SJK006 - Máster U. en Sistemas Inteligentes

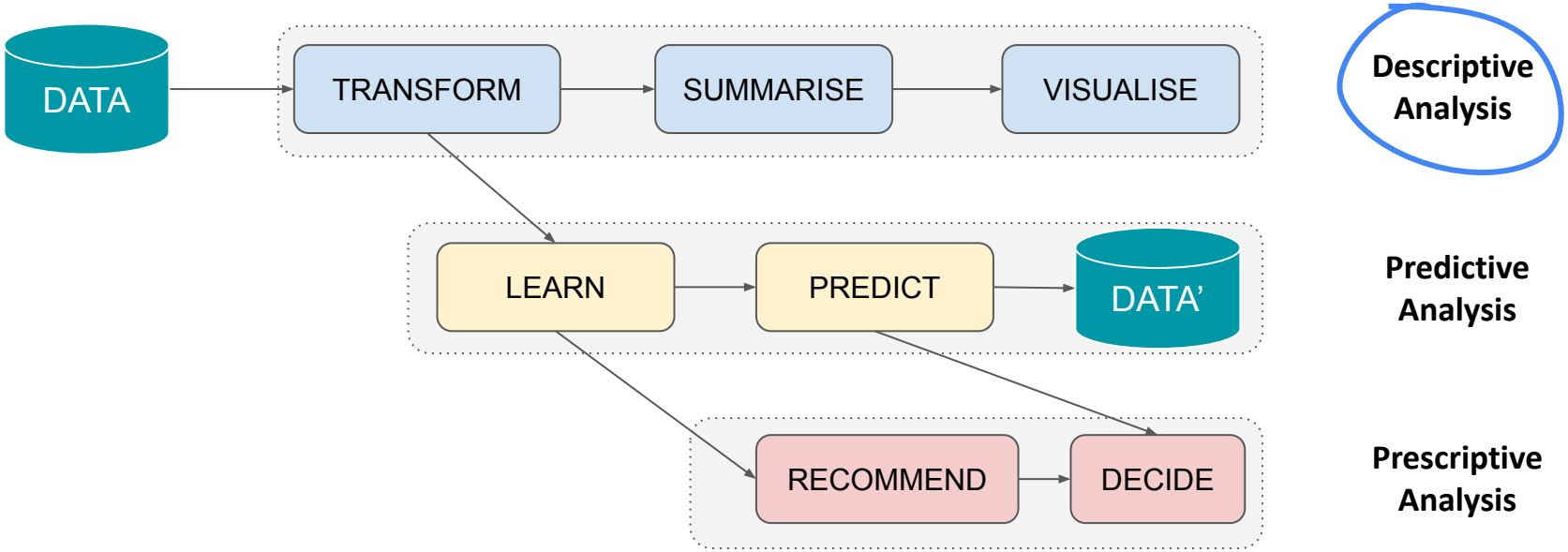
Descriptive Analysis

The multidimensional model

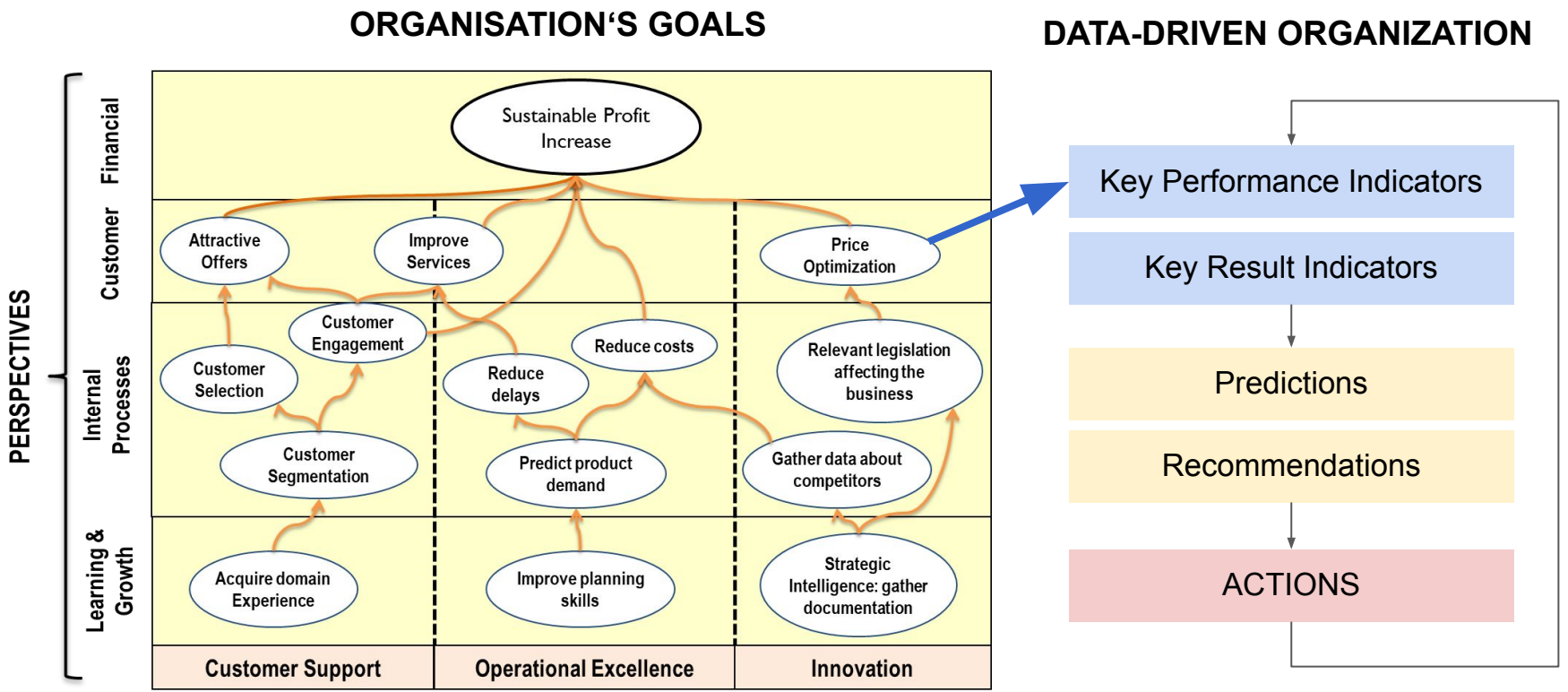
We will learn

- The relation of strategic goals, indicators and facts
- To transform data to analysis facts
- To obtain analytical cubes from facts
- To apply OLAP with Python and Pandas

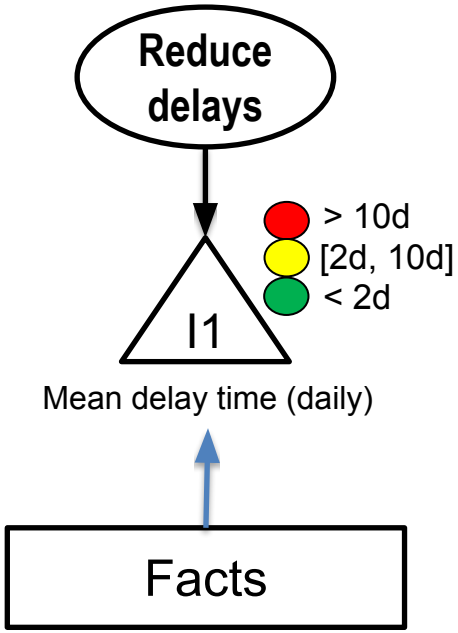
Reminder: DATA VALUE CHAINS



Reminder: Balanced Scorecard



Goals, indicators, facts



An indicator monitors the degree of achievement of a goal.

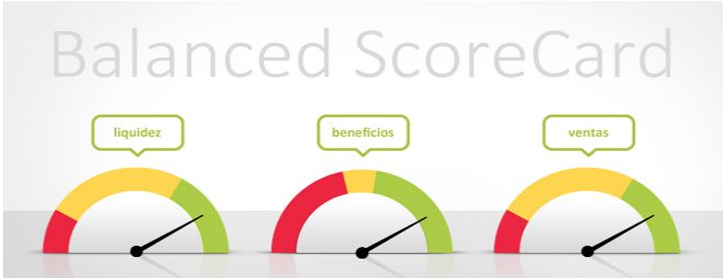
It is expressed as a ratio, a percentage, a proportion, etc.

It is preferably normalized.

It is always TEMPORAL (to monitor its evolution)

Qualitative values (traffic lights) are set by defining a series of reference values (thresholds).

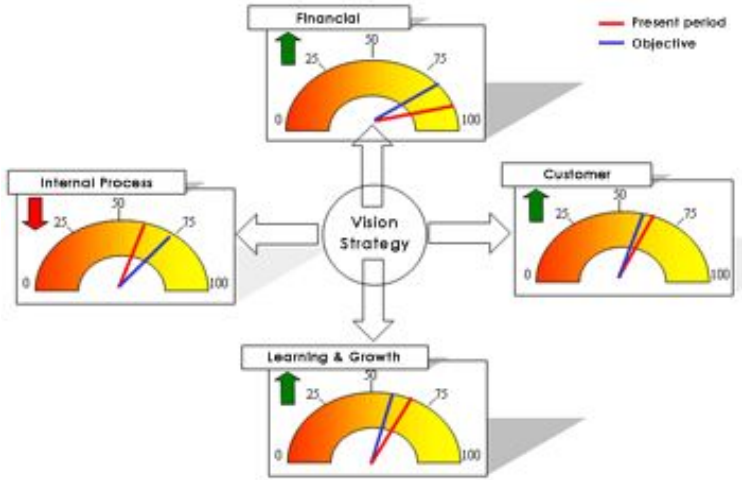
Visual tools: Control Panels



https://www.intrafocus.com/

Scorecard (W) More

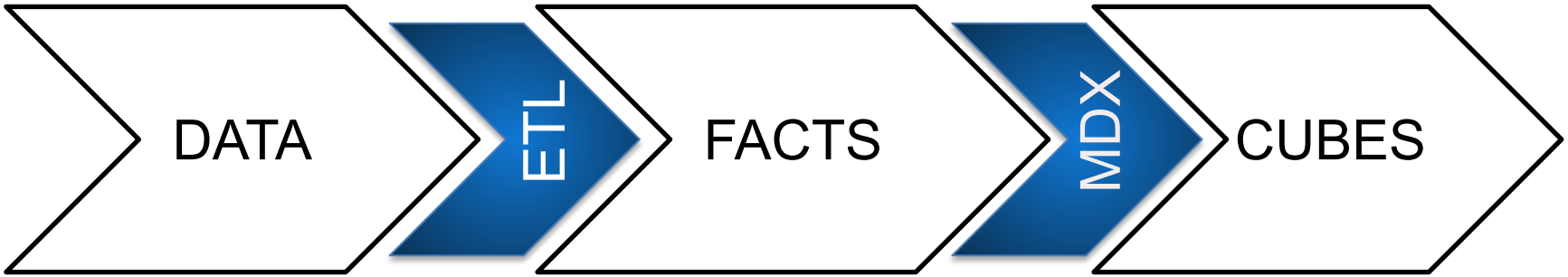
Name	Latest	Trend	% Chg	Tgt	% Tgt
Number Of Sales Number Of Sales	£27 May-12		-4 %	£23	17 %
Spend Advertising & Leads	£6,373 May-12		-2 %	£6,600	-3 %
Invoiced Financials	£49,001.00 May-12		-4 %	£50,000.00	-2 %



Key Performance Indicator (KPI)
- Definition of strategic measures
- Definition of goals and brands
- Continuous monitoring

Visual tools: Dashboards

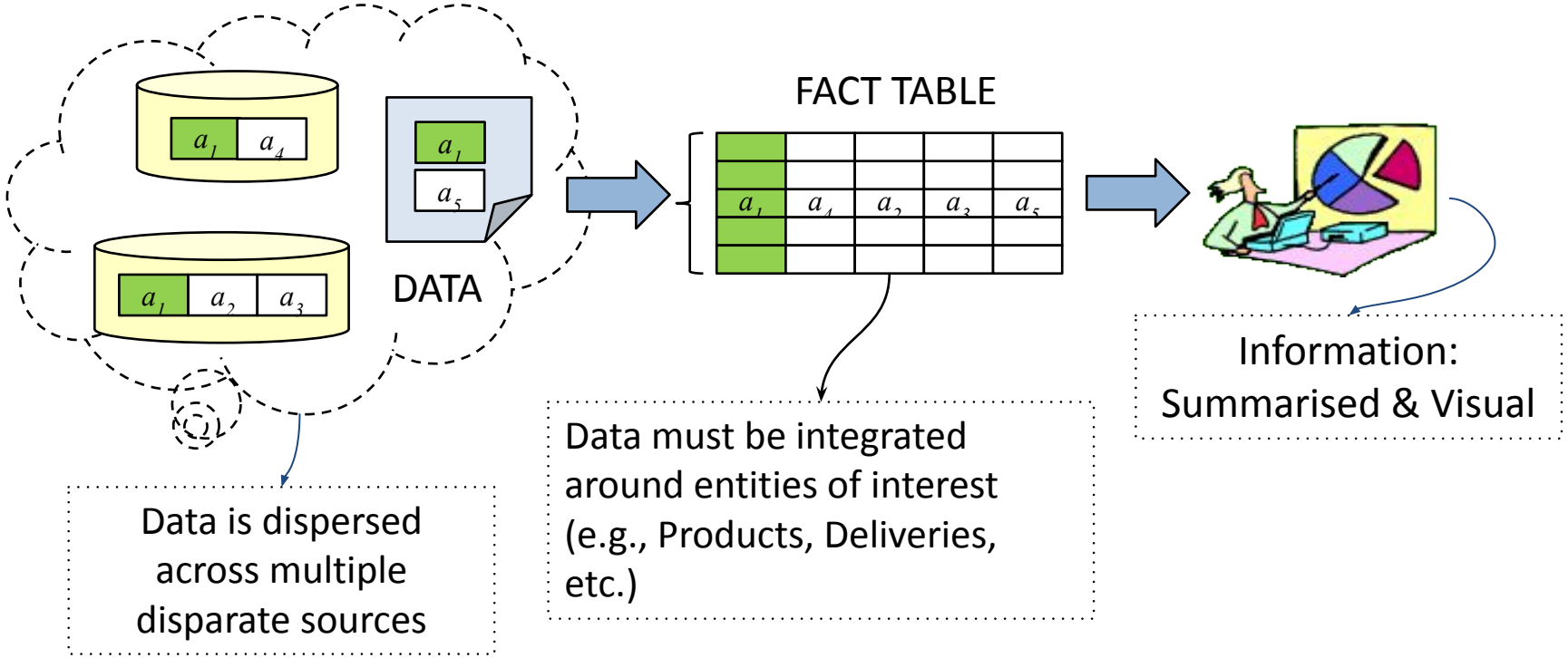
Multidimensional Analysis Process



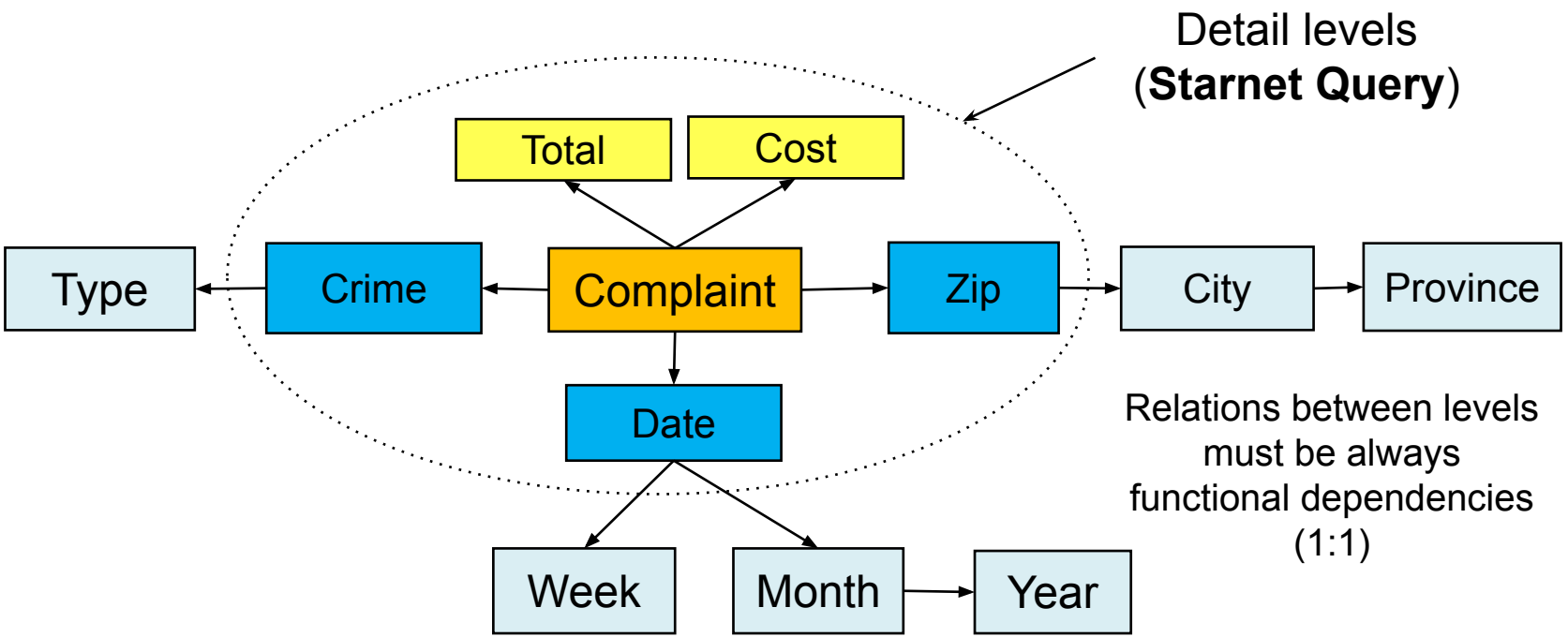
Extract, Transform & Load

OLAP
OnLine Analytical Processing

Data Value Chain for Descriptive Analysis



FACTS = DIMENSIONES + MEASURES



FACT TABLES

https://graphic-walker.kanaries.net/

DIMENSIONS MEASURES

FACTS

Complaint	Type	Zip	City	Prov.	Date	Cost	Total
theft	T1	12003	Castellón	CS	12/10/2019	1520	5
vandalism	T2	12004	Castellón	CS	12/10/2019	3400	1
assault	T1	12005	Almassora	CS	14/10/2019	12000	1
theft	T1	12003	Castellón	CS	12/10/2019	30	1
burglary	T1	12023	Castellón	CS	13/10/2019	105	3

Dimensions compound the primary key of the table (all details)

Aggregation?

CUBES: arranging facts into a multidimensional model

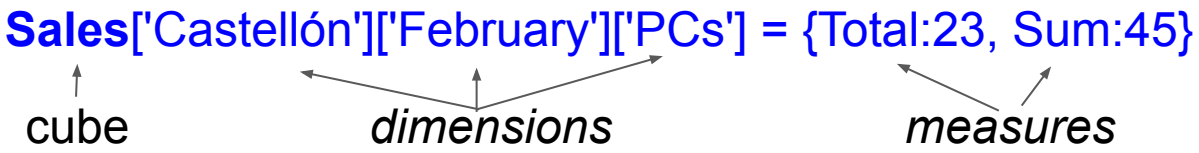
The multidimensional model is based on the data structure called **CUBE**.

The cells of the cube are accessed by a series of dimensions which:

- must be orthogonal (non dependent)
- can have associated different levels of detail

Each cube cell contains the observed measures for its dimension values.

It works similar to a multidimensional dictionary:



Mesures

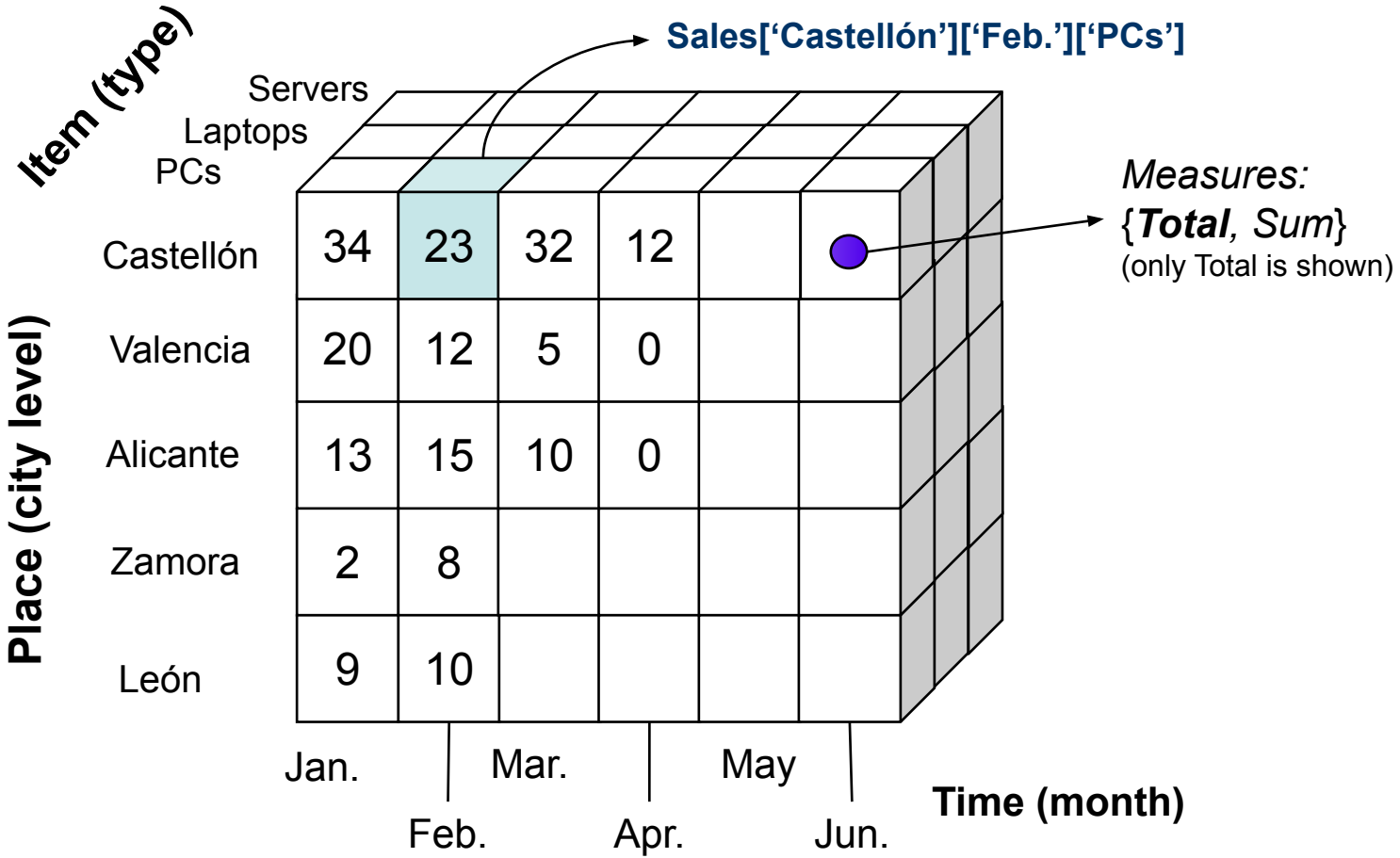
A measure is a **numerical function** that can be evaluated in each cell of the cube.

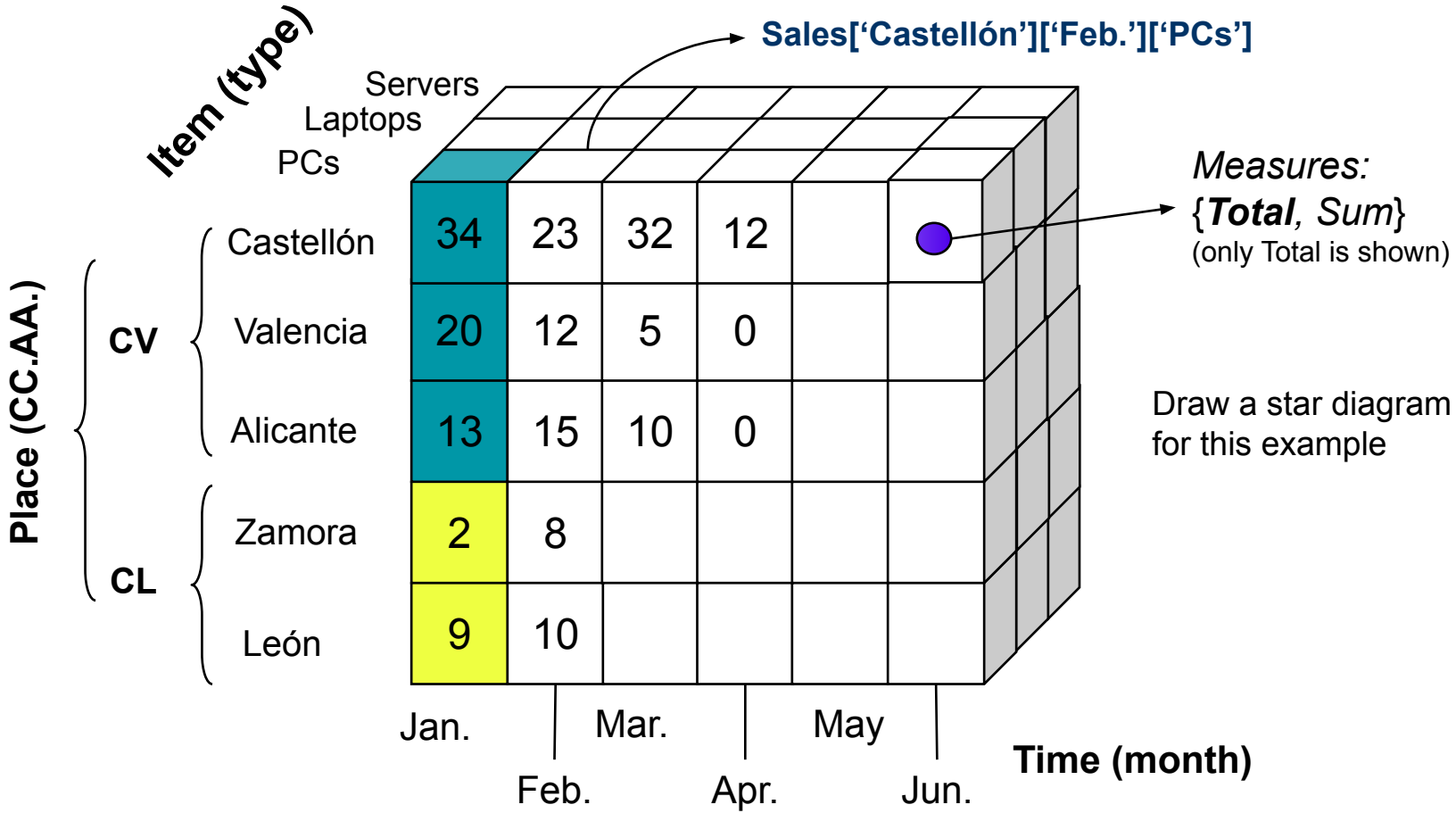
The measure value of a cell is computed by **aggregating** the cells contained in it.

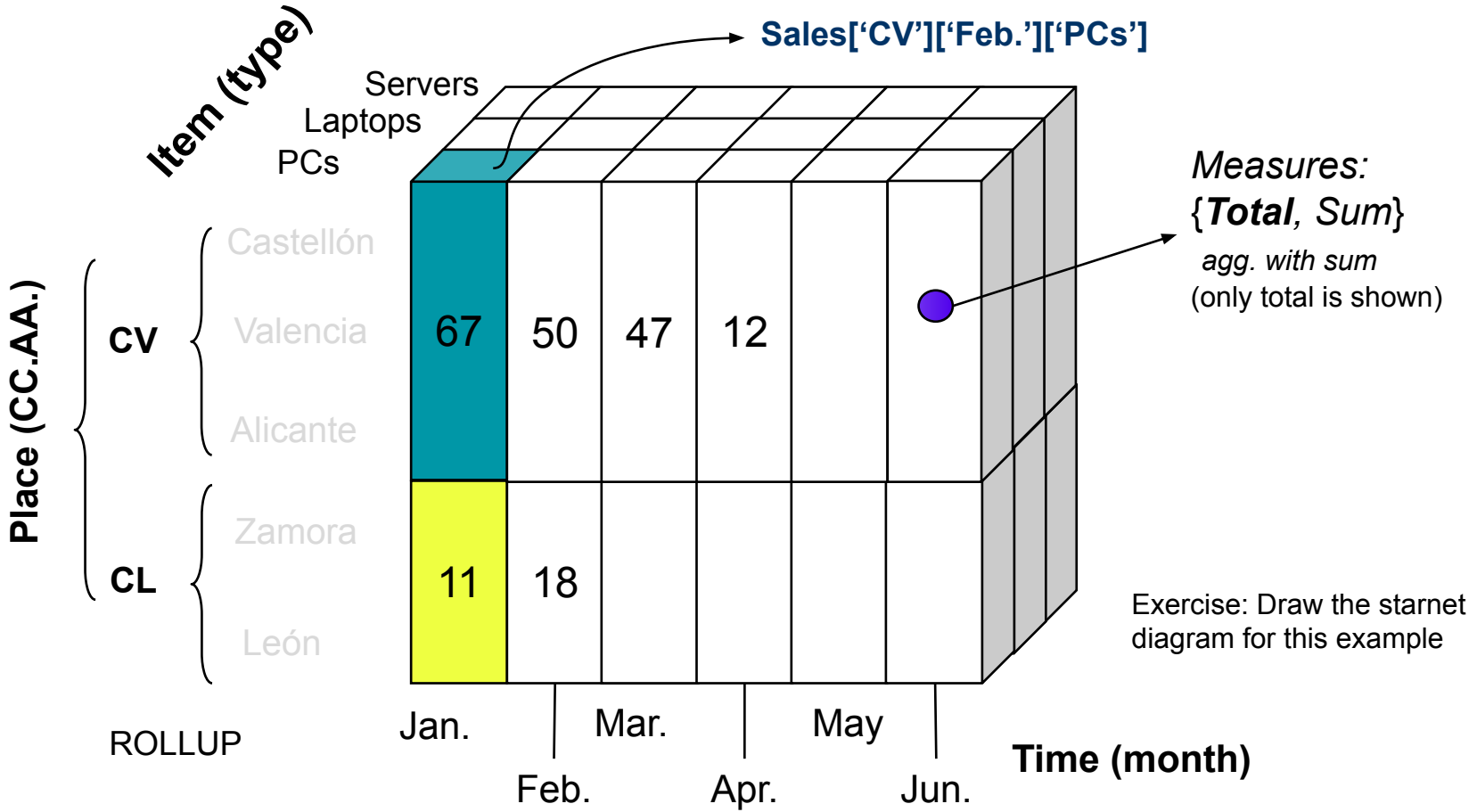
Types of mesure functions:

- **Distributive**: We can directly aggregate the sub-cells.
  count(), sum(), min(), max()
- **Algebraic**: we need to derive the measure combining distributive measures, for example avg() = sum()/count(), range() = max() – min(), standard_deviation()...
- **Holistic**: other measures that cannot be aggregated from partial results
  median(), mode(), etc.

Exercise: Build an algebraic aggregator for the standard deviation, which can be expressed as $\sqrt{E(x^2) - (E(x))^2}$


Sales['Castellón']['Feb.']['PCs']
Item (type)
Servers
Laptops
PCs
Place (city level)
Castellón
Valencia
Alicante
Zamora
León
34 23 32 12
20 12 5 0
13 15 10 0
2 8
9 10
Measures:
{Total, Sum}
(only Total is shown)
Jan.
Feb.
Mar.
Apr.
May
Jun.
Time (month)


Item (type)
Servers
Laptops
PCs
Sales['Castellón']['Feb.']['PCs']
Measures:
{Total, Sum}
(only Total is shown)
Draw a star diagram
for this example
Place (CC.AA.)
CV
CL
Castellón
Valencia
Alicante
Zamora
León
34
23
32
12
20
12
5
0
13
15
10
0
2
8
9
10
Jan.
Feb.
Mar.
Apr.
May
Jun.
Time (month)


Item (type)
Servers
Laptops
PCs
Sales['CV']['Feb.']['PCs']
Measures:
{Total, Sum}
agg. with sum
(only total is shown)
Place (CC.AA.)
CV
Castellón
Valencia
Alicante
CL
Zamora
León
67
50
47
12
11
18
ROLLUP
Jan.
Feb.
Mar.
Apr.
May
Jun.
Time (month)
Exercise: Draw the starnet diagram for this example

OnLine Analytical Processing (OLAP)

OLAP defines the following basic operations on cubes:

- **ROLL UP**(Cube, Dim.) → Reduce details
- **DRILL DOWN**(Cube, Dim.)
- **ROLL ACROSS**(Cube, Dim.)
- **DRILL ACROSS**(Cube, Dim.)
- **SLICE**(Cube, Dim., Level)
- **DICE**(Cube, D1, L1, D2, L2, …)
- **PIVOT**(Cube, Rows, Columns)

All operators return a new cube with some change: reduce details, augment details, change axis, etc.

These operations are applied interactively by using data visualization tools.

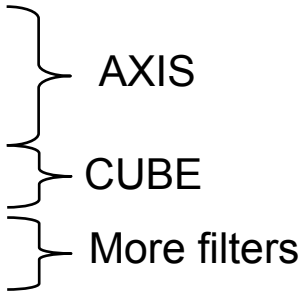
TABLEAU & POWER BI

It allows you to explore and define indicators, as well as to detect anomalies and interesting points for decision making.

Test: https://graphic-walker.kanaries.net/

MDX: the OLAP query language

```
SELECT
{ [Measures].[total_sales]}  ON COLUMNS
{ [Time].[Q1], [Time].[Q2]}  ON ROWS
FROM Sales
WHERE ([Item].[PCs])
```



Dimension values and measures are called **MEMBERS**
(e.g., [Time][Q1][Jan] is a member of the Time dim. at month level)

Members are ordered and they can be accessed with special functions like FIRST, LAST, CHILD, etc.

All measures are members of [Measures].

Pandas and Cubes

```
SELECT {[Store Type].[Store Type].MEMBERS} ON COLUMNS,
       {[Store].[Store State].MEMBERS} ON ROWS
FROM [Sales]
WHERE (Measures.[Store Sales], [Time].[Year].[1997])
```

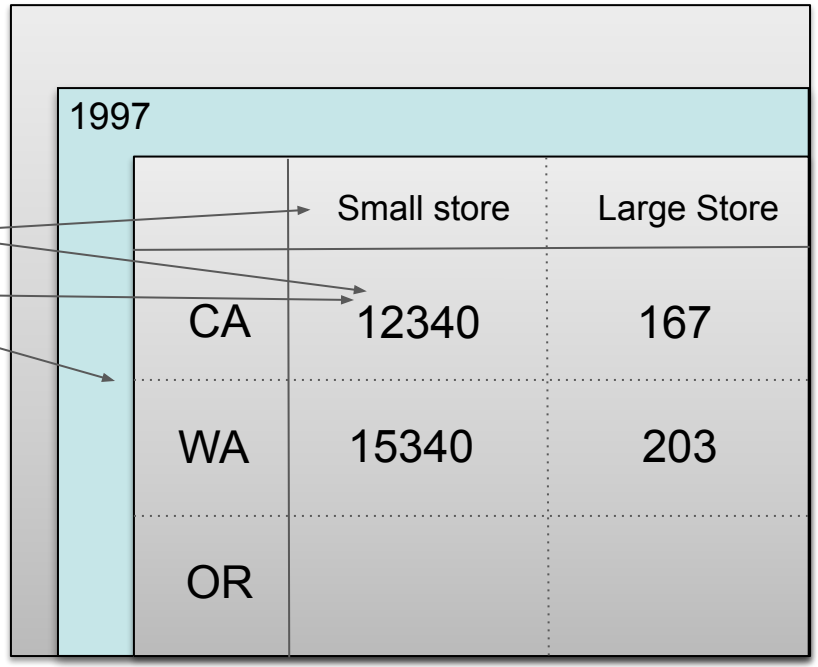
```
# Filter data before pivoting (WHERE)

# Add new columns (calculated members)

pandas.pivot_table(data,
    values= [store_sales],
    index= [store_state],
    columns= [store_type],
    aggfunc='sum',
    fill_value=None,
    margins=False,
    dropna=True,
    margins_name='All',
    observed=False,
    sort=True)

# Perform calculated measures (profit, perc.)

# Filter data on aggregated data
# Concat cubes if necessary
```



ACTIVITIES

Learn MDX language from examples (chatGPT)

Prompt 1. Explain the following MDX query for a non-expert user:

```
SELECT
{ [Measures].[total_sales]}  ON COLUMNS
{ [Time].[T1], [Time].[T2]}  ON ROWS
FROM Sales
WHERE ([Item].[PCs])
```

Prompt 2. Can you draw an example of the result of the previous MDX query?

Generate MDX queries from a schema and an analysis specification (chatGPT)

Prompt. Given the following multidimensional cube

Sales(Time.Quarter.{Q1,Q2,Q3,Q4},
Location.Country.State.{CA,NY,AR}, total_sales)

build an MDX query for **showing the quarter incremental changes (in percentage) of the total sales per state**

Implement MDX queries with Pandas over dataframes

(see notebook)