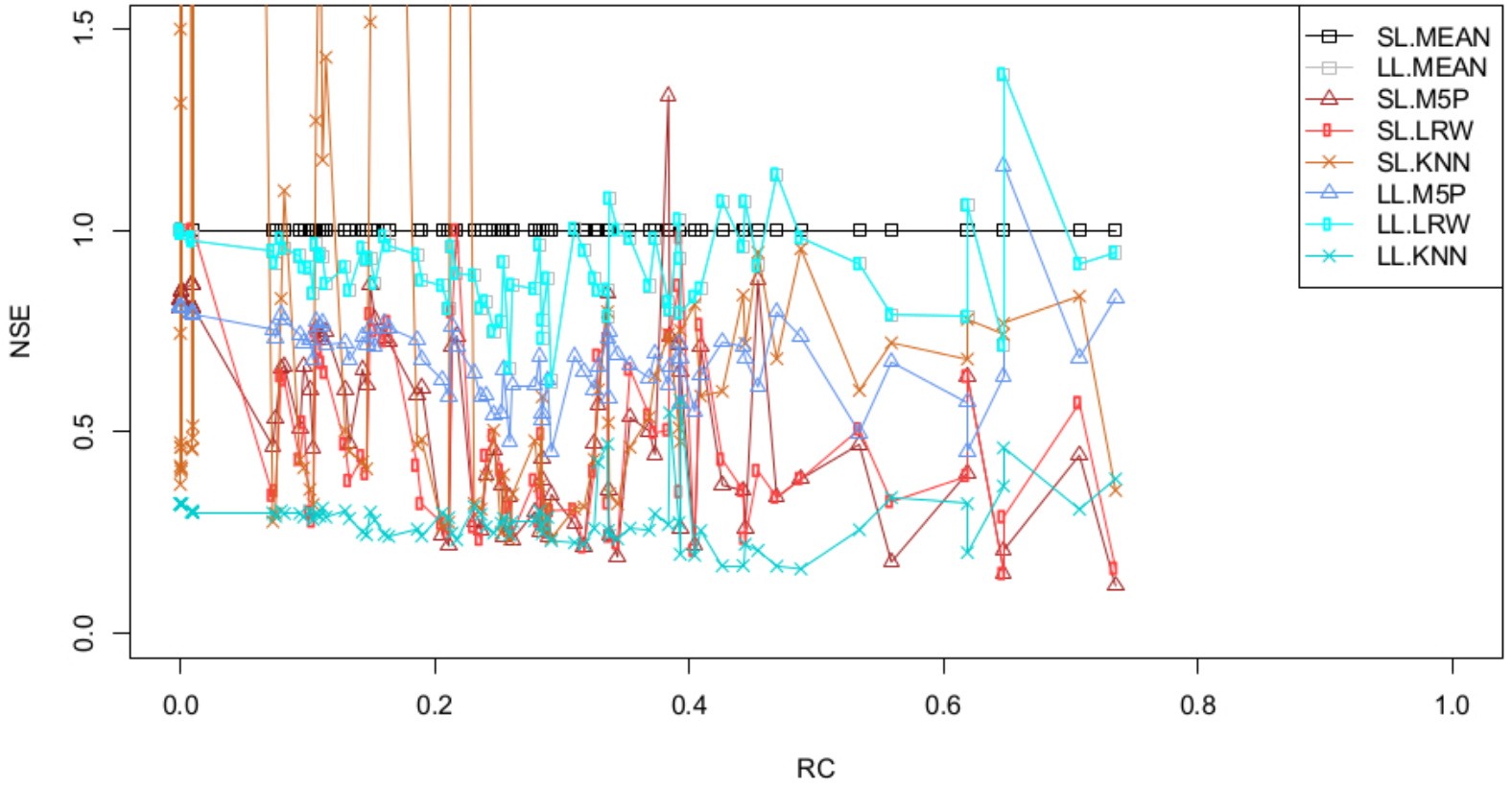
Graphical results: GENOMICS



Graphical results: AROMA



Graphical results: CARS



- No sparseness → this allows the models to make better predictions

Comparing LL and SL

	GENOMICS		AROMA		CARS	
	LL	SL	LL	SL	LL	SL
MEAN	0.97	1.00	1.31	1.00	0.93	1.00
M5P	1.14	1.03	5.81	3.44	0.71	0.57
LRW	0.91	0.94	1.56	1.24	0.93	0.60
KNN	0.90	0.89	1.36	1.74	0.29	1.31
Overall	0.98	0.97	2.51	1.86	0.71	0.87
N.Cubes	79	56	77	243	224	150

NSE Average

- LL: any bias in these predictions will accumulate further up and will lead to high error.
- SL: the models are learnt from aggregated data, and many rows will be aggregated into single rows with measures that are no longer zero.

chist-era

REFRAME

Conclusions

- MD data: the same task can change significantly depending on the level of aggregation
- Reframing vs. Retraining dilemma
- New plots and metrics
- Best choice depends on the dataset/technique but …
  - if !(sparse) → LL
  - LL–KNN generally works fine … M5P generally loses
- Resources are an important criterion
  - LL is more versatile and economical

Future work

- Find more datamarts and other ways of splitting the data
- Propose new plots and metrics
- LL approach using a quantification procedure
- Disaggregation: work at an upper level and then disaggregate
- Specific techniques devised for the MD setting:
  - MD kNN, MD Decision Trees, MD Naive Bayes, etc.

chist-era
REFRAME

Predictive models for multidimensional data when the resolution context changes

Thank you

chist-era

REFRAME