From analytic intentions to data stories via highlights

Panos Vassiliadis

Dept. of Computer Science & Engineering

Univ. Ioannina, Hellas



Univ. of Ioannina

Research reported here is funded by

H.F.R.I. Hellenic Foundation for Research & Innovation

DataStory

https://www.cse.uoi.gr/~pvassil/projects/datastory/

Caught somewhere in time



- **Query specific.= properties of desired data**
- **Query result = (just) a set of tuples**

- No difference from the 70's when this assumption was established and tailored for
  - what people had available then
    - … a green/orange monochrome screen
    - … a dot-matrix(?) printer
    - … nothing else
  - users being programmers



Photos copied from http://en.wikipedia.org/

The *DataStory* project

- *Goal: revise the definition of what a query is both with respect to its specification, and with respect to its produced result!*
- *Query spec: Specify Analytical Intentions instead of data specification*

  …via an algebra of operators that allow the user to assess the situation presented by a subset of a data space via comparisons of the retrieved values to similar ones, explain and analyze the demonstrated situation with more detailed data, predict future values etc.
- *Query result: replace the current recordset-as-an-answer query result with a Data Story, i.e., a report that complements the traditional query answer with*
  - *… results of auxiliary queries that contextualize and explain the answer as well as with concise & meaningful highlights and commentaries, packaged with appropriate visualization and textual descriptions that ultimately enhance user insights.*

The research project is implemented in the framework of H.F.R.I. call "3rd Call for H.F.R.I. Research Projects to Support Faculty Members & Researchers" (H.F.R.I. Project Number: 23640)

Website: https://www.cse.uoi.gr/~pvassil/projects/datastory/

Software: github.com/DAINTINESS-Group/DelianCubeEngine

Intentional Analytics Model



I want the system to …

*… DESCRIBE the data in different detail **WHAT** the data tell us!*

*… ASSESS the situation **COMPARED TO** a reference benchmark!*

*… EXPLAIN **WHY** the situation is as it is!*

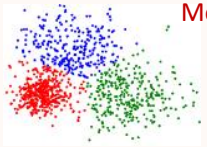
*… PREDICT the **FUTURE STATE** of the situation!*

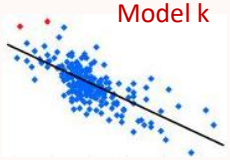
1 Compute (SEVERAL) **relevant data**

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**

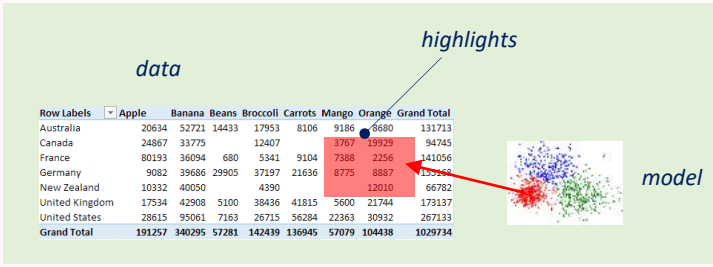
Products	10248	10249	10250	10251	Grand Total
Bagels				6	6
Celery	8				8
Clam Chowder			10		10
Dried Apples	35	24	35		94
Hot Pepper Sauce			15	20	35
Marmalade	11			40	51
Mozzarella		26			26
Ravioli				15	15
Spaghetti	10				10
Tofu	5	9			14
Grand Total	**69**	**59**	**60**	**81**	**269**

2 Construct **models**



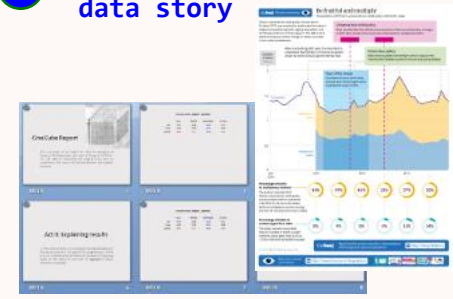


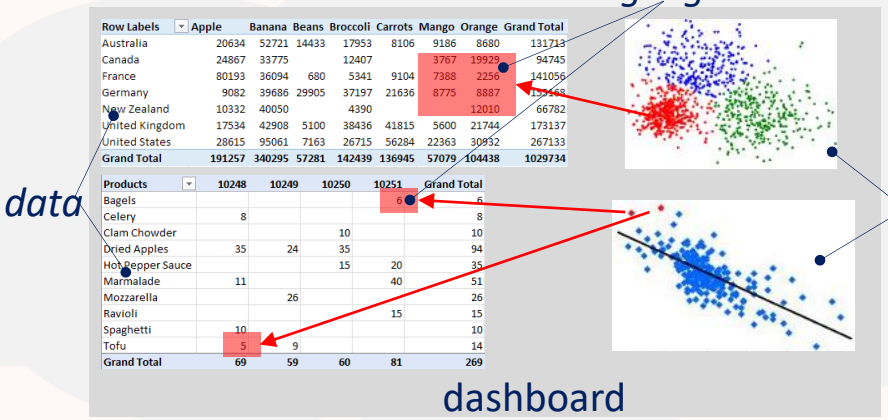
3 Compute significant highlights



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)

4 Synthesize a **data story**





NEW answers include…

Data

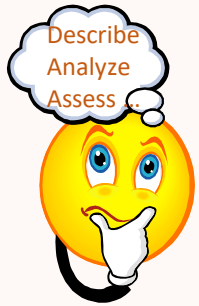
Models

Highlights

ASK *Why Data & Models*
(not what)

SELECT *important Highlights*

GET *a Data Story*
(not tuples)



ALGEBRA OF INTENTIONAL OPERATORS

ORIGINAL + AUXILIARY QUERIES

ML Models

USER PROFILE

PAST MODELS

Each Intentional Operator generates **auxiliary queries** and passes all query results through **ML models**

HIGHLIGHTS

SCORED, PRUNED, STRUCTURED FINDINGS

INTERESTINGNESS MEASURES

Novelty, relevance, Surprise, …

Highlights, produced in the previous stage, are scored according to **interestingness measures**, pruned and structured as **findings**

Story

Acts

Episodes

STRUCTURED STORY

ANNOTATED (text, video, …) DATA STORY

PNG

EXPORTED RESULTS

Findings structure a **story** of **episodes & acts**, which is visualized, textually commented, and packaged as a **Data Story**, exportable in various formats

Concepts

Software Architecture

NL

SQL-like

CLI

GUI

API

QUERY ENGINE API

OPTIMIZER

DBMS

ML EXECUTION ENGINE

HIGHLIGHTS

CACHING

RESTRUCTURING

RANKING

SCORING

PRUNING

MAJOR FINDINGS

Story

Acts

Episodes

TEXT WRITER

VISUALIZER

STORY CONSTRUCTOR

DISCOURSE CONSTRUCTOR

PNG

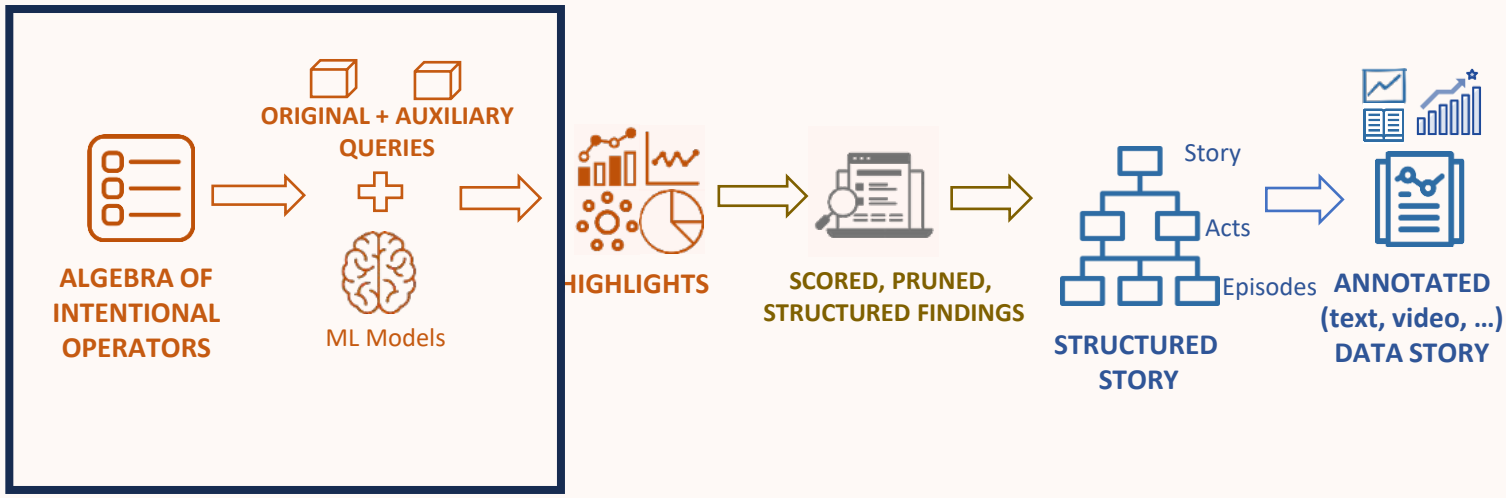
EXPORTER

DataStory

QUERY ENGINE

HIGHLIGHTS SELECTOR

STORYTELLER



Intentional Operators

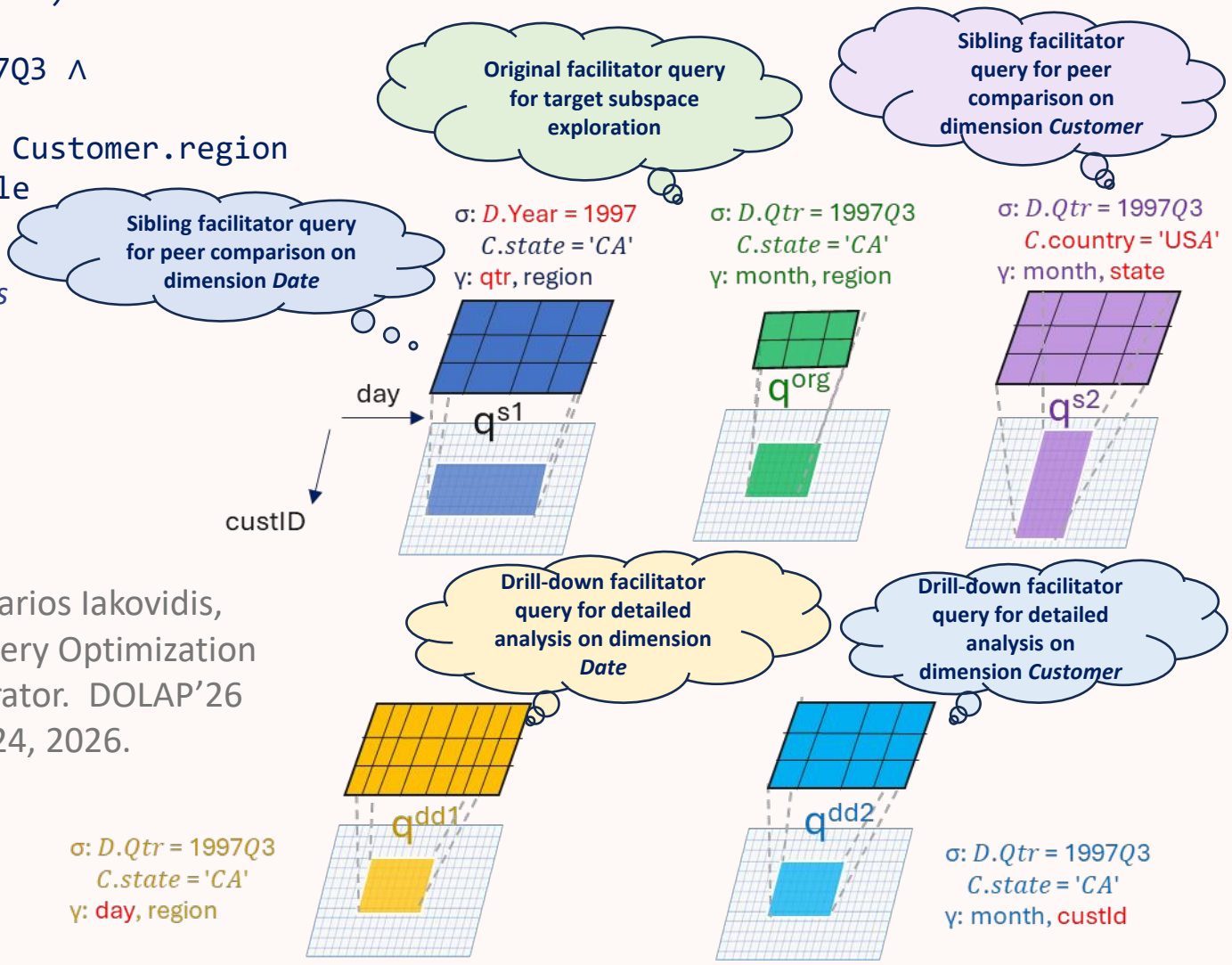
How to query intentionally?

Two examples…

Intentional operators: ANALYZE

```
ANALYZE sum(Store_Sales) as SumSales
FROM Sales
FOR Date.quarter=1997Q3 ∧
Customer.state=CA
GROUP BY Date.month, Customer.region
AS AnalyzeQueryExample
```

Concrete example: store sales in CA, Q3-1997, grouped by month × region

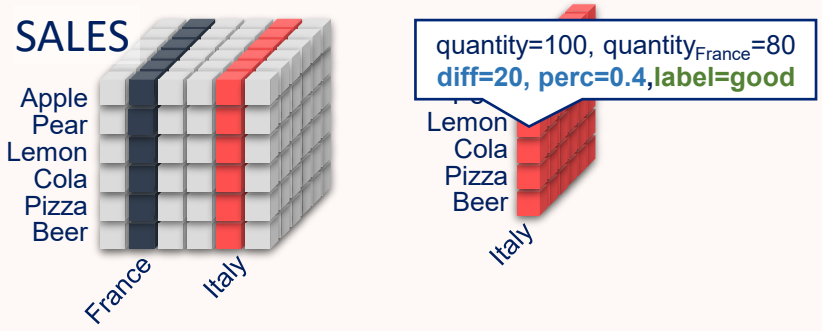


The ANALYZE Operator: Marios Iakovidis, Panos Vassiliadis. Multi-Query Optimization for the Novel Analyze Operator. DOLAP'26 Tampere, Finland - March 24, 2026.

Intentional operators: ASSESS

The ASSESS Operator: Matteo Francia, Matteo Golfarelli, Patrick Marcel, Stefano Rizzi, Panos Vassiliadis. Assess Queries for Interactive Analysis of Data Cubes. 24th International Conference on Extending Database Technology (EDBT 2021)

```
with SALES for country = 'Italy'
by product, country
assess quantity against country = 'France'
using percOfTotal(difference(quantity,
benchmark.quantity))
labels {[-inf, -0.2): bad, [-0.2,0.2]: ok,
(0.2, inf ]: good}
```





Highlights

What happens once we have the data from all these operators?

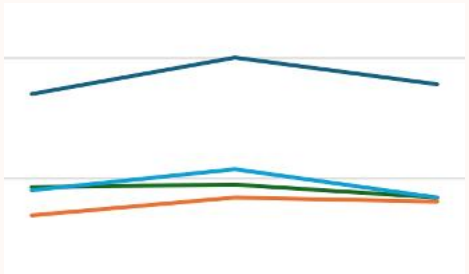
We compute properties of the data!

Example

Month	Athens	Rhodes	Chania	Thera	Total
April 2023	500	50	85	80	715
May 2023	1000	70	90	120	1280
June 2023	600	65	70	70	805
Total	2100	185	245	270	**2800**

Query:
Total sales for
- the product Wine
- the 2nd quarter of 2023
- the country Greece,

grouped by month and city



- Highlights:
  - The city Athens dominates all other cities
  - The month May 2023 dominates all other months
  - The city of Athens is a mega-contributor to the total sales (75%)
  - No trend in time-series but unimodality (a peak in May)

Can we uniformly model all these highlights?

Model

Panos Vassiliadis, Veronika Peralta, Patrick Marcel, Dimos Gkitsakis, Angeliki Dougia, Faten El Outa. **A Conceptual Model for Data Analysis Highlights.** ER Forum 2025, Poitiers, France, 20-23 October 2025.

- **Archetype Property**: a specific pattern of internal relationships

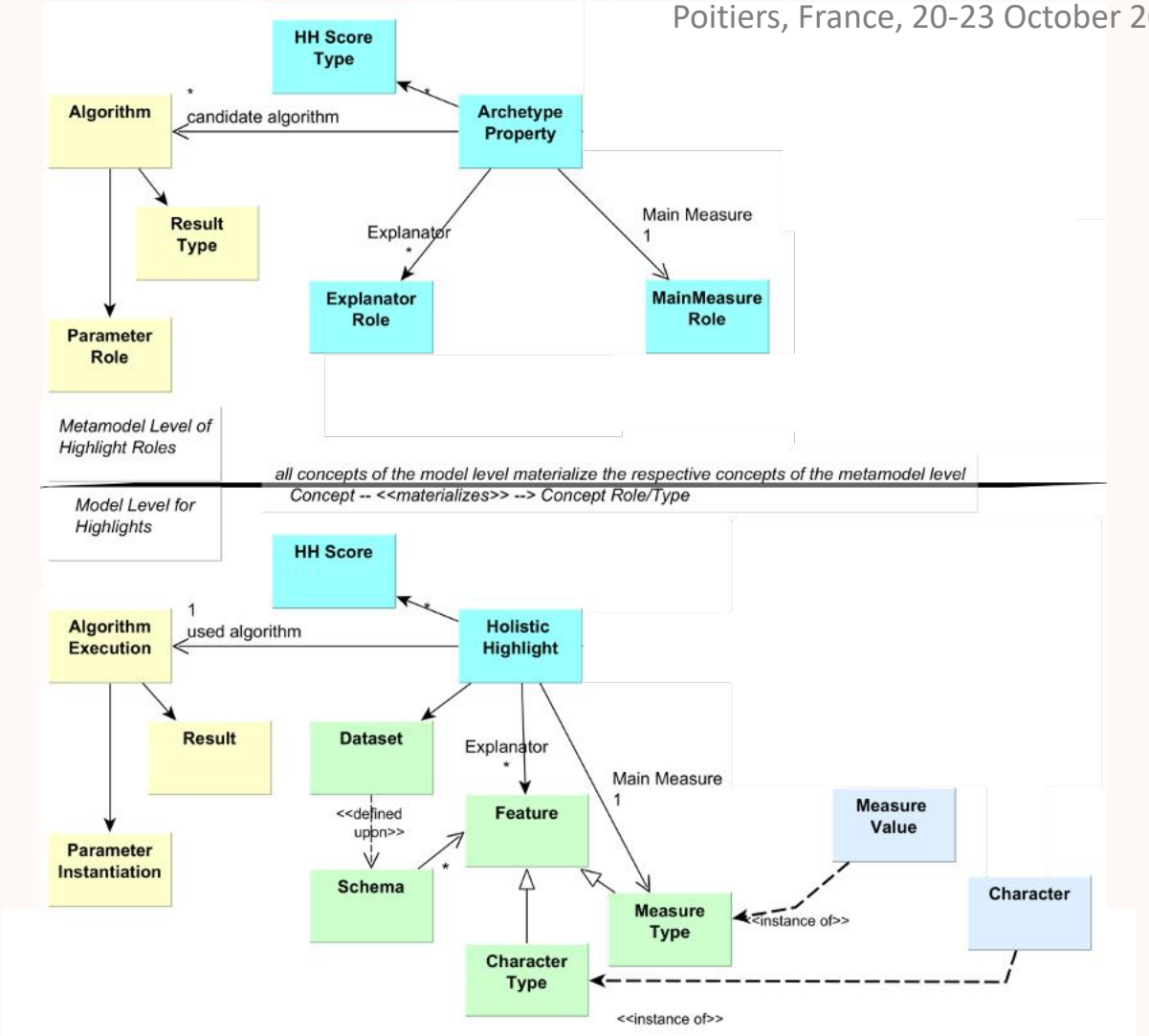
  Example:
  - Monotonous trend
  - Outlier Existence

- **Holistic Highlight:** materializes an Archetype Property over a specific dataset

Examples: the existence of
- a trend in a timeseries
- the correlation of two Measures
- a mega contributor

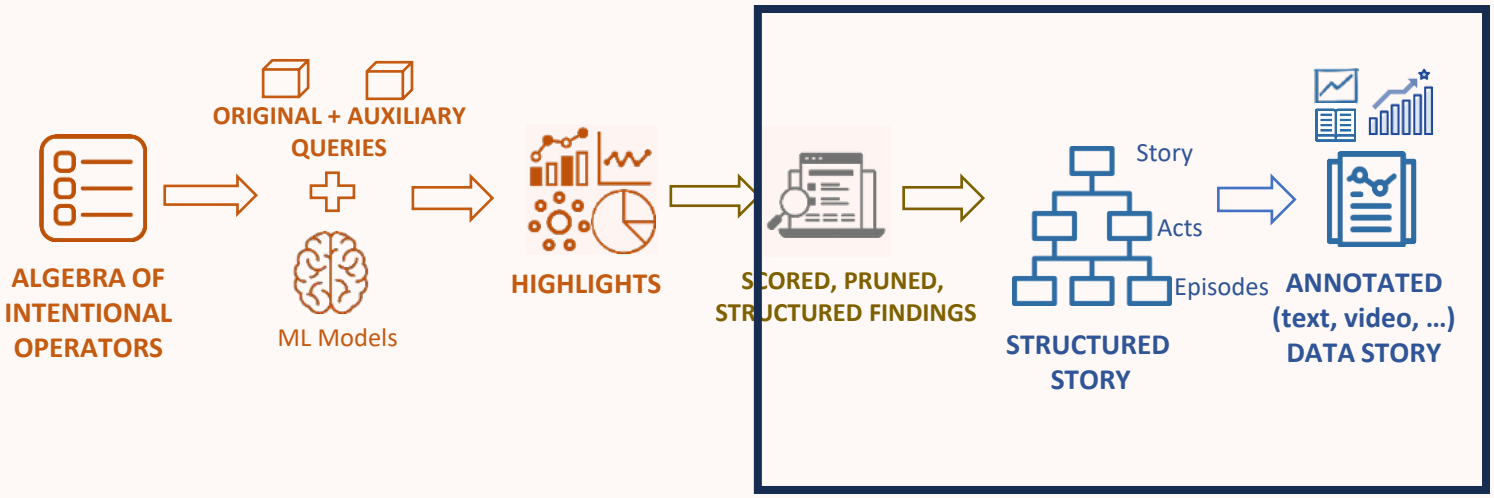
- **Elementary Highlight:** a fact which plays an important role to the interpretation of a Holistic Highlight

Example: the time point where a time series reaches a peak



At the Metamodel Level						At the Model Level	
The **Archetype Property**	to be tested over **Main Measure Role** *<constraint>*	wrt **Explanator**<supporting txt> *<constraint>*	via candidate **Algorithms** *(parameters omitted)*	having **Result Type**	and **Elementary Highlight Roles**	for measure **Main Measure**	wrt **Explanator**
Normal Distribution	*Any measure*	-- --	Shapiro-Wilk, Kolm.-Smirnov	- {W, D} Statistic - p-value		`SumSales`	
Correlation	*Any measure*	"wrt measure" *Any measure*	Pears., Spear., Kendall	- {r,ρ,τ}, - p-v, df, 95%Conf.		`SumSales`	`cityPopulation`
Regression formula	*Any measure*	"wrt feature" *Any Feature*	SimpleRegr	- Intercept, Slope - R2, p-value, MSE		`SumSales`	`cityPopulation, month`
Top-k values	*Any measure*	*Any measure*	topK check	k	Top-k points, i=1..k	`SumSales`	
Trend, Seasonality	*Any measure*	"over" *Time-related Char.Type*	STL	- {season, trend, rem.} - #iterations		`SumSales`	`month`
Uni(bi) modality	*Any measure*	"over" *Time-related Char.Type*	DiffCheck, BoolCheck, …	Exists?	Point(s) of peak(s)	`SumSales`	`month`
Peer-domination	*Additive*	"over" *{Char.Type}**	Peer Dom Check	Exists?	Dominant Characters	`SumSales`	`city, month`
Total sum Mega-Contribution	*Additive*	"over" *{Char.Type}**	Mega Contrib	Exists?	Mega-contributor fact	`SumSales`	`city, month`
Surprise	*Any measure*	"over" *{Char.Type}**	3PastYears Extrapolator	Exists?	Surprising fact	`SumSales`	`city, month`

The **Character set**	*over* ***character types***	With a **Measure value**	For **measure type**	serves as the **Highlight Type**	with a **Score Type**	having **Score**
<April 2023>	Month	68	SumSales	1st peak of unimodality	*Peak Rank*	1
<June 2023>	Month	62	SumSales	2nd peak of bimodality	*Peak Rank*	2
<Athens, June 2023>	City, Month	54	SumSales	Top-k point	*Rank*	2
<Athens>	City	54	SumSales	peer-dominator over its sibling characters	*Pct of dominated peers*	100%
<June 2023>	Month	32	SumSales	peer-dominator over its sibling characters	*Pct of dominated peers*	80%
<June 2023, Athens>	City, Month	12	SumSales	a mega-contributor of the total sum	*Pct of total sum*	25%
<Athens>	City	2100	SumSales	a mega-contributor of the marginal sum of cities	*Pct of total sum*	75%
<April 2023, Rhodes>	City, Month	10	SumSales	Surprising cell	*Abs Diff Ratio*	28%



Data Storytelling

What the analyst finally tells other people about his findings.

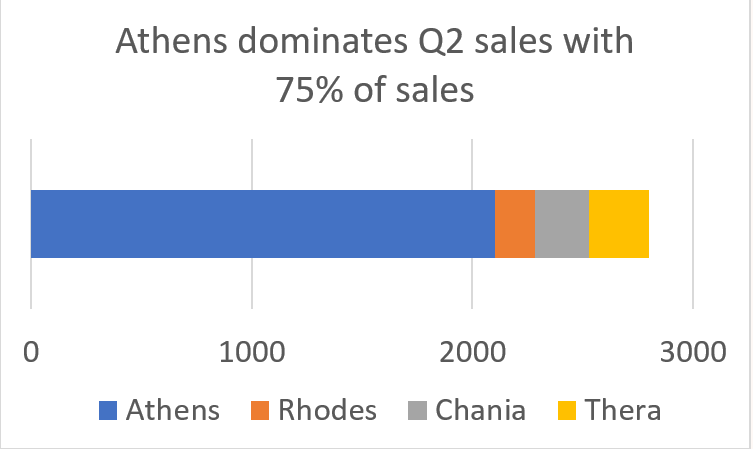
Data stories

- *A **Data Story** is a structured presentation of the analyst's **annotated messages***
- *An annotated message is the visual apodosis of an **info-message**, i.e., an (ideally structured) expression of analytical findings, combining relevant data and selected highlights*
- ***Annotated messages** apply visualizations, and textual explanations to communicate the essence of an info message to the targeted audience.*
- ***Data stories** combine annotated messages in a structured fashion, depending on the context (e.g., written report, infographic, comic, …) and the content.*

2023 Q2 with a pronounced peak

In terms of geography, Athens dominates all other cities, in every month. In fact, Athens is a mega-contributor to total sales, by contributing 75% of all sales. In terms of time, the progression in time shows a peak in May; in fact, May dominates all other months in terms of total sales. No trend or seasonality were detected.





Month	*Athens*	*Rhodes*	*Chania*	*Thera*	*Total*
April 2023	500	50	85	80	*715*
May 2023	1000	70	90	120	*1280*
June 2023	600	65	70	70	*805*
Total	*2100*	*185*	*245*	*270*	***2800***

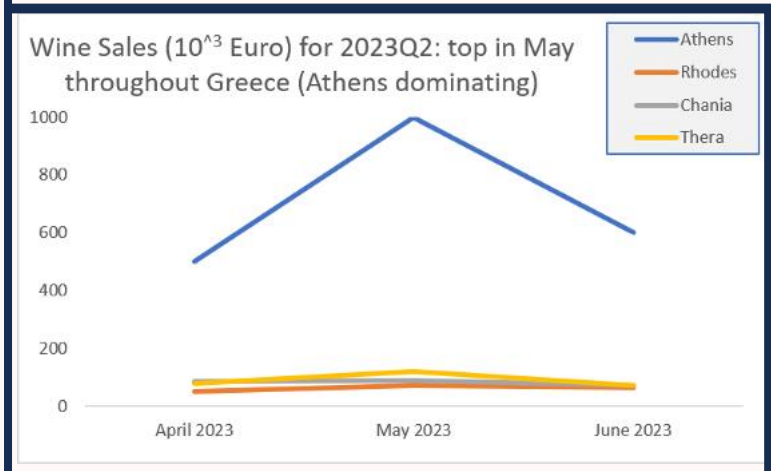
May dominates all other months, with a pronounced peak in Athens

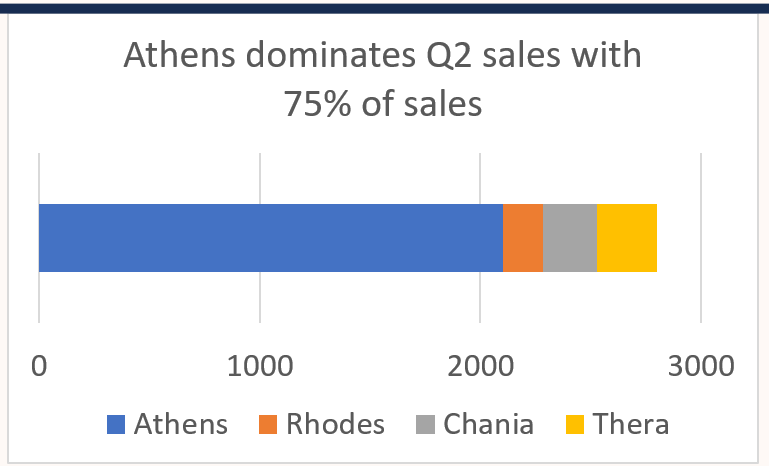
Structure of a story

2023 Q2 with a pronounced peak



In terms of geography, Athens dominates all other cities, in every month. In fact, Athens is a mega-contributor to total sales, by contributing 75% of all sales. In terms of time, the progression in time shows a peak in May; in fact, May dominates all other months in terms of total sales. No trend or seasonality were detected.





Month	Athens	Rhodes	Chania	Thera	Total
April 2023	500	50	85	80	715
May 2023	1000	70	90	120	1280
June 2023	600	65	70	70	805
Total	2100	185	245	270	**2800**

May dominates all other months, with a pronounced peak in Athens

Act III with 1 episode: summary

Act II with 2 episodes
- Athens mega-contr.
- Chart of data

Act I with 3 episodes
- Data
- May
- Outlier

Linking to the main theme…

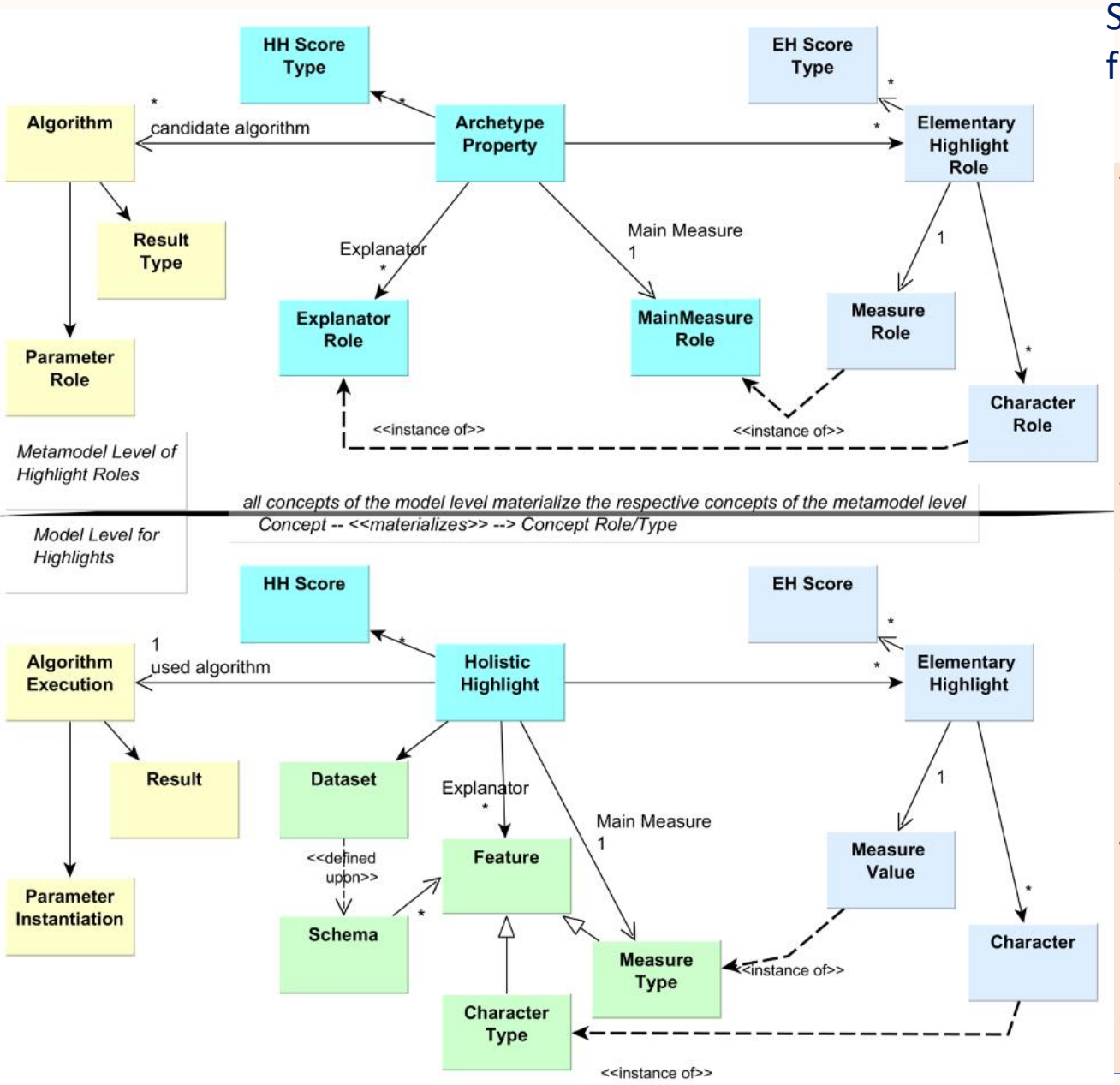
Semantic foundations…

Novel paradigm of services …

… for intentional data storytelling



Storytelling with highlights



Story fragment

Who are these characters?

	<April 2023>	<June 2023>
The **Character set** *over character types* with a **Measure value**	Month 68	Month 62
For **measure type**	SumSales	SumSales
serves as the **Elementary Highlight Role**	1st peak	2nd peak
with **EH Score Type**	*Peak Rank Significance*	*Peak Rank Significance*
having **EH Score**	80%	50%
via <**formula**>	*htan(measureValue / avgValue)*	

The need for a reference set of concepts and story characters

- To be able both
  - to query
  - to tell a story
- … you need a reference set of characters for your highlights. Different complementary views:
  - **Enterprise Ontology** at the meta level: business meaning, entities, relationships, constraints
  - **Cube Dimensional Model** at the cube level: hierarchies, measures, members
  - **Dimension Bus** at the database level: consistent, conformed data

Enterprise Ontologies for enhanced meaning

- Analytical intentions such as explanation and prediction require understanding what the data represents and how its elements relate to one another.
- Research opportunities: semantic models and ontologies could provide the business context needed to **interpret** results, **score** highlights, and **narrate** meaningful data stories.

Towards a new paradigm of services with…



A richer input abstraction: Must move from specifying which records to retrieve toward specifying what analytical objective should be pursued.

A richer data retrieval mechanism: Must move from solo-query execution to multi-query analytical workflows & model execution

A richer finding production: Must move from simply returning tuples or their aggregates to producing highlights, scored by significance.

A richer output production: Move from tuples and aggregates toward structured analytical answers with data, models, highlights, and narratives.

DataStory

We envision an intelligent analytical (DBMS-external) infrastructure that transforms high-level intentions into evidence-grounded findings expressed as significant highlights that are communicated via structured data stories.

Many thanks to our organizers!



The *DataStory* project

Website: https://www.cse.uoi.gr/~pvassil/projects/datastory/

Software: github.com/DAINTINESS-Group/DelianCubeEngine

The research project is implemented in the framework of H.F.R.I. call "3rd Call for H.F.R.I. Research Projects to Support Faculty Members & Researchers" (H.F.R.I. Project Number: 23640)



H.F.R.I.
Hellenic Foundation for Research & Innovation