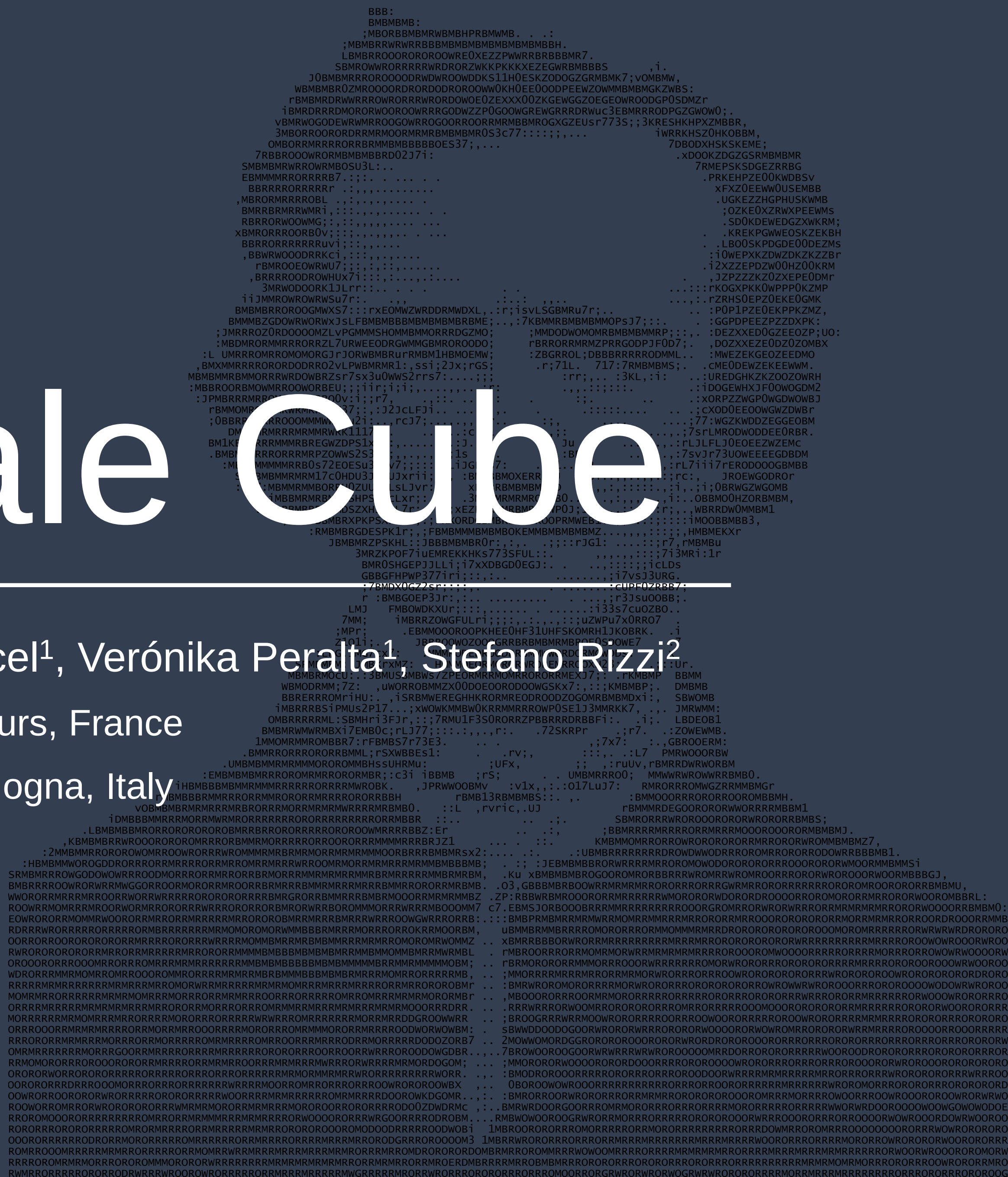
# The Tell-Tale Cube



Antoine Chédin[1], ***Matteo Francia***[2], Patrick Marcel[1], Verónika Peralta[1], Stefano Rizzi[2]

[1]LIFAT, University of Tours, France

[2]DISI, University of Bologna, Italy

# Intentional Analytics Model

Intentional Analytics Model (IAM) [1]

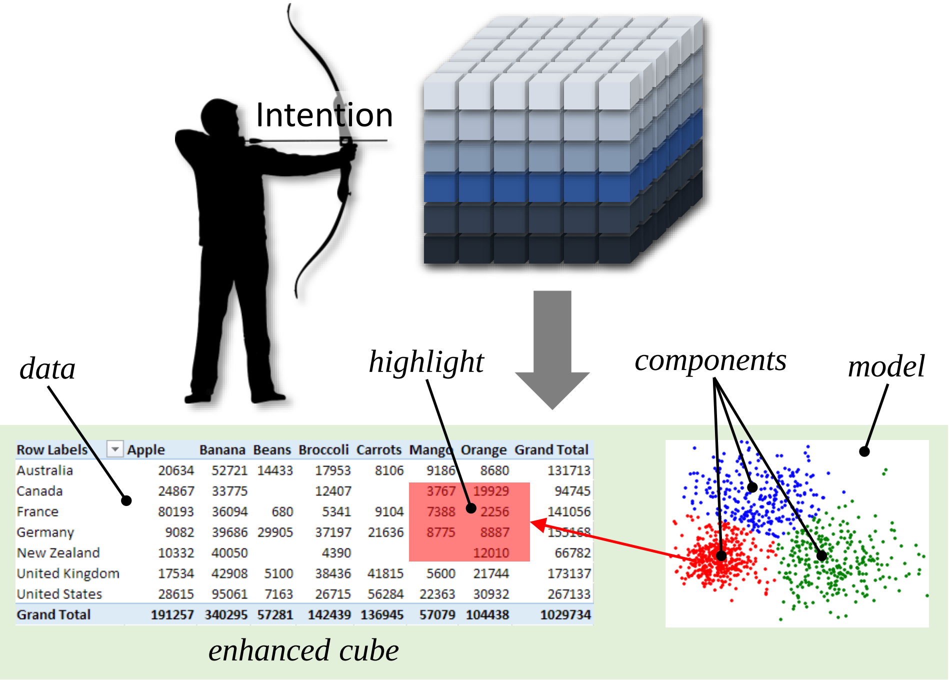
- Facilitate analysis of multidimensional cubes
- Bring OLAP to higher abstraction level
- Escape from query answers as sets of tuples

Express high-level *intentions*, not queries

- <u>Describe</u>, Assess, Explain, Predict, Suggest

Get cubes enhanced with insights

- Apply (mining/ML) models to data
- Highlight interesting components

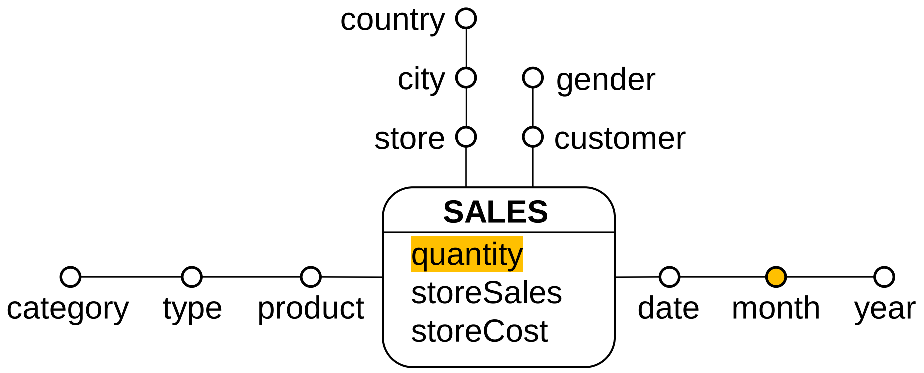



| Row Labels | Apple | Banana | Beans | Broccoli | Carrots | Mango | Orange | Grand Total |
|---|---|---|---|---|---|---|---|---|
| Australia | 20634 | 52721 | 14433 | 17953 | 8106 | 9186 | 8680 | 131713 |
| Canada | 24867 | 33775 | | 12407 | | 3767 | 19929 | 94745 |
| France | 80193 | 36094 | 680 | 5341 | 9104 | 7388 | 2256 | 141056 |
| Germany | 9082 | 39686 | 29905 | 37197 | 21636 | 8775 | 8887 | 155168 |
| New Zealand | 10332 | 40050 | | 4390 | | | 12010 | 66782 |
| United Kingdom | 17534 | 42908 | 5100 | 38436 | 41815 | 5600 | 21744 | 173137 |
| United States | 28615 | 95061 | 7163 | 26715 | 56284 | 22363 | 30932 | 267133 |
| **Grand Total** | **191257** | **340295** | **57281** | **142439** | **136945** | **57079** | **104438** | **1029734** |

[1] Panos Vassiliadis, Patrick Marcel, Stefano Rizzi: Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP. Inf. Syst. 85: 68-91 (2019)

# Classical OLAP

Query the cube, get a plain table




Identify interesting patterns / cells

- What if we have thousands of cells?
- Can we have an effective representation?

```
select month, sum(quantity)
from sales_ft join date_dt on (…)
group by month
```

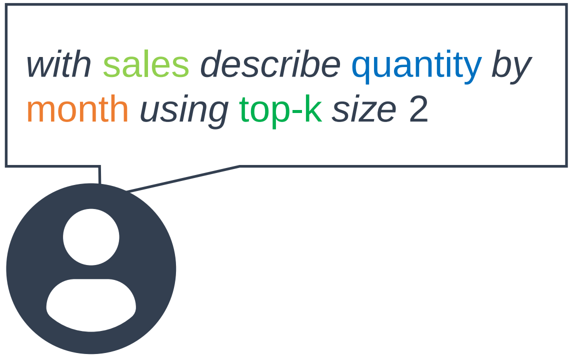
| Month | sum(quantity) |
| --- | --- |
| Jan | 125 |
| Feb | 132 |
| Mar | 12 |
| Apr | 15 |
| May | 50 |

# Intentional OLAP: Describe

`Describe` intention:

*with* cube *describe* $m_1, \ldots, m_z$ [ *for* P ] [ *by* l ] [ *using* $t_1, \ldots, t_r$ ] [ *size* k ]

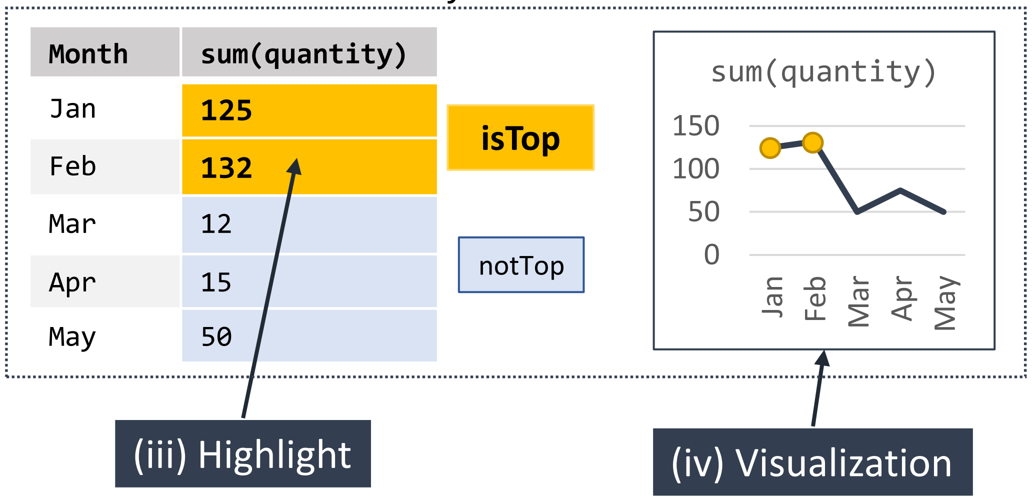
Cube measures: $m_1, \ldots, m_z$
Selection predicate: P
Cube level: l
Models: $t_1, \ldots, t_r$ (apply all models if omitted)
Model size: k (automatically tune k if omitted)




Execution flow

i. Execute query for given cube, measures, predicate, level
ii. Apply models (e.g., clustering, top-k) with given k (e.g., how many clusters)
iii. Estimate interestingness
iv. Pick effective charts

Analytic dashboard



# Model

*with sales describe quantity by month using top-k size 2*

| Month | sum(quantity) |
|---|---|
| Jan | 125 |
| Feb | 132 |
| Mar | 12 |
| Apr | 15 |
| May | 50 |

isTop

notTop

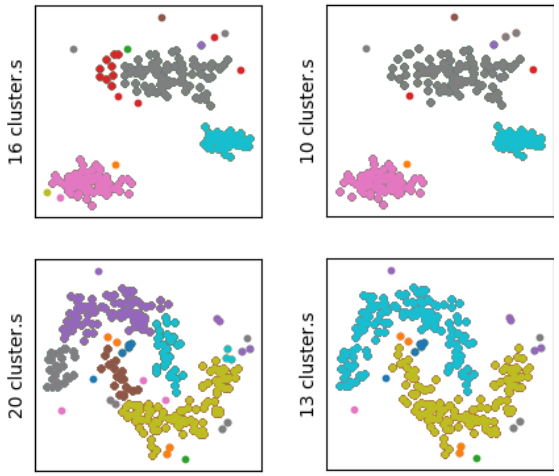
Information-rich patterns

- Computed on levels/measures
- Partition cube cells into *components*
- We consider: Top/Bottom-k, Skyline, Outliers [2], Clustering [3]

Tuning *k*

- Clustering: k clusters
  - Find *knee* in a curve
  - L-method [4]: slower, shift the knee
  - Kneedle [5]: faster, consistent results
- Top/Bottom-k: k points with higher/lower values
- Outliers: k points with higher outlierness

L-method  Kneedle




[2] Liu, F. T., Ting, K. M., & Zhou, Z. H.: Isolation forest. In: *Proc. ICDM.* pp. 413-422 (2008)
[3] S. Lloyd.:, Least squares quantization in PCM. In. IEEE Trans. Inf. Theory, vol. 28, pp. 129-137 (1982)
[4] Salvador, S., Chan, P.: Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In: Proceedings of ICTAI. pp. 576–584 (2004)
[5] Satopaa, V., Albrecht, J.R., Irwin, D.E., Raghavan, B.: Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: Proceedings of ICDCS. pp.166–171 (2011)

# Interestingness

*with sales describe quantity by month using top-k size 2*

| Month | sum(quantity) |
|---|---|
| Jan | **125** |
| Feb | **132** |
| Mar | 12 |
| Apr | 15 |
| May | 50 |

**isTop**

notTop

## Enhance cube with highlights

- Cells in *most-interesting* component

## Interestingness [6]: average component surprise

- Surprise: change in belief before (C) after (C′) applying intention
  - z-score of current cell - avg z-score of *corresponding cells*
- Corresponding cells
  - Intention either changes group-by set/selection predicate
  - If GBS changes, determined via part-of order
  - If P changes, same cell if part of C, whole cube otherwise

*with sales describe quantity by type*

(C)

| type | quantity | *z-score* |
|---|---|---|
| Beer | 116 | -0.6 |
| Bologna | 192 | -0.2 |
| Deli Meats | 211 | -0.1 |
| Fresh Chicken | 64 | -0.9 |
| Wine | 448 | 1.1 |
| … | | |

*with sales describe quantity by category*

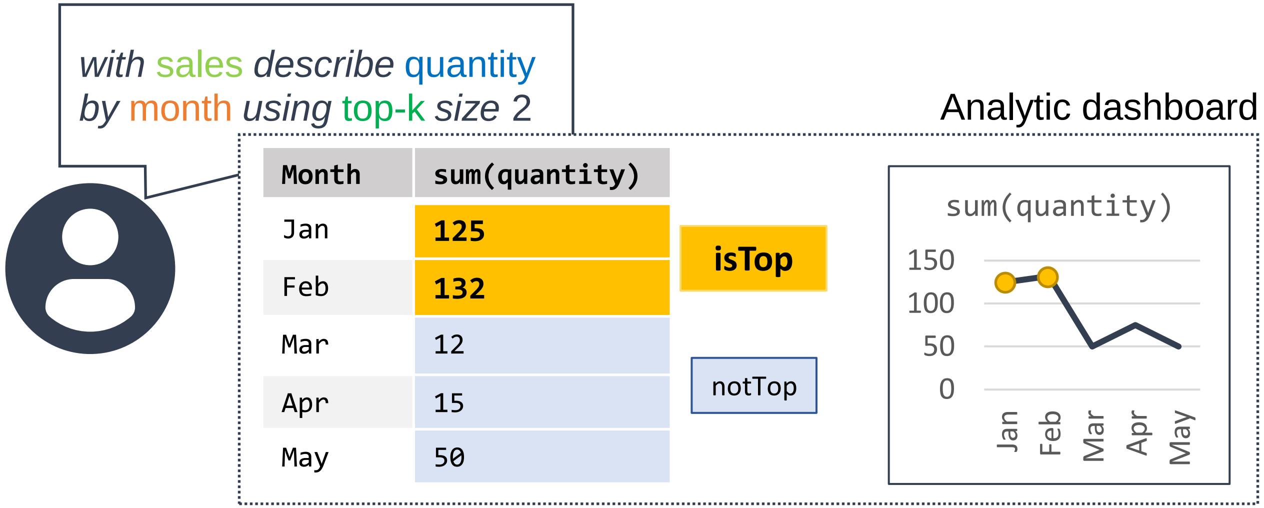
(C′)

| category | quantity | *z-score* | *surprise* |
|---|---|---|---|
| Beer and Wine | 564 | -0.9 | -1.1 |
| Meat | 999 | 1.1 | 1.5 |
| … | | | |

[6] Vassiliadis, P., Marcel, P., Rizzi, S.: Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP. Information Systems 85, 68–91 (2019)

# Visualization

Pivot table and graphical representation
- Color code links data/component




Guidelines:
- Visualizations from multiple points of view
- Visualizations for lay and skilled users

Choose charts depending on:
- Cardinality (e.g., grouped column chart for low cardinality)
- Dimensionality (e.g., scatter plot requires 2 levels)
- Data type (e.g., line chart for time series)

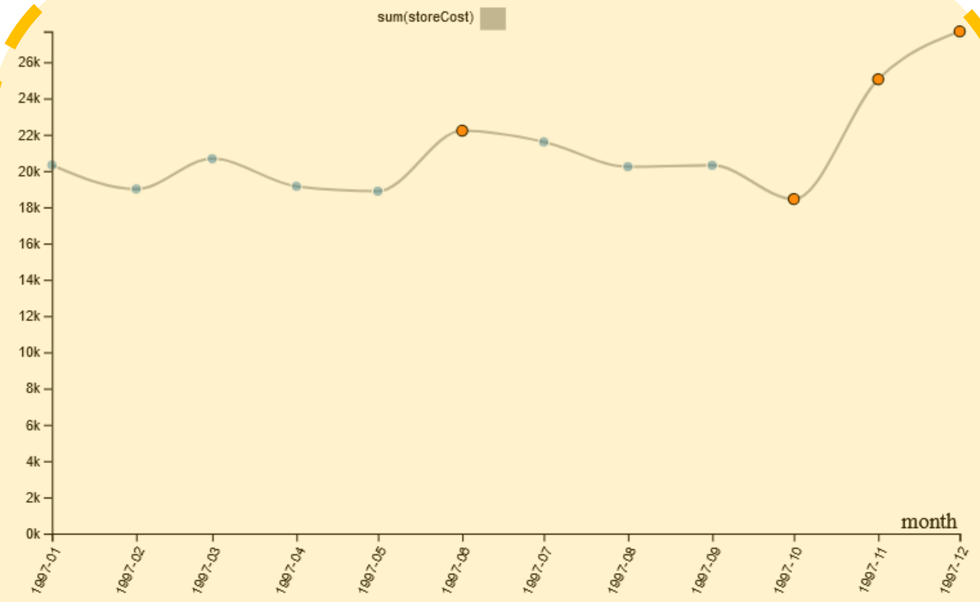
# Describe in action

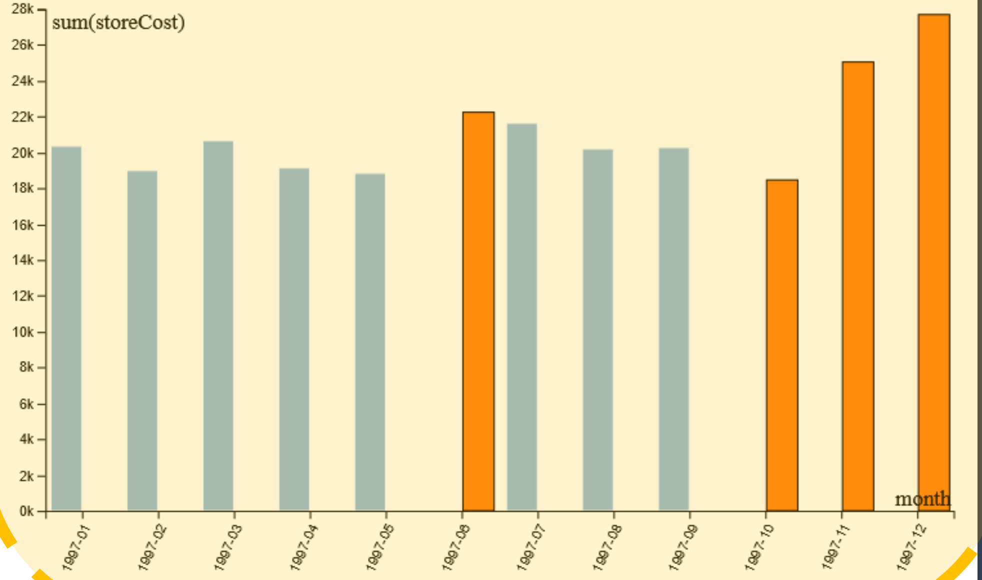
http://semantic.csr.unibo.it/describe/

with Sales describe storeCost by month

| | |
|---|---|
| Measures | sum(storeCost) |
| Rows | month |
| Columns | |
| 1997-01 | 20327.421 |
| 1997-02 | 18999.6688 |
| 1997-03 | 20674.0791 |
| 1997-04 | 19151.8369 |
| 1997-05 | 18880.6687 |
| 1997-06 | 22219.9698 |
| 1997-07 | 21605.791 |
| 1997-08 | 20235.9608 |
| 1997-09 | 20310.8052 |
| 1997-10 | 18449.4667 |
| 1997-11 | 25055.3934 |
| 1997-12 | 27693.7875 |

| component | interest | properties |
|---|---|---|
| cluster 1 | 1.994 | centroid: 26374.59 |
| top sum(storeCost) | 1.467 | avgZscore: 1.47 |
| outliers | 0.845 | outlierness: -1.0 |
| not bottom sum(storeCost) | 0.299 | avgZscore: 0.3 |
| cluster 0 | -0.399 | centroid: 20085.57 |
| not outliers | -0.423 | outlierness: 1.0 |
| not top sum(storeCost) | -0.489 | avgZscore: -0.49 |
| bottom sum(storeCost) | -0.897 | avgZscore: -0.9 |



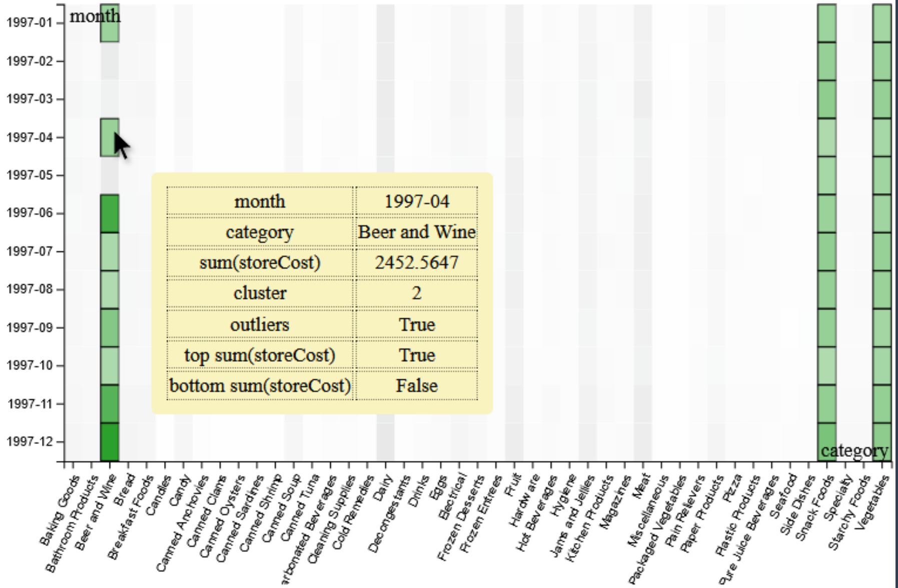
# Describe in action

http://semantic.csr.unibo.it/describe/

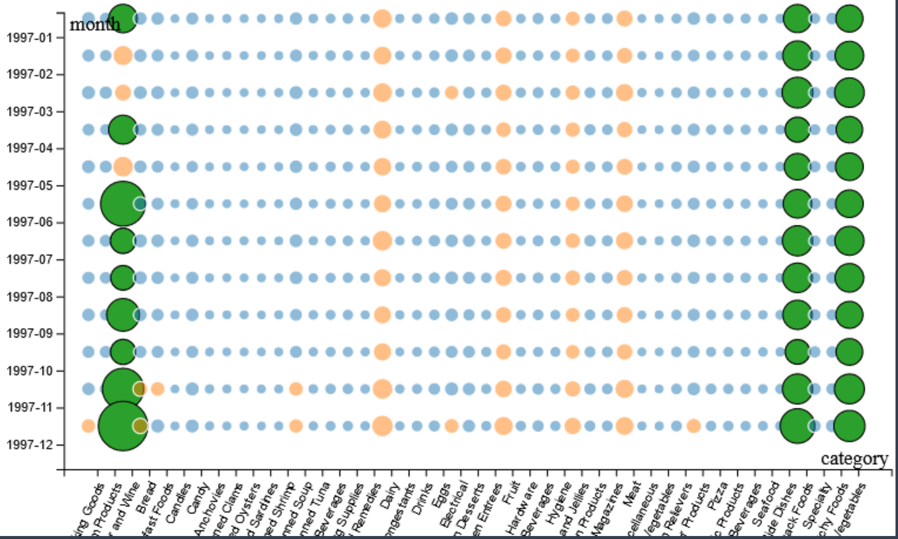
with Sales describe storeCost by category

| | |
|---|---|
| Measures | sum(storeCost) |
| Rows | month |
| Columns | category |

| | Baking Goods | Bathroom Products | Beer and Wine | Bread | Breakfast Foods | Candles | Candy | Canned Anchovies |
|---|---|---|---|---|---|---|---|---|
| 1997-01 | 556.4229 | 421.3061 | 2389.5654 | 520.7016 | 520.7085 | 40.7027 | 464.4884 | 48.8279 |
| 1997-02 | 476.3389 | 354.9201 | 1365.8729 | 519.3708 | 459.4693 | 24.7807 | 458.2083 | 57.2324 |
| 1997-03 | 582.483 | 500.8205 | 968.3671 | 569.7393 | 576.8542 | 49.1184 | 517.756 | 92.5708 |
| 1997-04 | 433.3722 | 344.8056 | 2452.5647 | 446.2024 | 475.4528 | 41.999 | 472.336 | 74.5717 |
| 1997-05 | 567.4029 | 459.0977 | 1441.135 | 506.5075 | 434.8817 | 55.5834 | 380.0482 | 58.3223 |
| 1997-06 | 430.5314 | 445.1673 | 4495.5083 | 500.7004 | 472.7688 | 60.638 | 401.1508 | 87.1755 |
| 1997-07 | 470.786 | 441.0914 | 2011.4115 | 592.2087 | 535.182 | 47.113 | 498.1104 | 73.5137 |
| 1997-08 | 477.8445 | 407.8353 | 1909.6441 | 562.7974 | 514.1978 | 44.5676 | 484.1587 | 64.2519 |
| 1997-09 | 519.8787 | 380.2993 | 2942.8915 | 514.9472 | 464.9874 | 61.7509 | 526.0353 | 80.3315 |
| 1997-10 | 446.2016 | 443.3619 | 1975.2111 | 458.3914 | 457.0872 | 29.1627 | 457.3782 | 61.37 |
| 1997-11 | 508.7303 | 464.7854 | 4045.0889 | 656.4131 | 663.5032 | 36.7406 | 589.9359 | 112.641 |
| 1997-12 | 653.3319 | 533.0684 | 5079.5247 | 716.1121 | 596.7529 | 48.1042 | 577.9768 | 103.0699 |




| component | interest | properties |
|---|---|---|
| cluster 2 | 3.296 | centroid: 2583.35 |
| outliers | 1.425 | outlierness: -1.0 |
| top sum(storeCost) | 1.028 | avgZscore: 1.2 |
| cluster 1 | 0.537 | centroid: 965.33 |
| not bottom sum(storeCost) | 0.179 | avgZscore: 0.21 |
| cluster 0 | -0.315 | centroid: 247.06 |
| not outliers | -0.32 | outlierness: 1.0 |
| not top sum(storeCost) | -0.342 | avgZscore: -0.4 |
| bottom sum(storeCost) | -0.538 | avgZscore: -0.63 |



# Conclusion & research directions

Describe intentional operator

- Extract interesting patterns from cube measures
- Feasibility (Sales from FoodMart, all models computed)

| Cardinality | Query (s) | Model (s) | Interestingness (s) | Pivot (s) | Total (s) |
|---|---|---|---|---|---|
| 323 | 0.88 | 1.45 | 0.03 | 0.03 | 2.39 |
| 77832 | 0.64 | 3.61 | 0.39 | 0.51 | 5.14 |
| **86829** | 0.69 | **3.66** | 0.48 | 1.56 | **6.38** |

Main research directions:

- Scale up to big cubes
  - As in Auto-ML [7], time-budget for optimal model computation
- Assess effectiveness with real data scientists
  - Understand if all meaningful visualizations are covered
  - Estimate how highlights correlate with user's insights
- Extend the approach to other intention operators

[7] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In *Advances in neural information processing systems* (pp. 2962-2970).

Thank you.

# Questions?