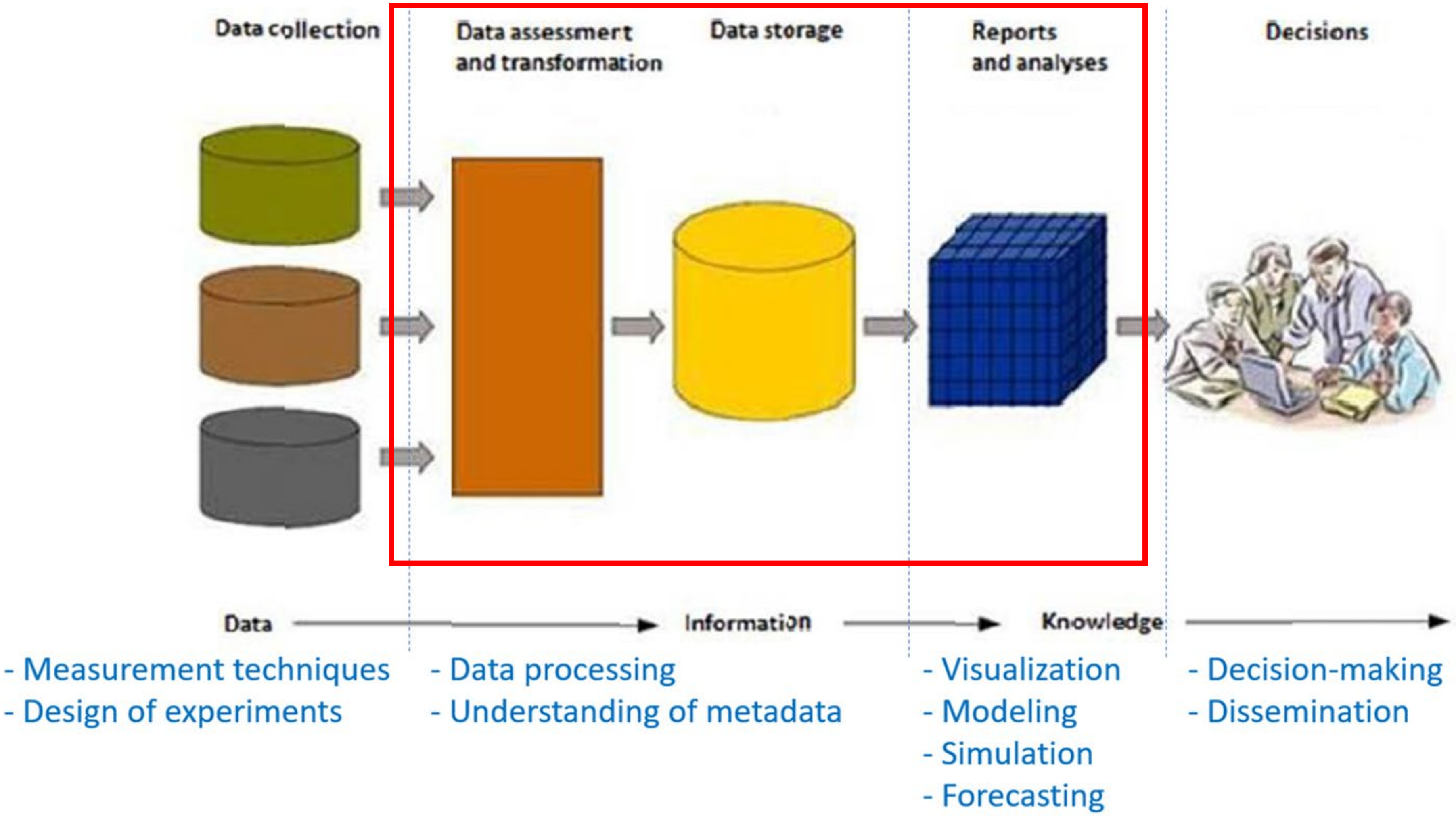
Data Warehousing

Prof. Yves Rybarczyk

Data analytics workflow



Outline

- **Lecture 1** (9 Dec, 10h-12h; room B301) – Introduction to DW
  - Research background
  - Definitions & overview
  - Database vs Data Warehouse
  - Architecture
  - Extract, Transform & Load (ETL) process
- **Lecture 2** (30 Dec, 10h-12h; room B301) – Multidimensional Modeling
  - Data Modelling
  - Online Analytical Processing (OLAP) vs OLTP (Online Transaction Processing)
  - Fact and dimension tables
  - Star and SnowFlake Schema
  - Data Mart

Outline (2)

- **Lecture 3** (7 Jan, 10h-12h; room B301) – Reporting and Data mining
  - Business Intelligence
  - Reporting
  - Data Mining
  - Machine Learning methods

Assessment

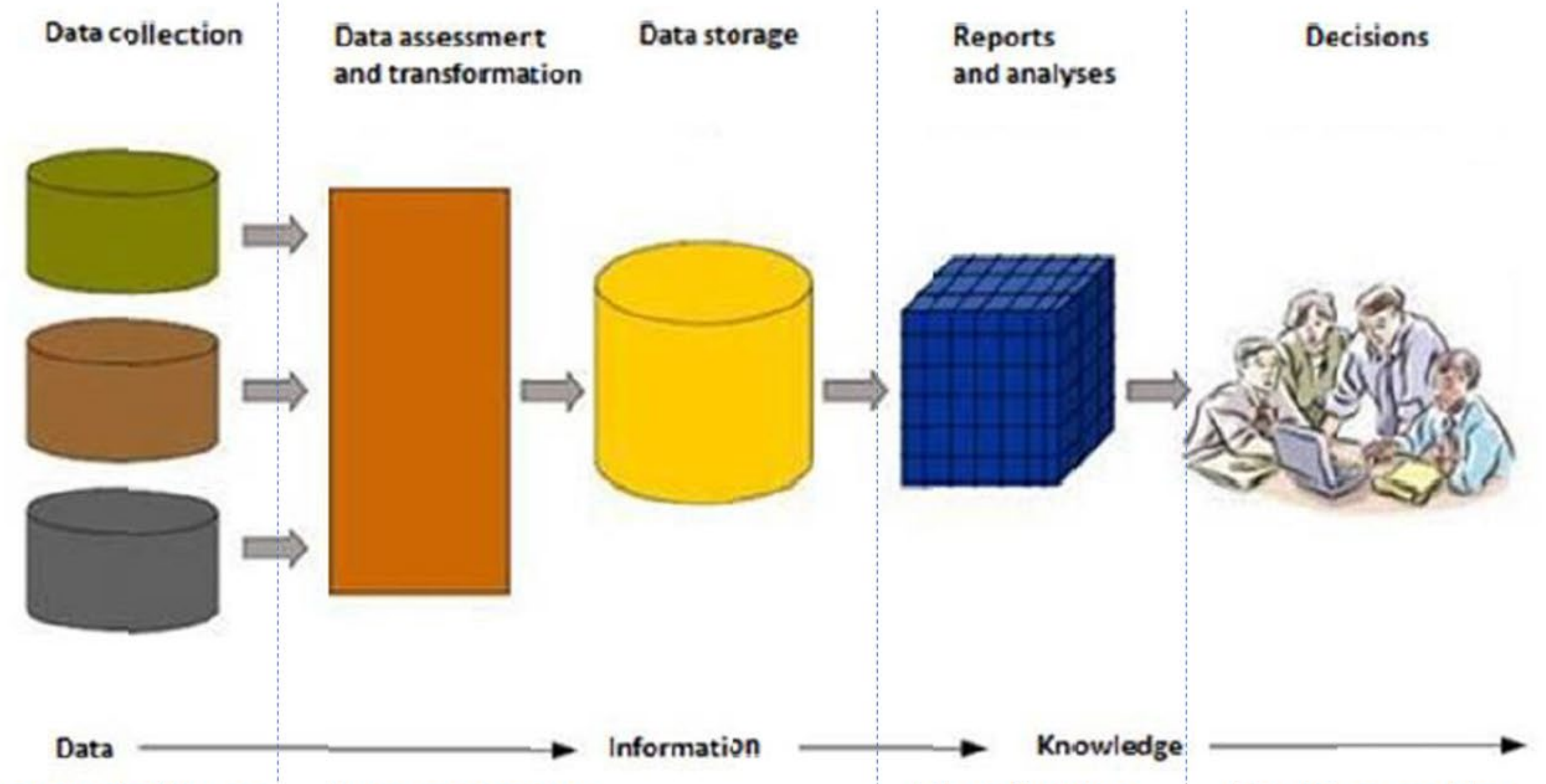
Current CANVAS groups

- **Labs** (groups of 2/3 people; room B101/B301/B302)
  - **Lab 1 – 1 session** (**10 Dec**, 8-11h/12-15h):
    Create an Excel pivot from a DW in Access
  - **Lab 2 – 3 sessions** (**19 Dec**, 10-13h/14-17h; **2 Jan**, 10-13h/14-17h; **8 Jan**, 10-13h/14-17h):
    Implementing a Data Warehouse integrated in a full BI flow (SQL Server and VS)
- **Seminar** (**18 Dec**, 10-13h/14-17h; room B301)
  - Lecture 1 break: choose one **<u>different</u>** printed paper for each group (3/4 people)
  - Presentation, discussion and report on the selected paper
- **Written exam** (**15 Jan**, 8-10h; room B401/B403)
  - Register for examination!!!
- **Lab 2 examination** (**16 Jan**, 8-12h; room B302)

Research background

Data Analytics: concept and application for air quality research

Data analytics workflow

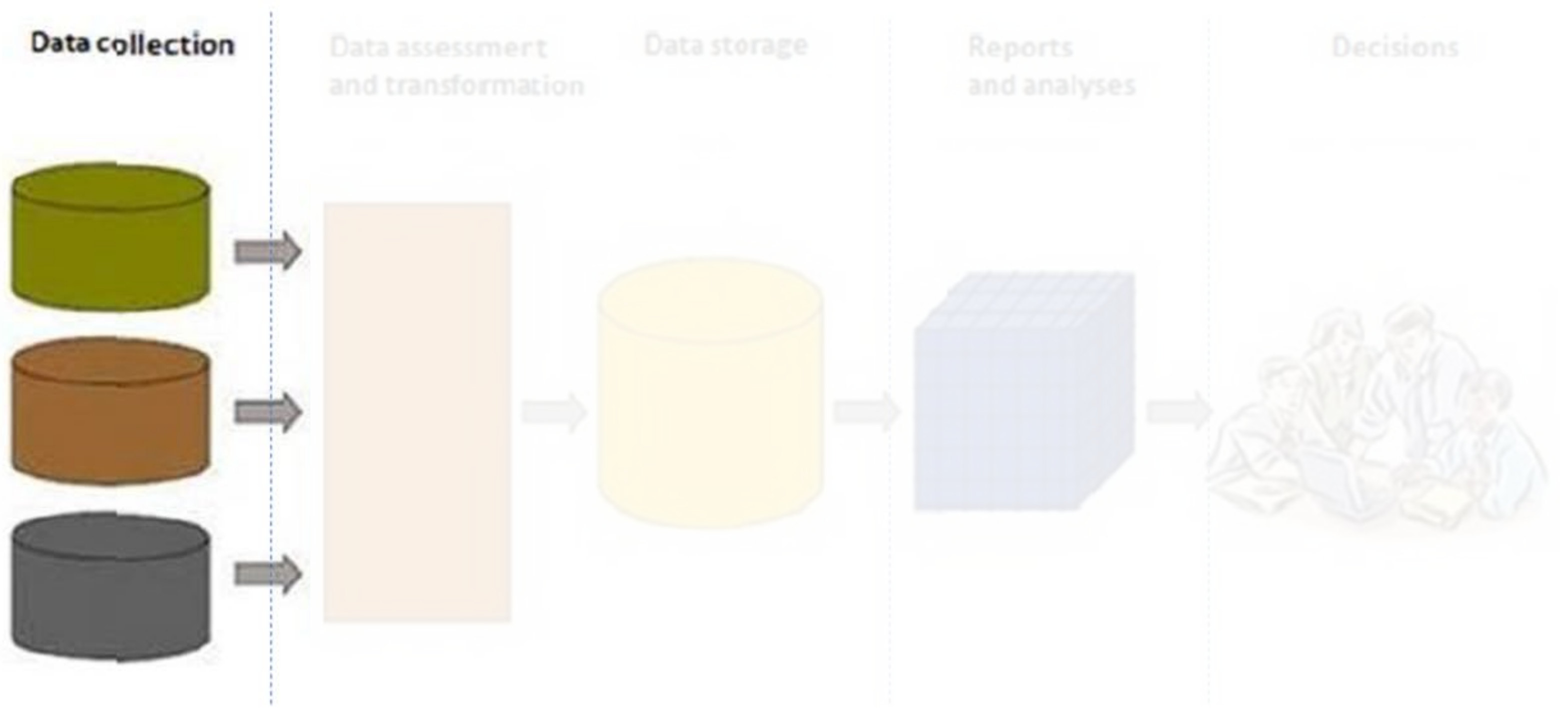


- Measurement techniques
- Design of experiments

- Data processing
- Understanding of metadata

- Visualization
- Modeling
- Simulation
- Forecasting

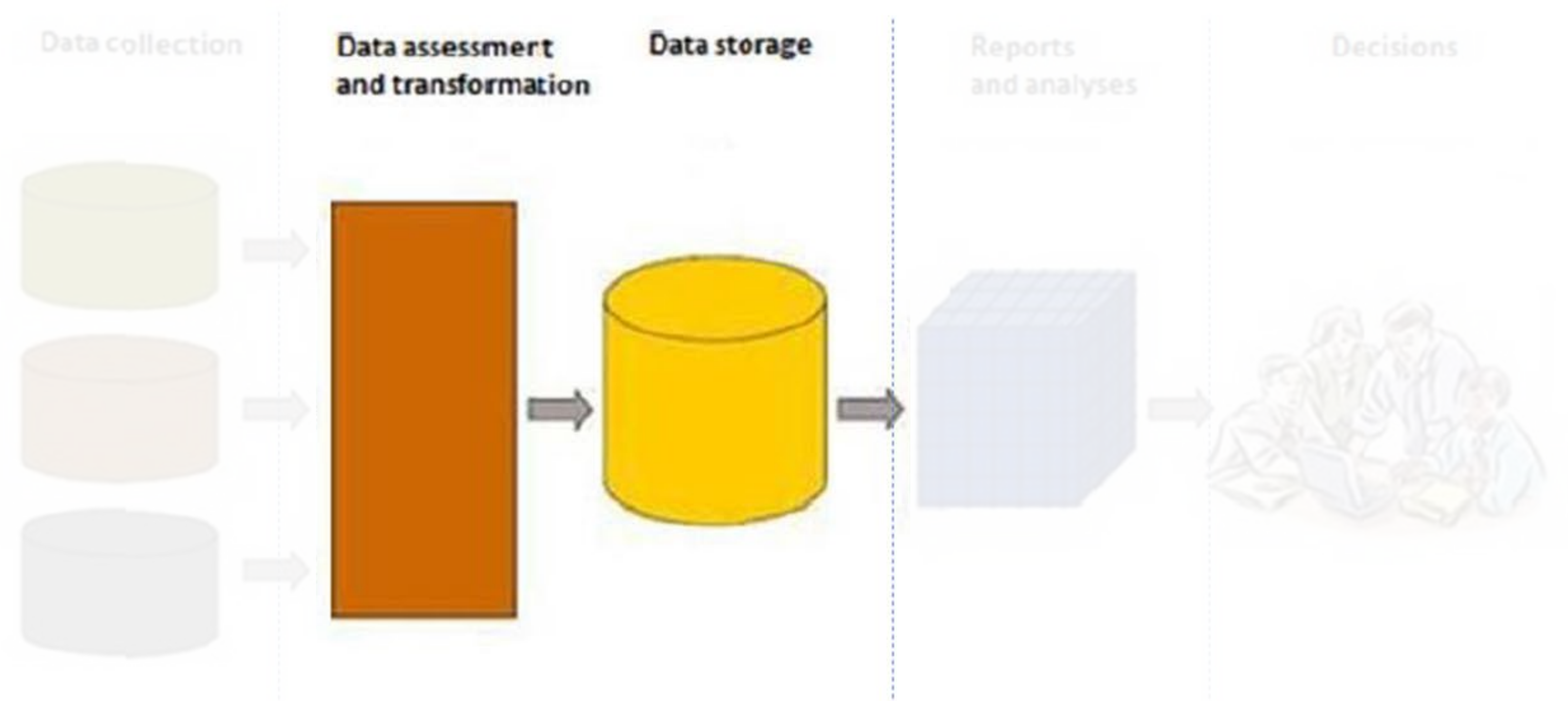
- Decision-making
- Dissemination



Step 1: Data collection

Data sources and data quality

- Sources:
  - Online databases (monitoring stations, satellite data, …)
  - Own measurements (camera, …)
  - Experiments
- Controlled (randomized controlled trials) vs **ecological (observation)** conditions
- Data quality
  - Garbage in, garbage out



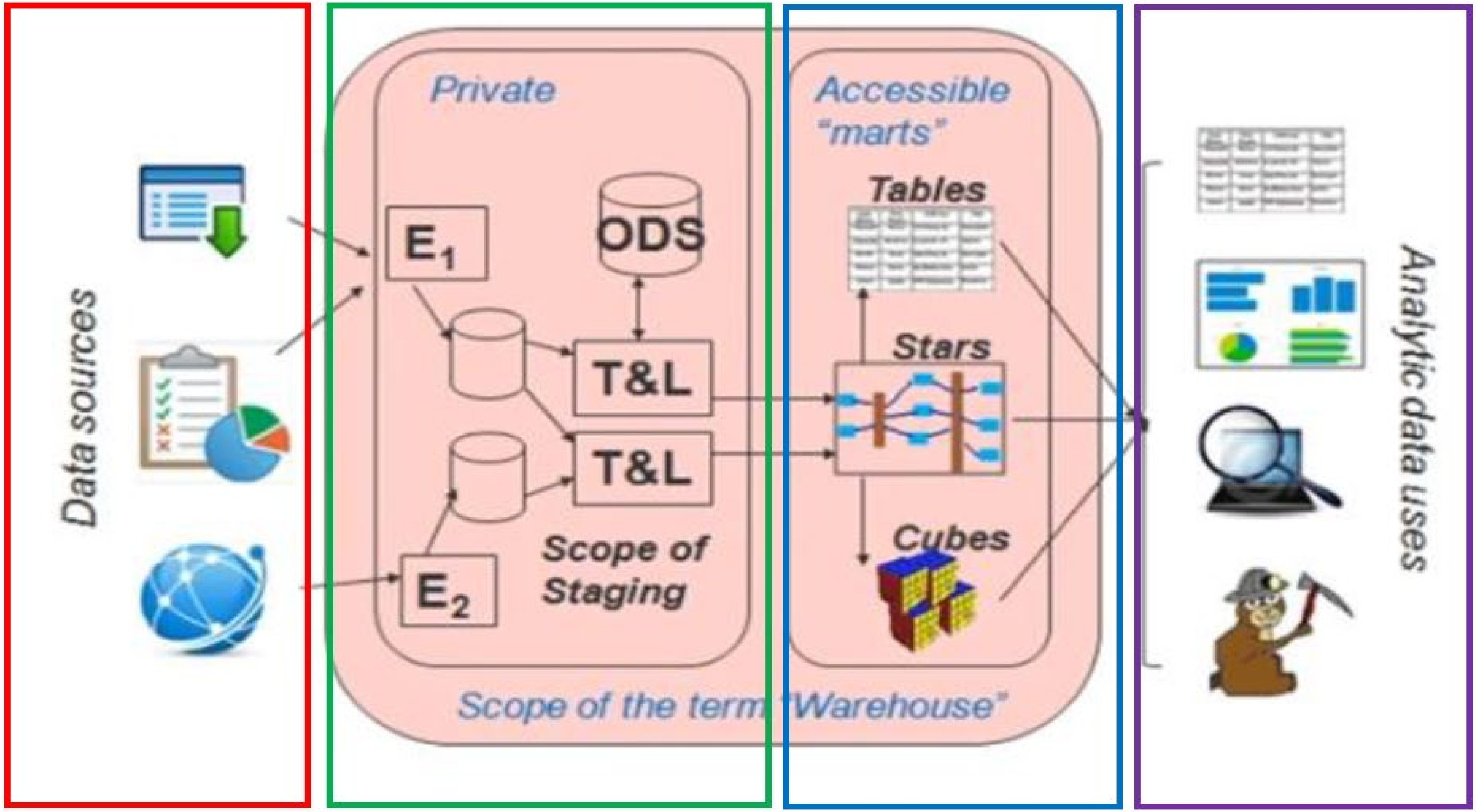
Step 2: Data curation

Data preparation and preprocessing

- Data cleaning (removing noise, identifying outliers, …)
- Missing values (data imputation, removing instances, …)
- Normalization?
- Rescaling (e.g., log)?
- **Dealing with big data:**
  - Huge number of features and observations
  - ↗ data => ↗ model accuracy (e.g., deep learning)

Data Warehousing (DW): concept

Heterogeneous data ETL Multidimensional model Analytic tools

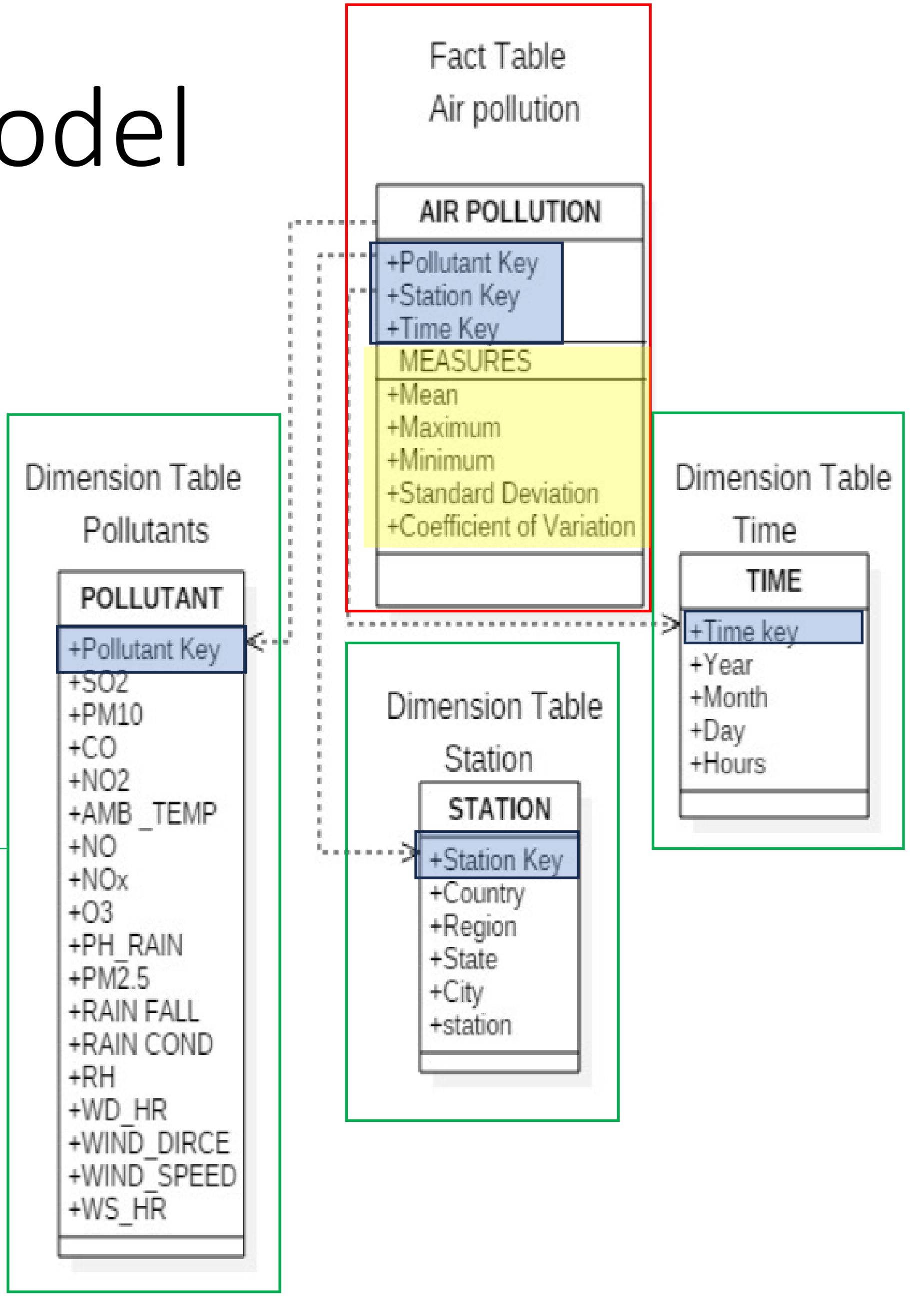


DW: multidimensional model

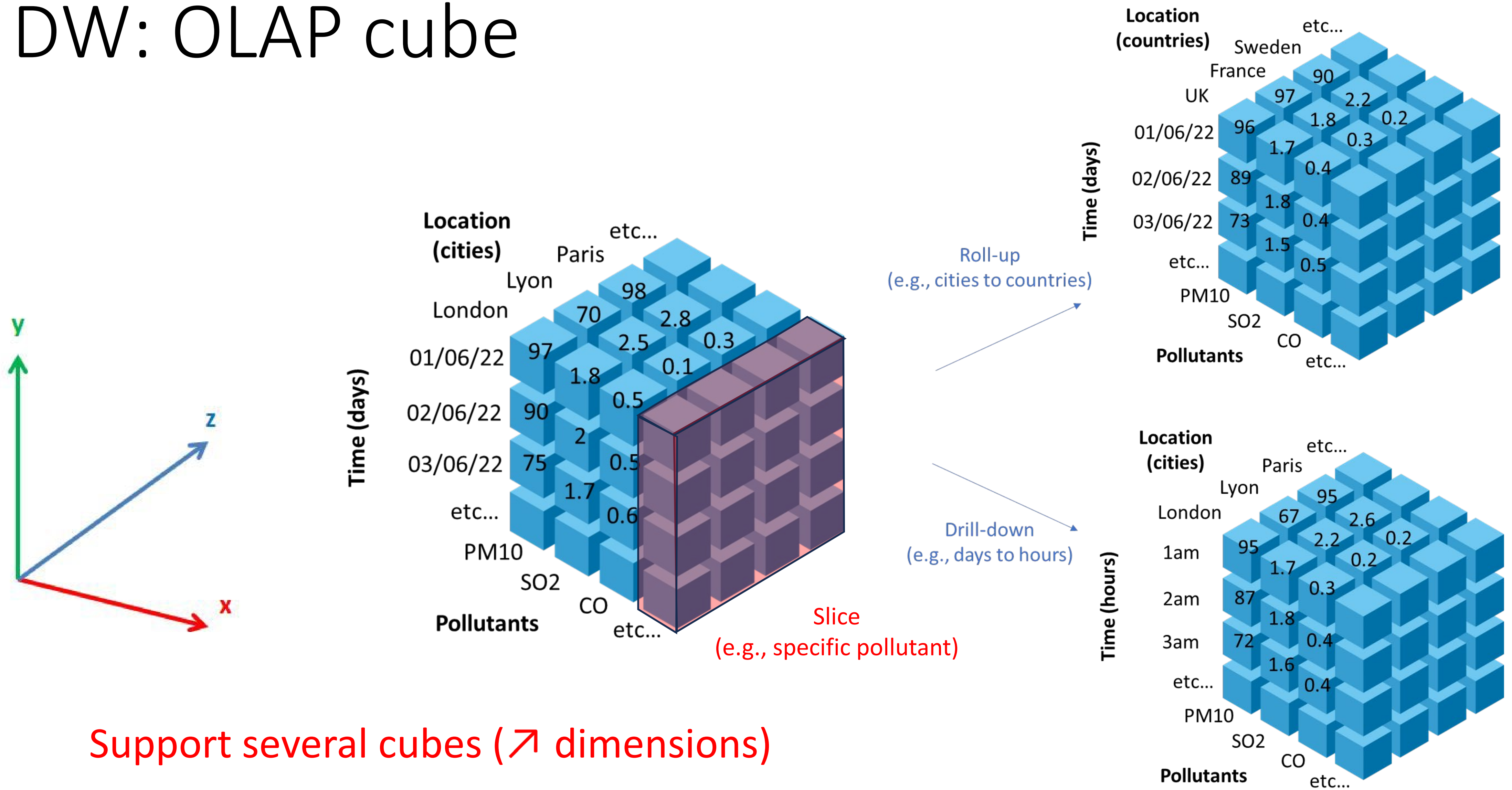
- Star schema

AIR QUALITY MEASURE UNITS AND DURATION

Sl.No	Air Pollutants	duration of Reading	Unit
1	Particulate Matter (PM 10)	1 hour	g/m3
2	Particulate Matter (PM 2.5)	1 hour	g/m3
3	Sulfur dioxide (SO2)	1 hour	PPM
4	Nitrogen dioxide (NO2)	1 hour	PPM
5	Carbon monoxide (CO)	1 hour	PPM
6	Ozone (O3)	1 hour	PPM
7	Ambient Temperature (T)	1 hour	°C
8	Nitric Oxide (NO)	1 hour	PPM
9	Nitrogen Oxides (NOx)	1 hour	PPM
10	Wind Direction (WD)	1 hour	In degrees
11	Wind Speed (WS)	1 hour	In kph
12	Wind Direction HR	1 hour	In degrees
13	Relative Humidity (RH)	1 hour	In %



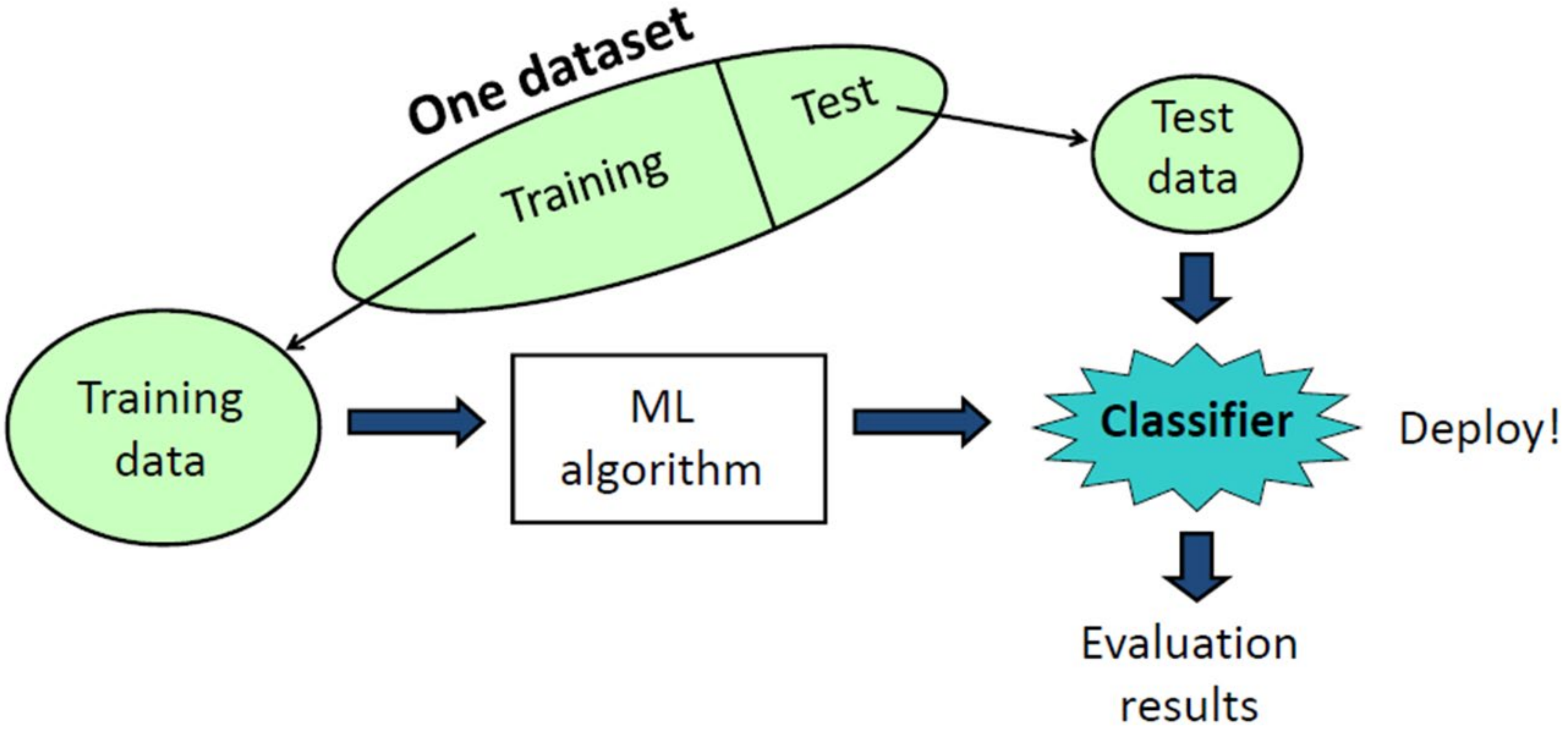
DW: OLAP cube



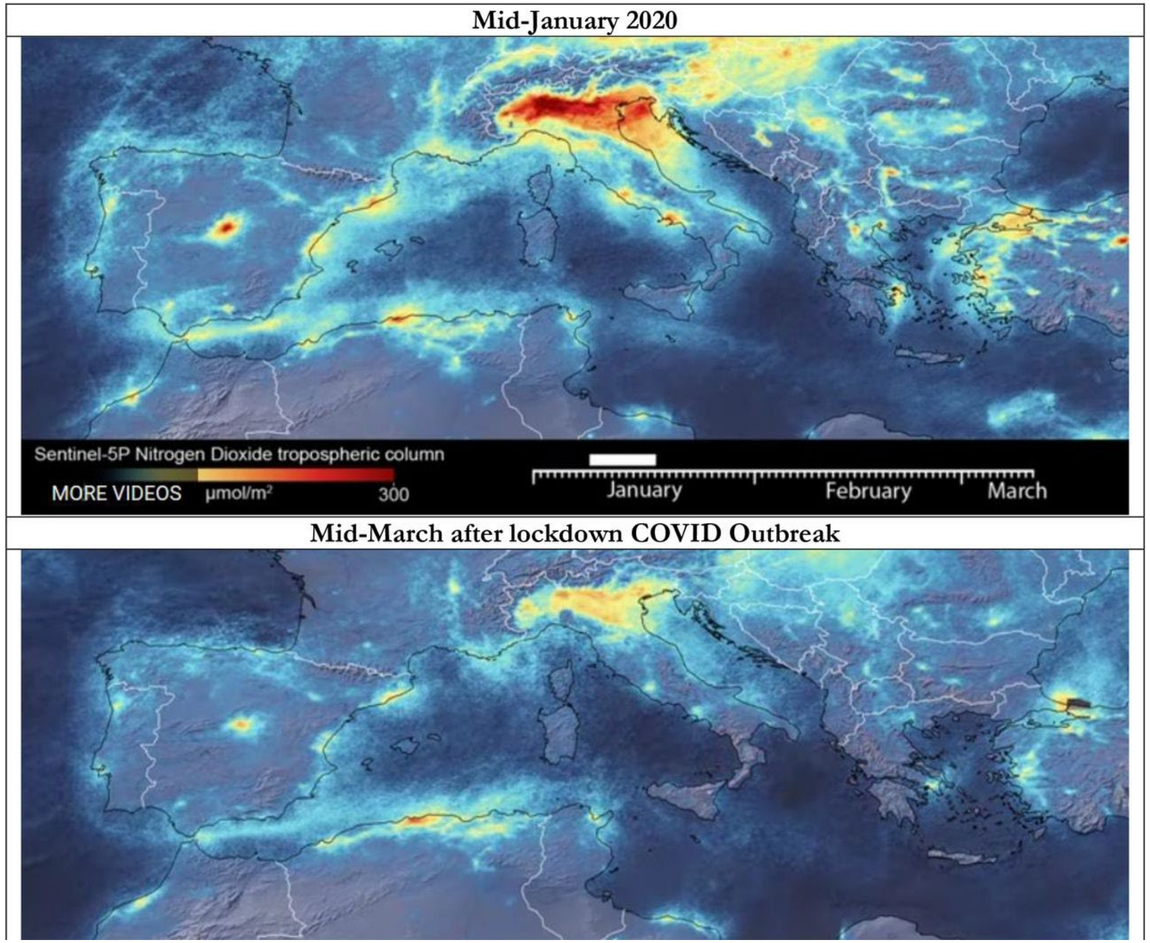
Support several cubes (↗ dimensions)

Step 3: Data analysis

Prediction


One dataset
Training
Test
Test data
Training data
ML algorithm
Classifier
Deploy!
Evaluation results

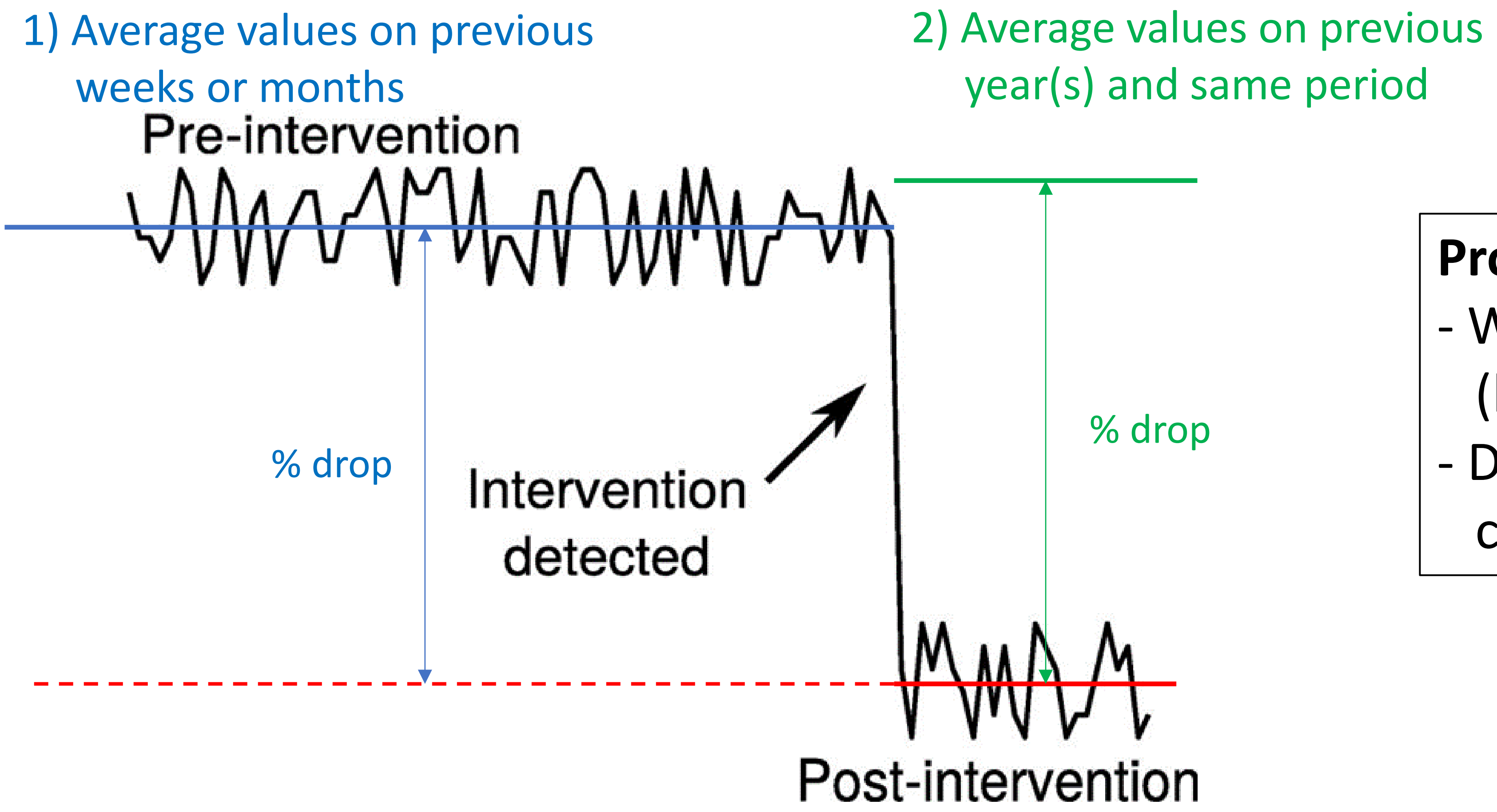
Assessing the COVID-19 Impact on Air Quality
(Rybarczyk & Zalakeviciute, 2021)



How to quantify the drop of air contamination?

=> by comparison to the pollution emission for business as usual (BAU)

How to get the BAU?

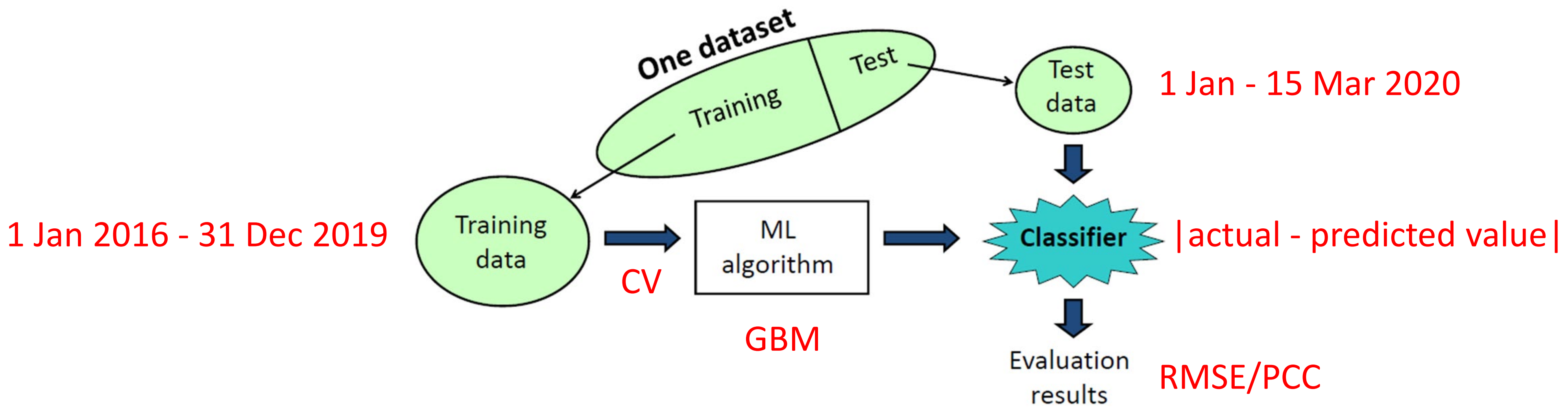


Problems:

- Which one is the true BAU (blue or green)?
- Does not consider the current meteorology!

Weather Normalized Model

- Since the concentration of pollutants is highly dependent on meteorology, the predictive model is built from meteorological features
- Case study: Quito (6 stations), Ecuador
  - 4 pollutants: NO2, CO, SO2, and PM2.5
  - 7 weather features: RH, precipitation, temperature, SR, pressure, WS, and WD
  - 4 temporal features: Julian day, weekday, hour, and date index (trend)

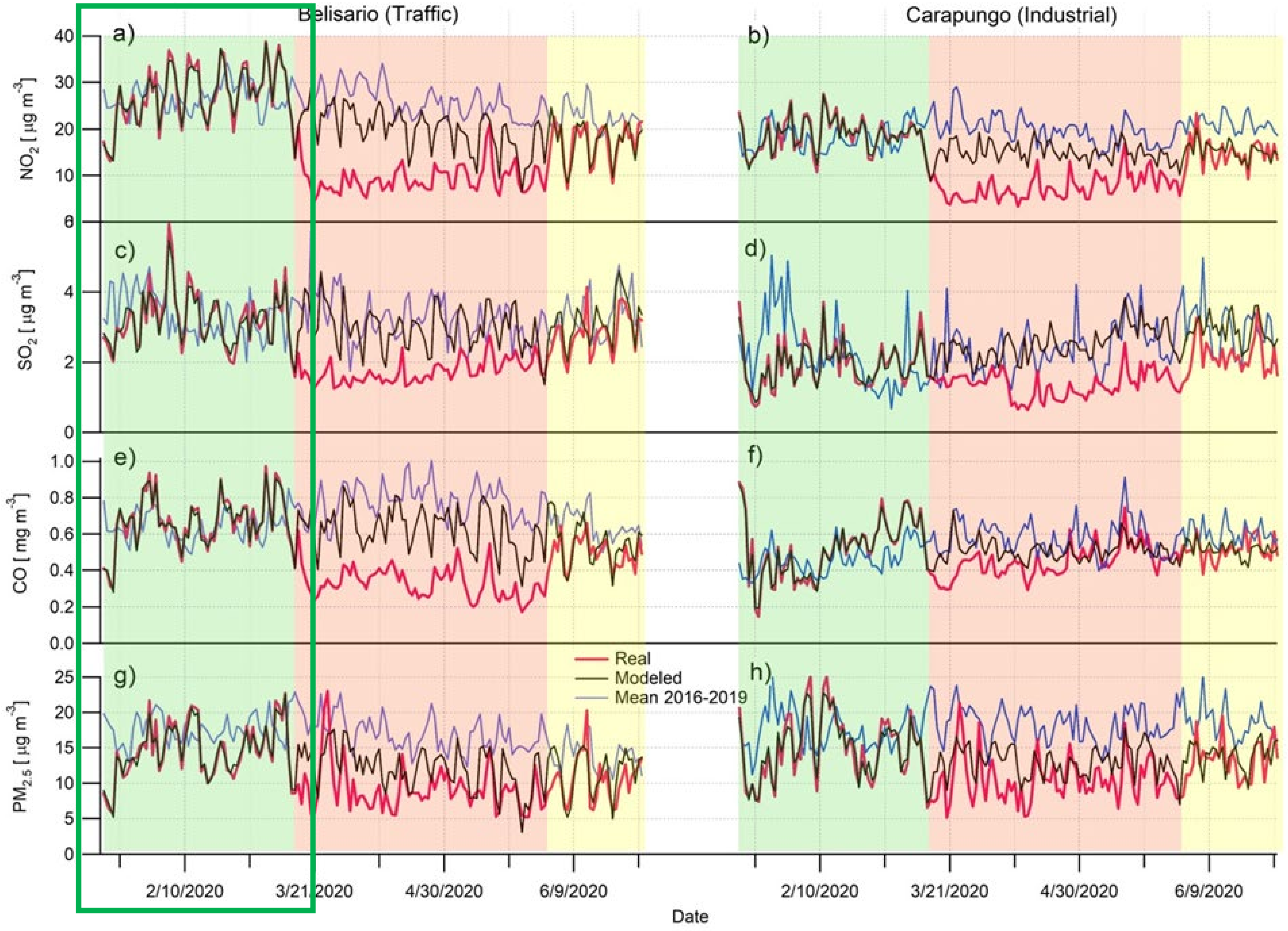


Method 2
(previous slide)

Results: model accuracy

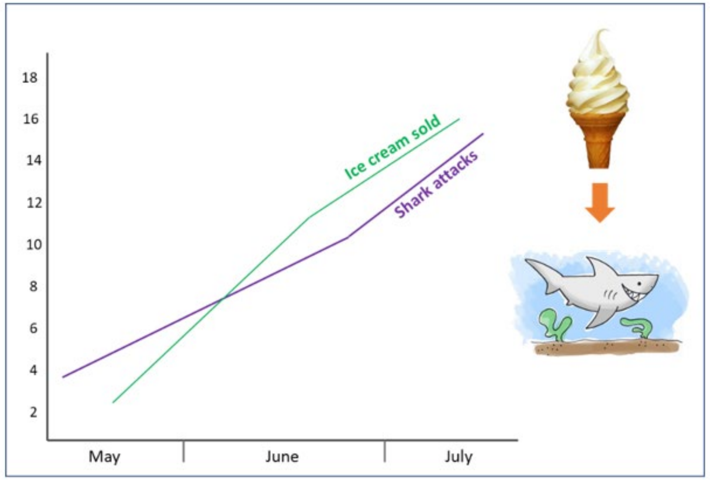
District	Type	Pollutant	Machine learning model: Number of trees	RMSE (train)	PCC (train)	RMSE (test)	PCC (test)	Mean 2016-2019: PCC
Belisario	Urban - traffic	NO2	8009	5.1	0.91	6.2	0.86	0.07
		SO2	7895	1.7	0.84	2.2	0.7	0.27
		CO	6644	0.1	0.93	0.2	0.88	0.09
		PM2.5	6585	5.2	0.85	6.5	0.75	0.22
Carapungo	Urban - industrial	NO2	6741	4.7	0.92	5.9	0.87	0.03
		SO2	6926	1.4	0.89	2.0	0.76	0.08
		CO	6803	0.1	0.92	0.2	0.85	0.29
		PM2.5	9852	7.2	0.83	9.1	0.72	0.04
Camal	Urban - industrial	NO2	8030	5.4	0.91	7.0	0.85	0.08
		SO2	10033	2.8	0.91	4.2	0.79	0.18
		CO	8222	0.2	0.93	0.2	0.86	0.13
		PM2.5	8725	9.7	0.76	12.1	0.59	0.08
Cotocollao	Suburban - traffic	NO2	12862	2.2	0.97	3.8	0.92	0.1
		SO2	13368	0.6	0.96	1.2	0.85	0.12
		CO	12253	0.1	0.97	0.2	0.88	0.06
		PM2.5	9291	5.6	0.88	9.5	0.61	0.18
Guamani	Suburban - agricultural	NO2	9483	4.3	0.94	5.8	0.88	0.23
		SO2	11099	1.1	0.85	1.7	0.6	0.25
		CO	9594	0.1	0.93	0.2	0.84	0.13
		PM2.5	9040	8.3	0.79	10.8	0.6	0.17
Chillos	Suburban - industrial	NO2	9202	4.5	0.9	5.9	0.82	0.15
		SO2	6494	4.7	0.84	6.5	0.66	0.07
		CO	7908	0.1	0.92	0.1	0.87	0.31
		PM2.5	8443	5.6	0.78	6.9	0.63	0.01

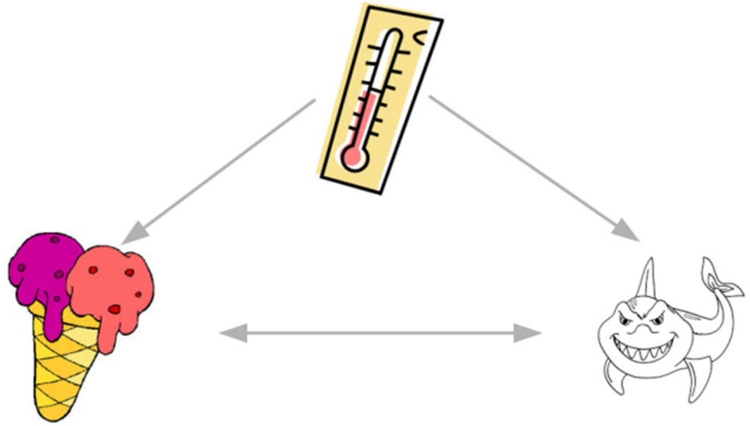
Results: time series



Causality

- “To predict is not to explain” (René Thom)





- How to infer causations in **observational studies** with a **data-driven approach**?
  - No experiment
  - No mechanistic model (“Data Analytics branding”)

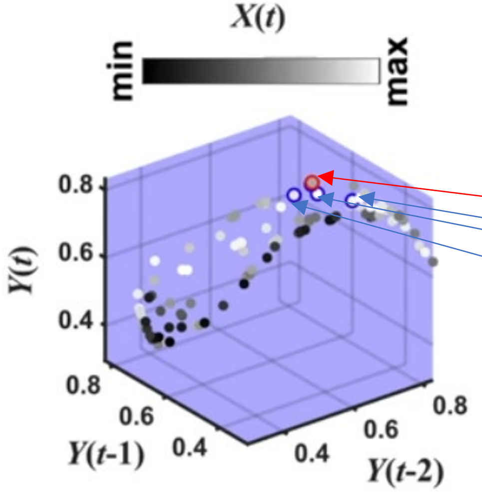
Dynamic causation

- Time series variables are causally related if they are coupled (perturbing one variable perturbs the other) and belong to the same dynamic system.

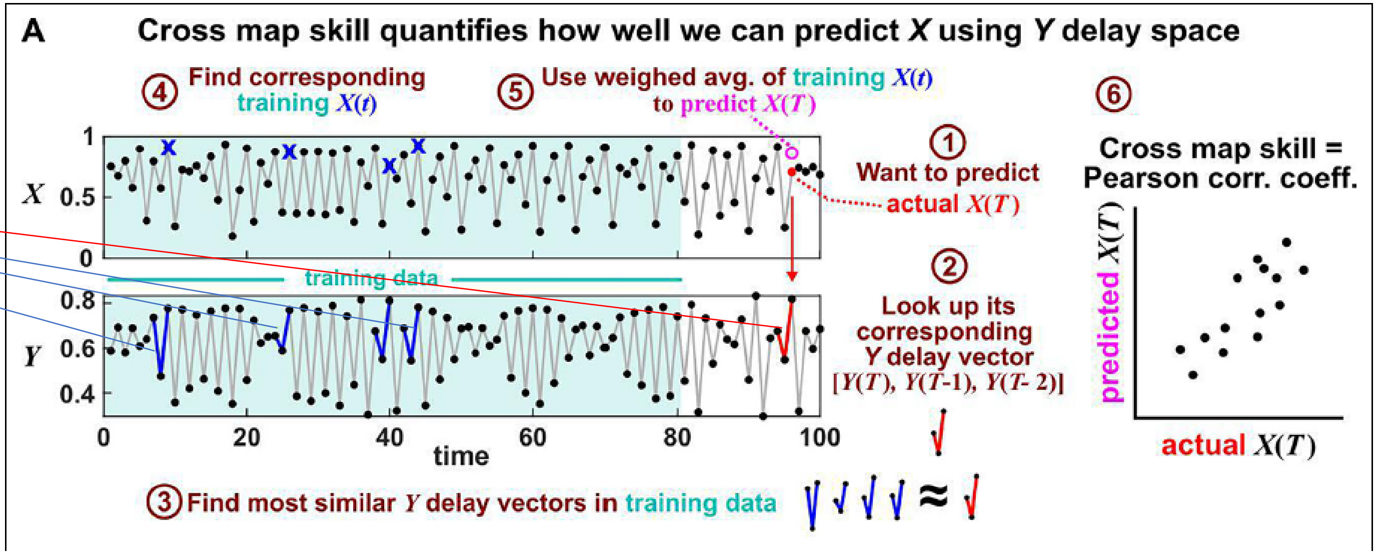
- If *X* causes (influences) *Y* then, *Y* contains information about *X* that can be used to predict (recover) *X*.

- That is, states of *X* can be recovered from the history of *Y*.

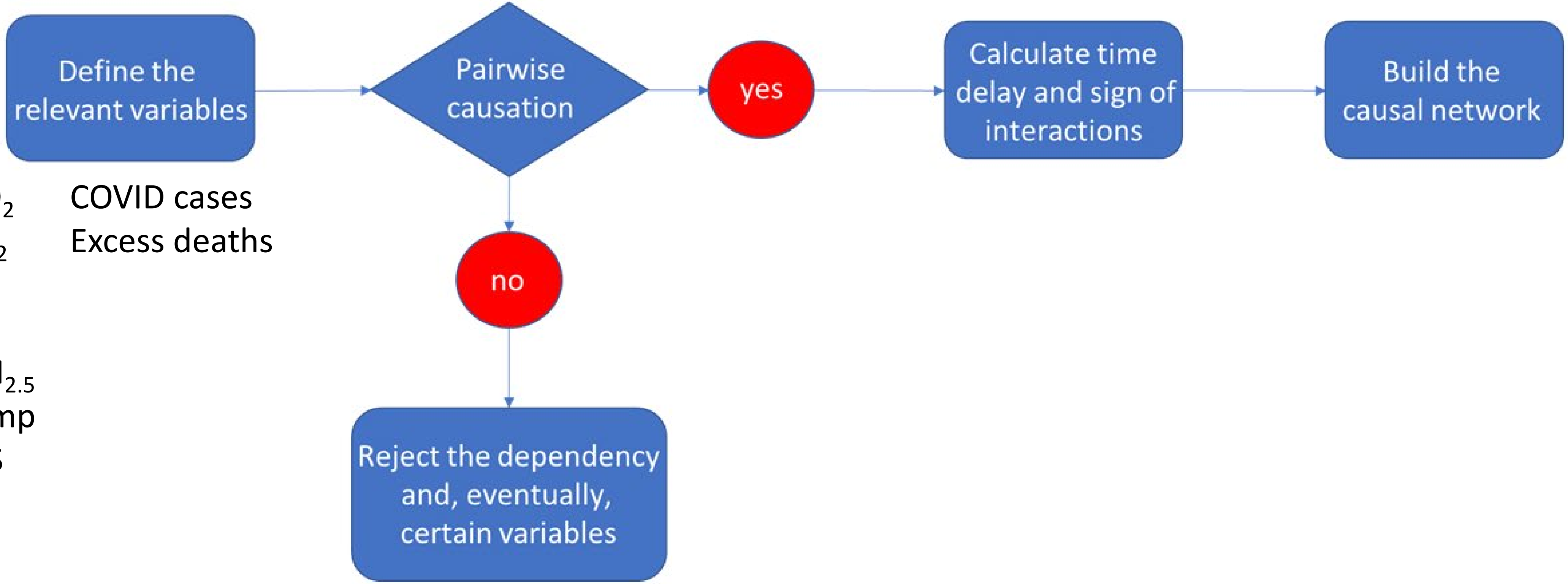
Convergent Cross Mapping (CCM)



State Space
Reconstruction



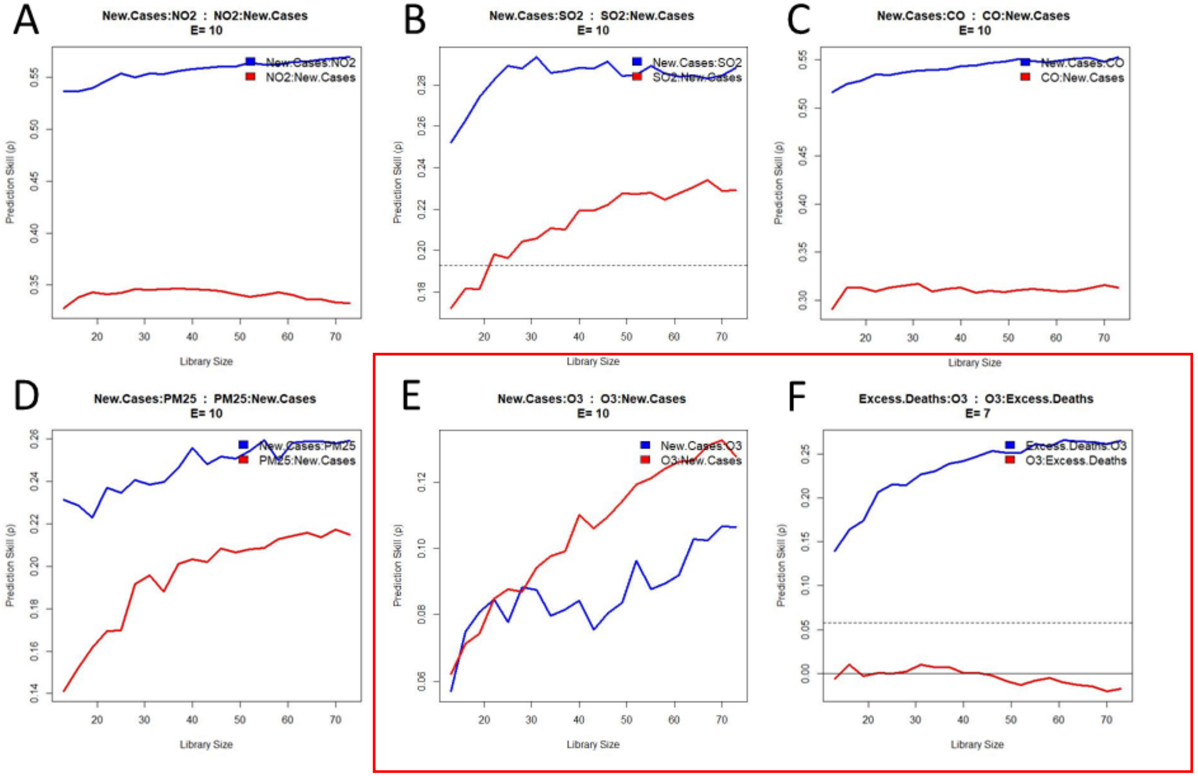
Causal effect of air pollution and meteorology on the COVID-19 pandemic (Rybarczyk et al., 2023)

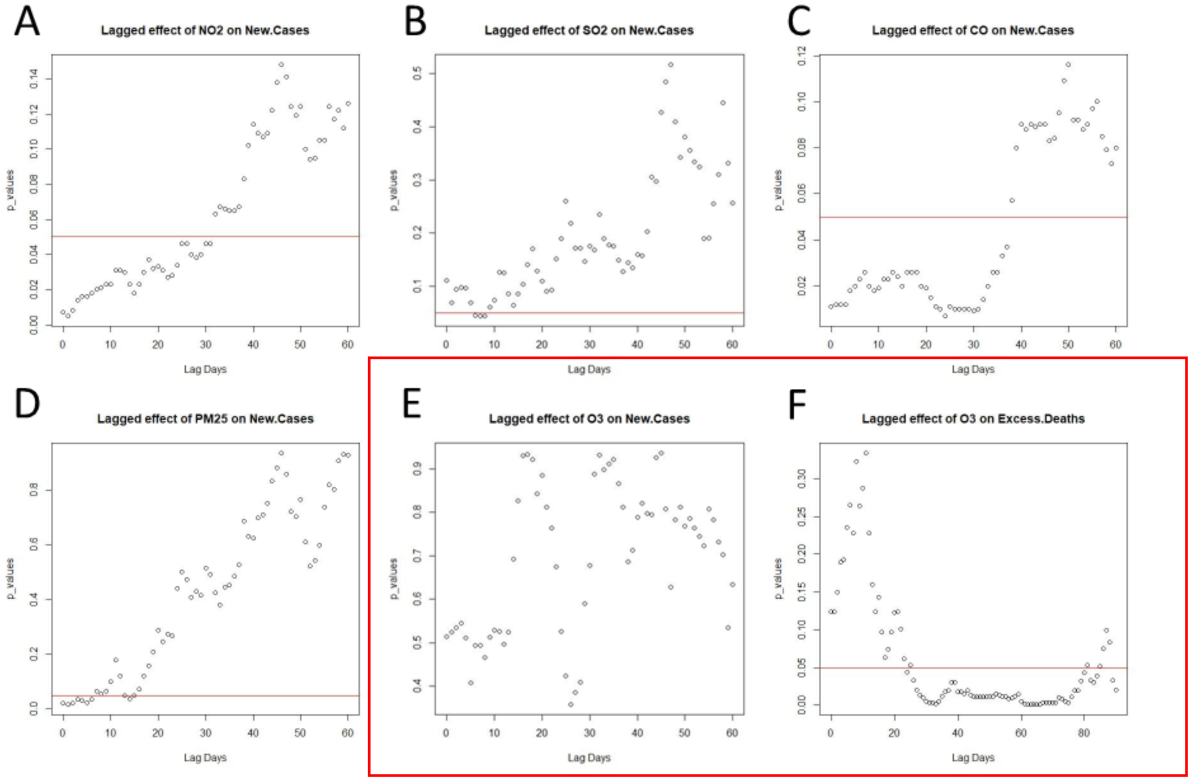


NO_2
SO_2
CO
O_3
$PM_{2.5}$
Temp
WS
RH

COVID cases
Excess deaths

Results: criteria 1-3 (pollutants only)



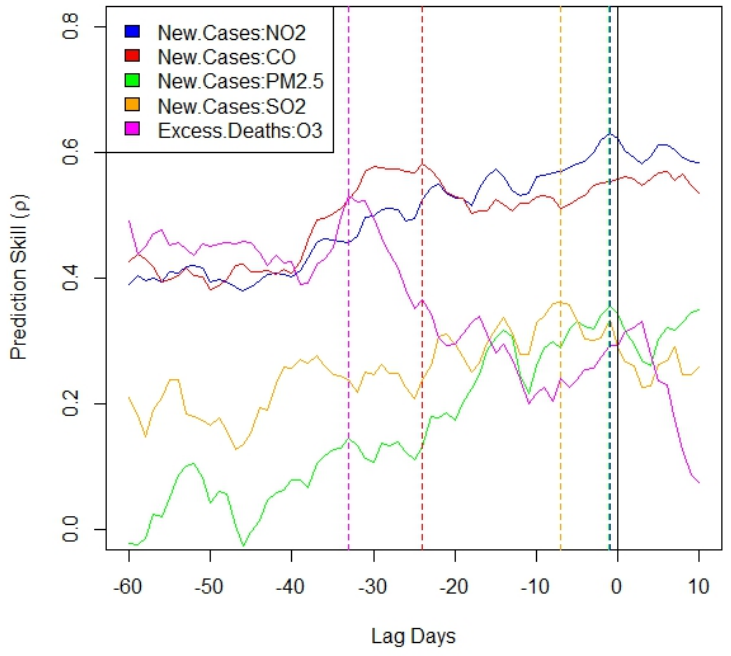


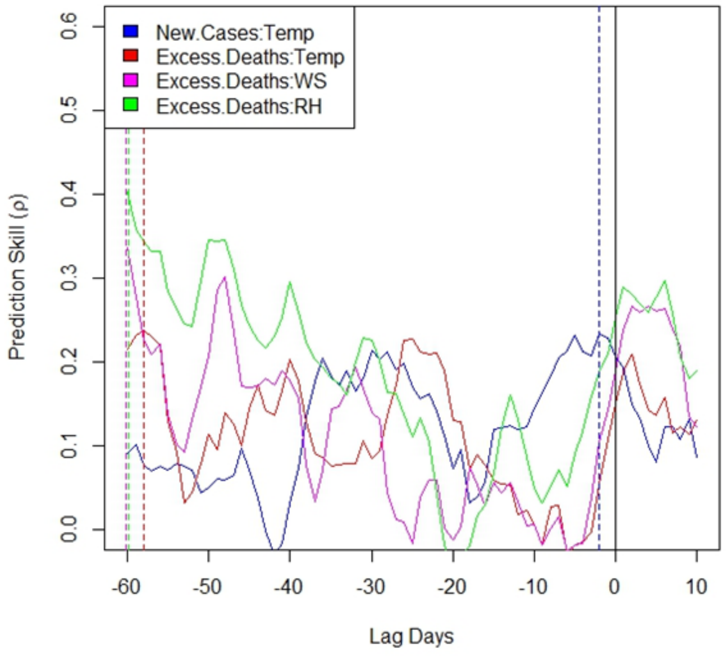
O3 -> COVID-19

COVID-19 -> O3

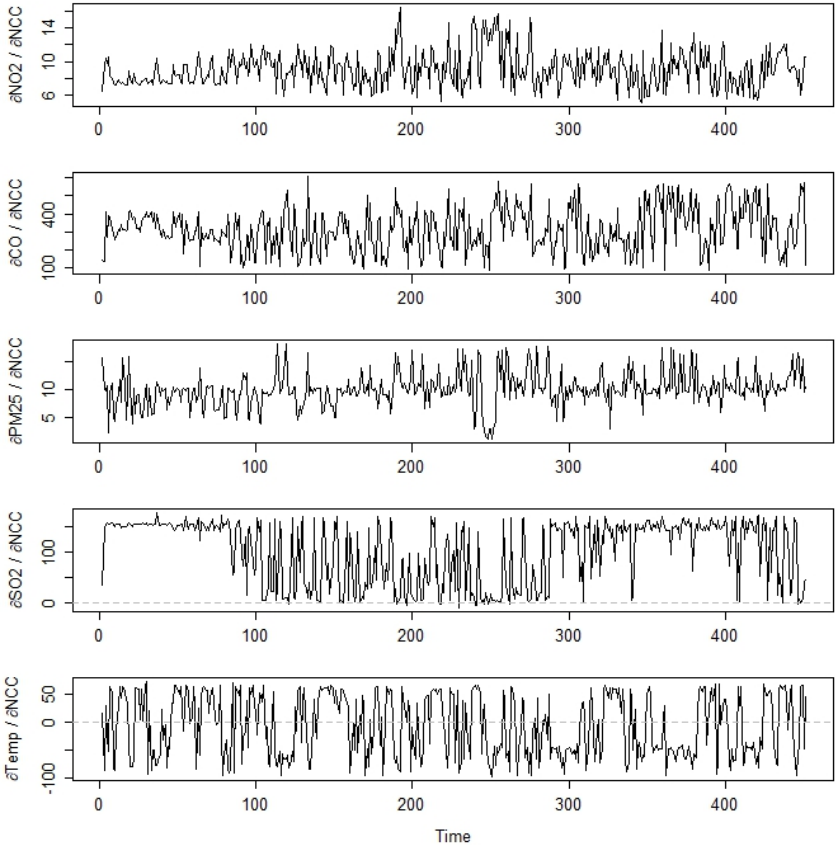
~~surrogate test~~ surrogate test

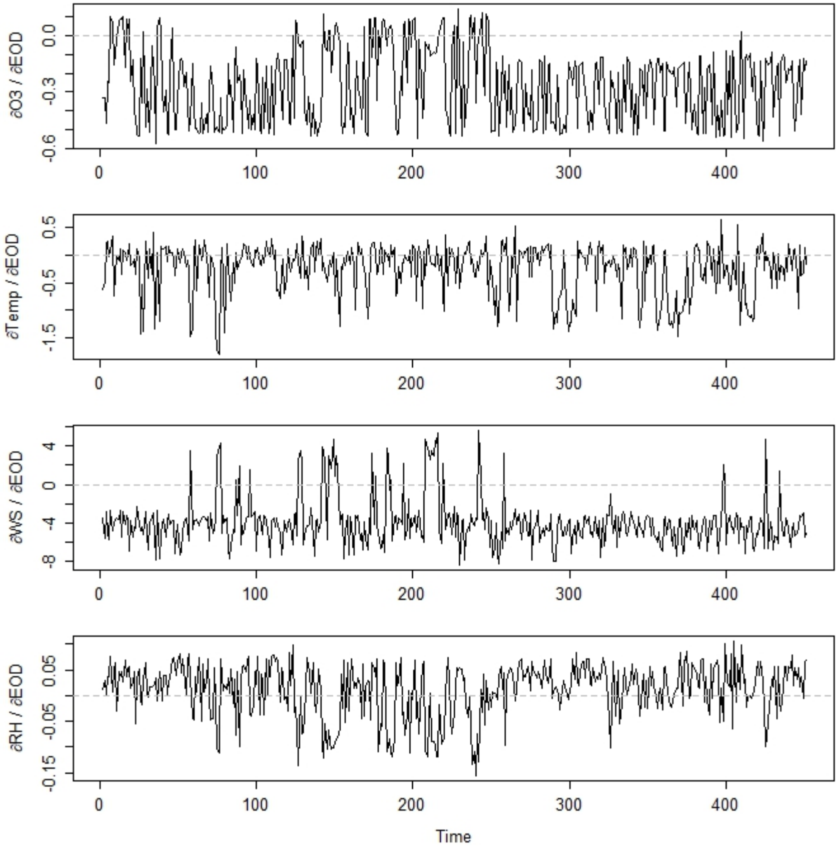
Results: time-delayed causation





Results: sign of interactions

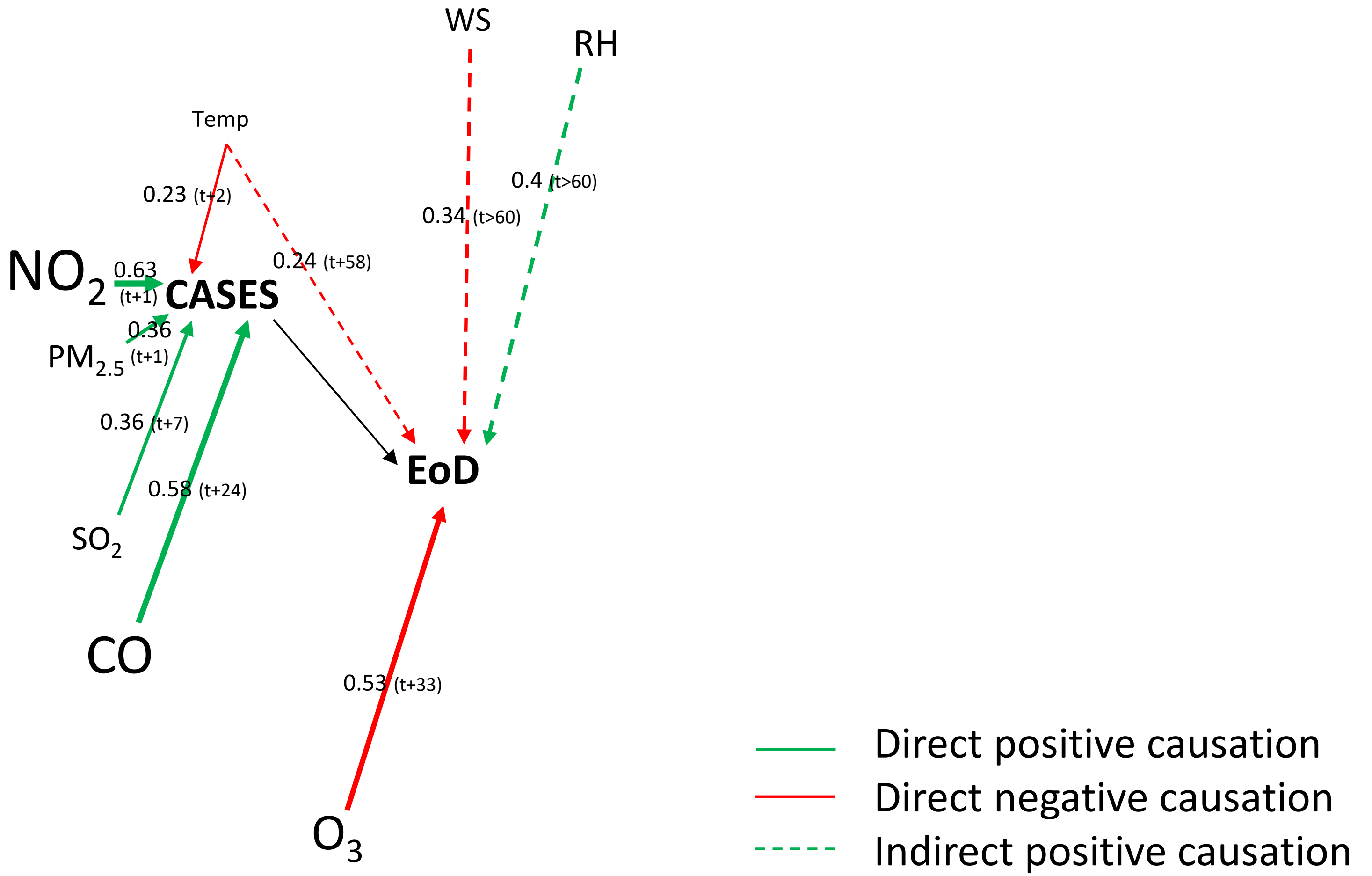


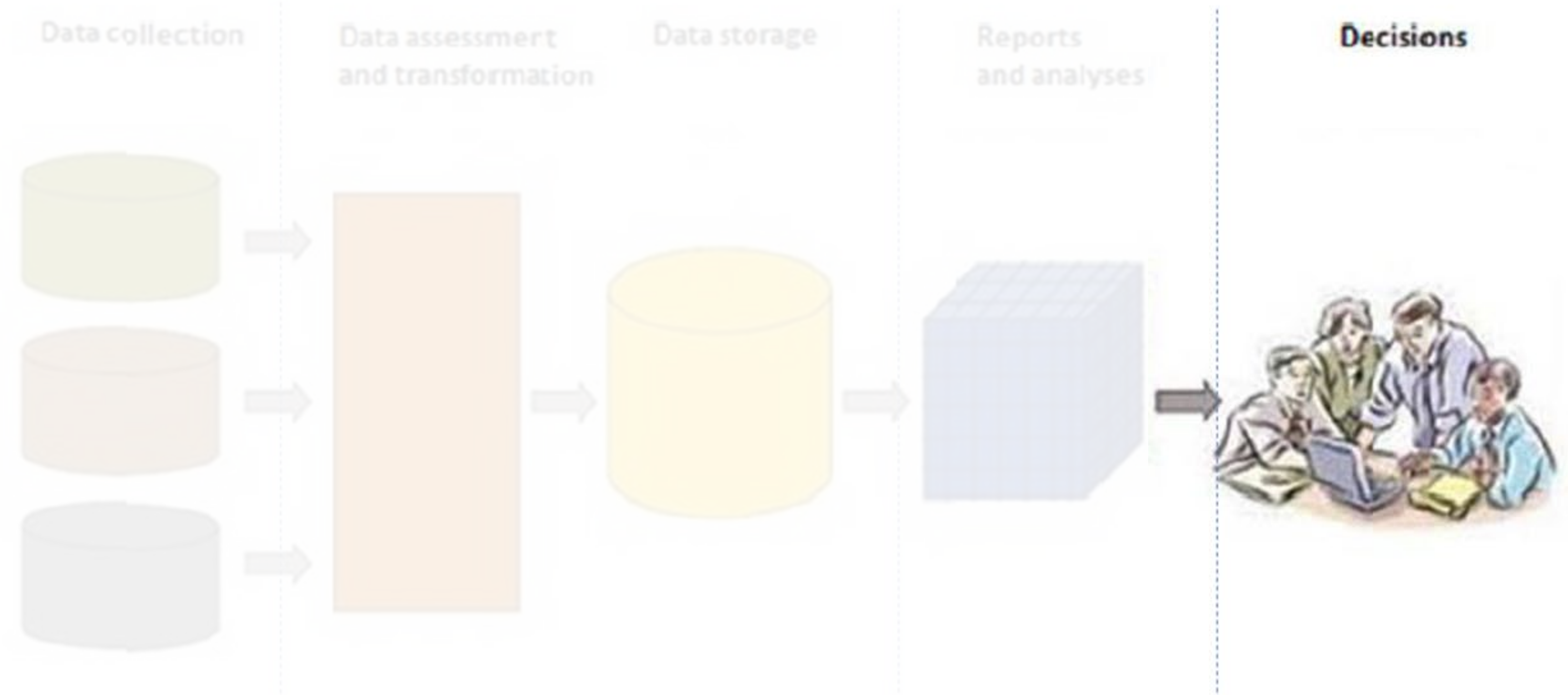


Weighted causal network:

arrow length = log(delay)

arrow width & name size = x.Xmap





Step 4: Decision-making

Decision and dissemination

- Depend on the domains of application (unlimited)
- Pollution: public health strategies
  - Anticipating potential outbreaks
  - Plan timely interventions
  - Social awareness (public outreach)
- Structured multidimensional data source on a specific theme (DW)

Conclusion

Science

RESEARCH ARTIC

Check for updates

Cite as: R. Lam *et al.*, *Science*
10.1126/science.adi2336 (2023).

Learning skillful medium-range global weather forecasting

Remi Lam[1*†], Alvaro Sanchez-Gonzalez[1*†], Matthew Willson[1*†], Peter Wirnsberger[1†], Meire Fortunato[1†], Ferran Alet[1†], Suman Ravuri[1†], Timo Ewalds[1], Zach Eaton-Rosen[1], Weihua Hu[1], Alexander Merose[2], Stephan Hoyer[2], George Holland[1], Oriol Vinyals[1], Jacklynn Stott[1], Alexander Pritzel[1], Shakir Mohamed[1*], Peter Battaglia[1*]

[1]Google DeepMind, London, UK. [2]Google Research, Mountain View, CA, USA.

*Corresponding author. Email: remilam@google.com (R.L.); alvarosg@google.com (A.S.); matthjw@google.com (M.W.); shakir@google.com (S.M.); peterbattaglia@google.com (P.B.)

†These authors contributed equally to this work.

Global medium-range weather forecasting is critical to decision-making across many social and economic domains. Traditional numerical weather prediction uses increased compute resources to improve forecast accuracy, but does not directly use historical weather data to improve the underlying model. Here, we introduce "GraphCast," a machine learning-based method trained directly from reanalysis data. It predicts hundreds of weather variables, over 10 days at 0.25° resolution globally, in under one minute. GraphCast significantly outperforms the most accurate operational deterministic systems on 90% of 1380 verification targets, and its forecasts support better severe event prediction, including tropical cyclones tracking, atmospheric rivers, and extreme temperatures. GraphCast is a key advance in accurate and efficient weather forecasting, and helps realize the promise of machine learning for modeling complex dynamical systems.

What is Data Warehousing?

Definitions & Examples

Definitions

- Technique for collecting and managing **data from varied sources** to provide **meaningful business insights**.
- The data warehouse is the **core of the BI** system, which is **built for data analysis and reporting**.
- An **architectural construct** using **historical data** to support decision-making information:
  - The traditional relational databases involve relation between many tables, which may slow down the response time of the query. A data warehouse provides a new design which can **help to reduce the response time and to enhance the performance of queries for reports and analytics**.

Synonymous

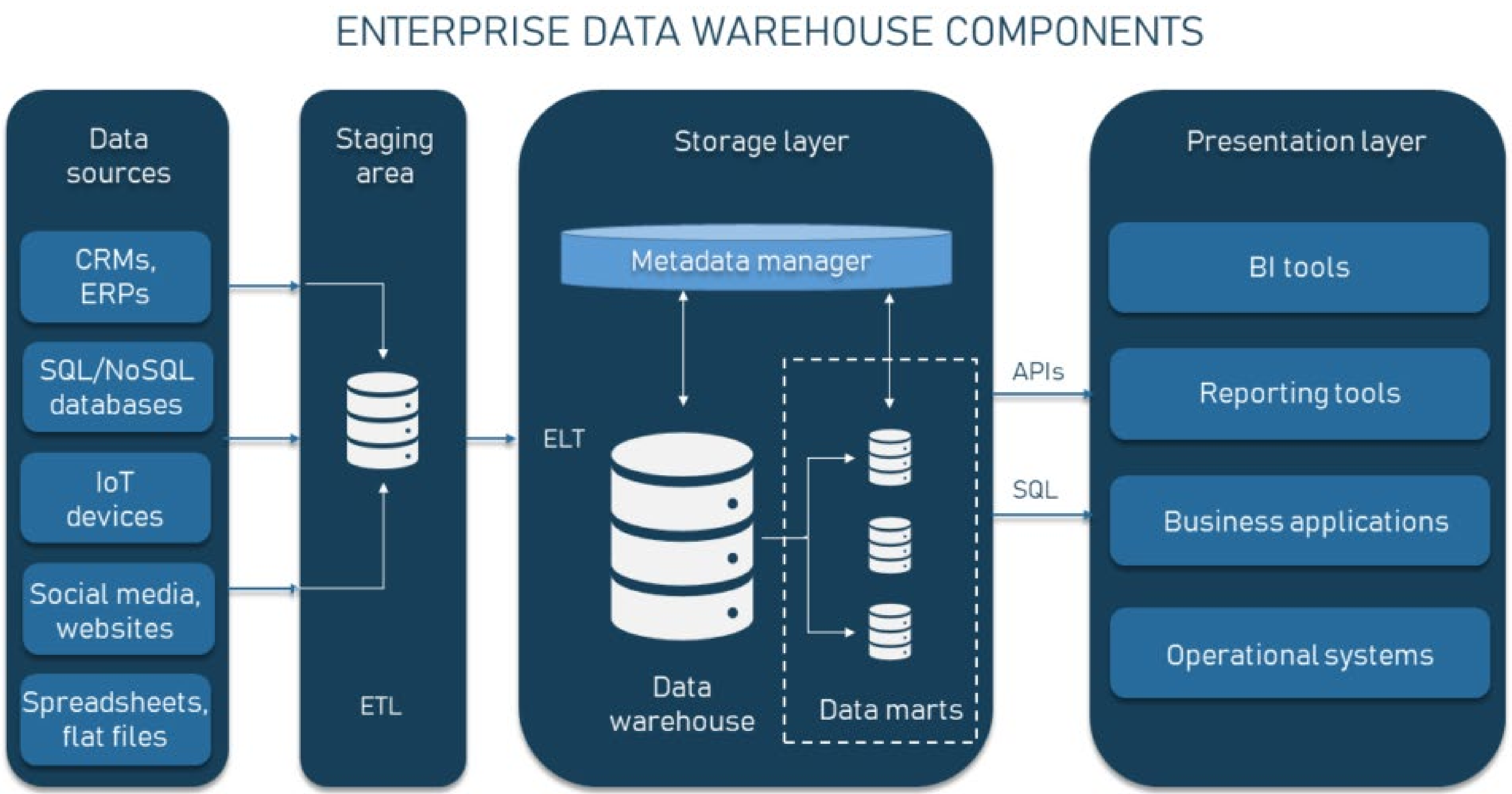
Decision Support System
Data Warehouse
Executive Information System
Other Names for Datawarehouse
Analytic Application
Management Information System
Business Intelligence Solution

How Data Warehouse (DW) works?

- A DW works as **a central repository** dealing with **heterogenous data:**
  - => Data are processed, transformed, and ingested.
  - End user accesses the processed data through Business Intelligence tools, SQL clients, and spreadsheets.
- By merging all this information in one place, an organization can **analyse its customers more holistically:**
  - Reporting.
  - Data mining.

The 4 components of a DW

- **Load manager**
  - Also called the **front component**.
  - Performs with all the operations associated with the **extraction, preparation and load of data** into the warehouse.
- **Warehouse manager**
  - Performs operations like analysis of data **to ensure consistency, creation of indexes and views, generation of denormalization and aggregations, transformation and merging of source data and archiving/backing-up data.**
- **Query manager**
  - Also known as **backend component**.
  - Performs all the operations related to the **management of user queries**.
- **End-user access tools**
  - Categorized into 5 different groups: 1) **Data Reporting**, 2) **Query Tools**, 3) **Application development tools**, 4) **EIS tools**, and 5) **OLAP/Data mining tools**.



Lecture 1 Lecture 2 Lecture 3

Who needs Data Warehouse?

- **Decision makers** who rely on mass amount of data.
- Users who use customized, complex processes **to obtain information from multiple data sources**.
- People who want **simple technology to access the data**.
- People who want a **systematic approach for making decisions**.
- Users who want **fast performance on a huge amount of data** for reports, grids or charts.
- **To discover 'hidden patterns'** of data-flows and groupings.

Sectors where DW is used

- **Airline**
  - For operation purpose like **crew assignment**, analyses of route profitability, frequent flyer **program promotions**…
- **Banking**
  - For the market research, **performance analysis of the products** and operations.
- **Healthcare**
  - To strategize and **predict outcomes**, generate **patient's treatment reports**, share data insurance companies, medical aid services…
- **Public sector**
  - Used by government agencies **to maintain and analyse tax records**, health policy records, … for every citizen.
- **Investment and insurance sector**
  - To analyse data patterns, **customer trends**, and to track market movements.
- **Retain chain**
  - Used for distribution and marketing, in order to track items, **customer buying pattern**, promotions and determining pricing policy.
- **Telecommunication**
  - Used for product promotions, **sales decisions**, and distribution decisions.
- **Tourism industry**
  - **To design and estimate advertising and promotion campaigns,** where clients are targeted based on their feedback and travel patterns.

Steps to implement a DW

Step	Tasks	Deliverables	
1	Need to define project scope	Scope Definition	Business scope
2	Need to determine business needs	Logical Data Model	Modelling
3	Define Operational Datastore requirements	Operational Data Store Model	
4	Acquire or develop Extraction tools	Extract tools and Software	ETL
5	Define Data Warehouse Data requirements	Transition Data Model	Implementation & Mapping
6	Document missing data	To Do Project List	
7	Maps Operational Data Store to Data Warehouse	D/W Data Integration Map	
8	Develop Data Warehouse Database design	D/W Database Design	
9	Extract Data from Operational Data Store	Integrated D/W Data Extracts	Loading data
10	Load Data Warehouse	Initial Data Load	
11	Maintain Data Warehouse	On-going Data Access and Subsequent Loads	Maintenance

Why we need DW? Advantages

- Allow business users to **quickly access critical data from several sources** in a single place.
- **Provide consistent information** on various cross-functional activities.
- **Reduce total time for analysis and reporting**.
- **Store a large amount of historical data** (help user to analyse trends and make future predictions).

Why we need DW? Disadvantages

- **Not an ideal option** for unstructured data.
- **Time consuming**.
- **Can be outdated** relatively quickly.
- **Difficult to make changes** in data types, data source schema, indexes, and queries.
- **Complex** for the average users.

To sum up

- DW is a **central repository** to **quickly access critical data** from several sources.
- **4 main layers:** data source, staging layer (ETL), storage layer (DW per se), and end-user tools (reports, data mining, …).
- **DW can be used in any sector:** industry, research, academy, …

Database vs Data Warehouse

Key Differences

DB & DW

- Database (DB)
  - A collection of **related data** which represents some elements of the real world.
  - It is designed to be built and populated with data **for a specific task**.
- Data Warehouse (DW)
  - **Stores historical data** in order **to analyse, report**, integrate transaction data from different sources.
  - Eases the analysis and reporting process of an organization **for decision making and forecasting process**.

Why using a Database?

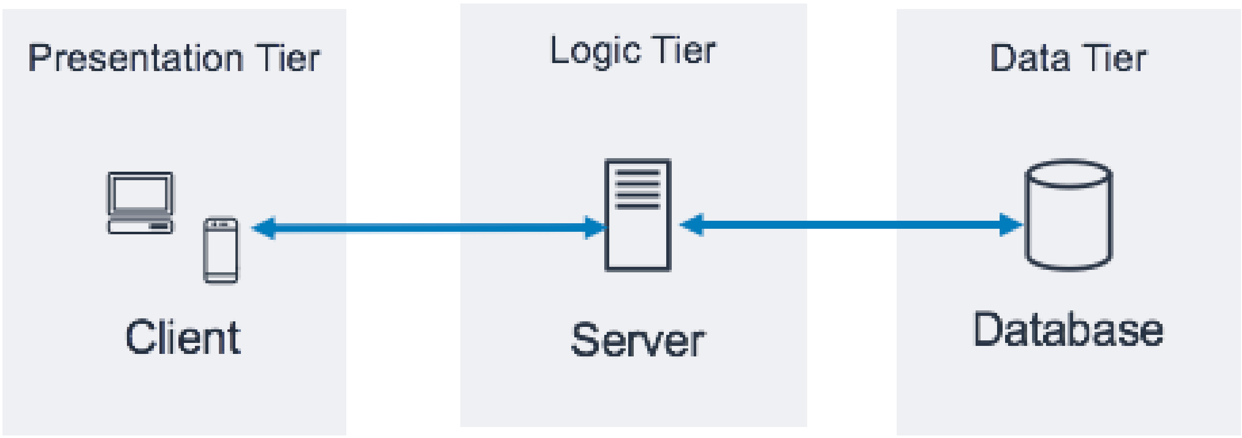
- **Security of data and its access**:
  - **DBMS** offers integrity constraints to get a high level of protection to prevent access to prohibited data.
- Offers a **variety of techniques to store and retrieve data** (CRUD).
- Allows you to **access concurrent data** in such a way that only a single user can access the same data at a time.

Why using a Data Warehouse?

- **Helps business users** to access critical data from several sources in a single place.
- Helps you to **reduce time for analysis and reporting**.
- Allows you to **stores a large amount of historical data to analyse trends and make predictions**.
- **Separates analytics processing from transactional databases**, improving the performance of both systems.
- **Enhances the value of operational business applications** and customer relationship management systems.

Characteristics of a Database

- Offers **security** and **removes redundancy**.
- Allows multiple views of the data.
- Database system follows the **ACID compliance** (Atomicity, Consistency, Isolation, and Durability).
- Allows **separation between programs and data** (3-tier architecture).



Characteristics of Data Warehouse

- A DW is **subject oriented** as it offers information related to themes instead of companies' ongoing operations.
- **The data also needs to be stored in common and unanimously acceptable manner**.
- The **time span for the DW is relatively extensive** compared with other operational systems:
  - A DW is **non-volatile**, which means the previous data is not erased when new information is entered in it.

DB vs DW

Parameter	Database	Data Warehouse
Purpose	Is designed to record	Is designed to analyze
Processing Method	The database uses the Online Transactional Processing (OLTP)	Data warehouse uses Online Analytical Processing (OLAP)
Usage	The database helps to perform fundamental operations for your business	Data warehouse allows you to analyze your business.
Tables and Joins	Tables and joins of a database are complex as they are normalized.	Table and joins are simple in a data warehouse because tl are denormalized.
Orientation	Is an application-oriented collection of data	It is a subject-oriented collection of data
Storage limit	Generally limited to a single application	Stores data from any number of applications
Availability	Data is available real-time	Data is refreshed from source systems as and when neede
Usage	ER modeling techniques are used for designing.	Data modeling techniques are used for designing.
Technique	Capture data	Analyze data
Data Type	Data stored in the Database is up to date.	Current and Historical Data is stored in Data Warehouse. M not be up to date.
Storage of data	Flat Relational Approach method is used for data storage.	Data Ware House uses dimensional and normalized appro for the data structure. Example: Star and snowflake schema.
Query Type	Simple transaction queries are used.	Complex queries are used for analysis purpose.
Data Summary	Detailed Data is stored in a database.	It stores highly summarized data.

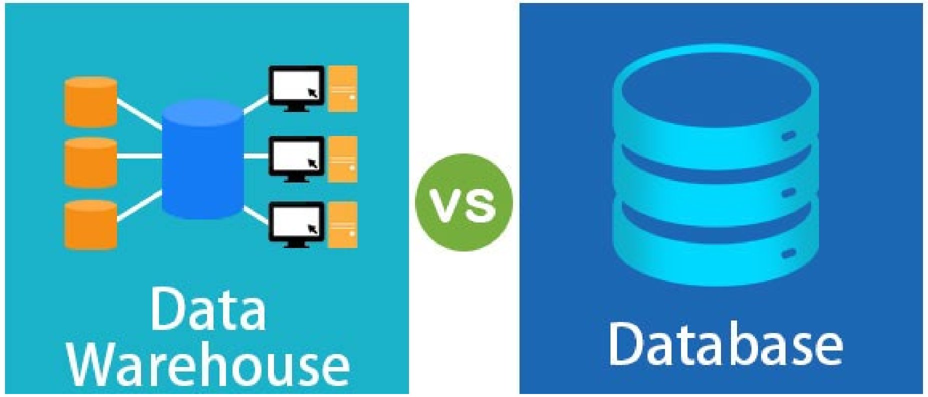
Applications of DB

Sector	Usage
Banking	Use in the banking sector for customer information, account-related activities, payments, deposits, loans, credit cards, etc.
Airlines	Use for reservations and schedule information.
Universities	To store student information, course registrations, colleges, and results.
Telecommunication	It helps to store call records, monthly bills, balance maintenance, etc.
Finance	Helps you to store information related stock, sales, and purchases of stocks and bonds.
Sales & Production	Use for storing customer, product and sales details.
Manufacturing	It is used for the data management of the supply chain and for tracking production of items, inventories status.
HR Management	Detail about employee's salaries, deduction, generation of paychecks, etc.

Applications of DW

Sector	Usage
Airline	It is used for airline system management operations like crew assignment, analyzes of route, frequent flyer program discount schemes for passenger, etc.
Banking	It is used in the banking sector to manage the resources available on the desk effectively.
Healthcare sector	Data warehouse used to strategize and predict outcomes, create patient's treatment reports, etc. Advanced machine learning, big data enable datawarehouse systems can predict ailments.
Insurance sector	Data warehouses are widely used to analyze data patterns, customer trends, and to track market movements quickly.
Retain chain	It helps you to track items, identify the buying pattern of the customer, promotions and also used for determining pricing policy.
Telecommunication	In this sector, data warehouse used for product promotions, sales decisions and to make distribution decisions.

To sum up



DB	DW
• Related data (normalized)	• Heterogenous/redundant data (denormalized)
• Detailed data	• Summarized data (Meta data)
• Transactional process (CRUD)	• Analytical process (CR only)
• Volatile data	• Non-volatile data
• Fundamental operations	• Analyze a business
• Simple queries	• Complex queries
• Application oriented	• Subject oriented