



TS. NGUYỄN QUỐC HÙNG

- Mobile: 0912 251 253
- Email: hungngq@ueh.edu.vn
- Website: <https://bit.ueh.edu.vn/nqhung/>

KHOA HỌC DỮ LIỆU

- Mã học phần: **25D1INF50905948**
- Thời gian: **11/04/2025 - 16/05/2025**
- Hệ: ĐH, Chính quy
- Số lượng: 48 sinh viên
- Số tín chỉ: 2.00

PHƯƠNG PHÁP KHAI THÁC DỮ LIỆU

Thứ 6, thời gian: 12g45-17g05, Giảng đường: N1-303

NỘI DUNG

- **GIỚI THIỆU KHAI THÁC DỮ LIỆU**
 - Khái niệm
 - Ứng dụng
- **CÁC BƯỚC KHAI THÁC DỮ LIỆU**
 - Các bước khai thác dữ liệu
 - Các kỹ thuật khai thác dữ liệu
- **TIỀN XỬ LÝ DỮ LIỆU**
 - Khái niệm
 - Các bước tiền xử lý dữ liệu
- **Thực hành bằng công cụ orange**

KHÁI NIỆM KHAI THÁC DỮ LIỆU

Định nghĩa

Là quá trình trích xuất, khám phá các tri thức từ một lượng dữ liệu lớn.
Là một kỹ năng đa ngành: thống kê, máy học, AI và CSDL

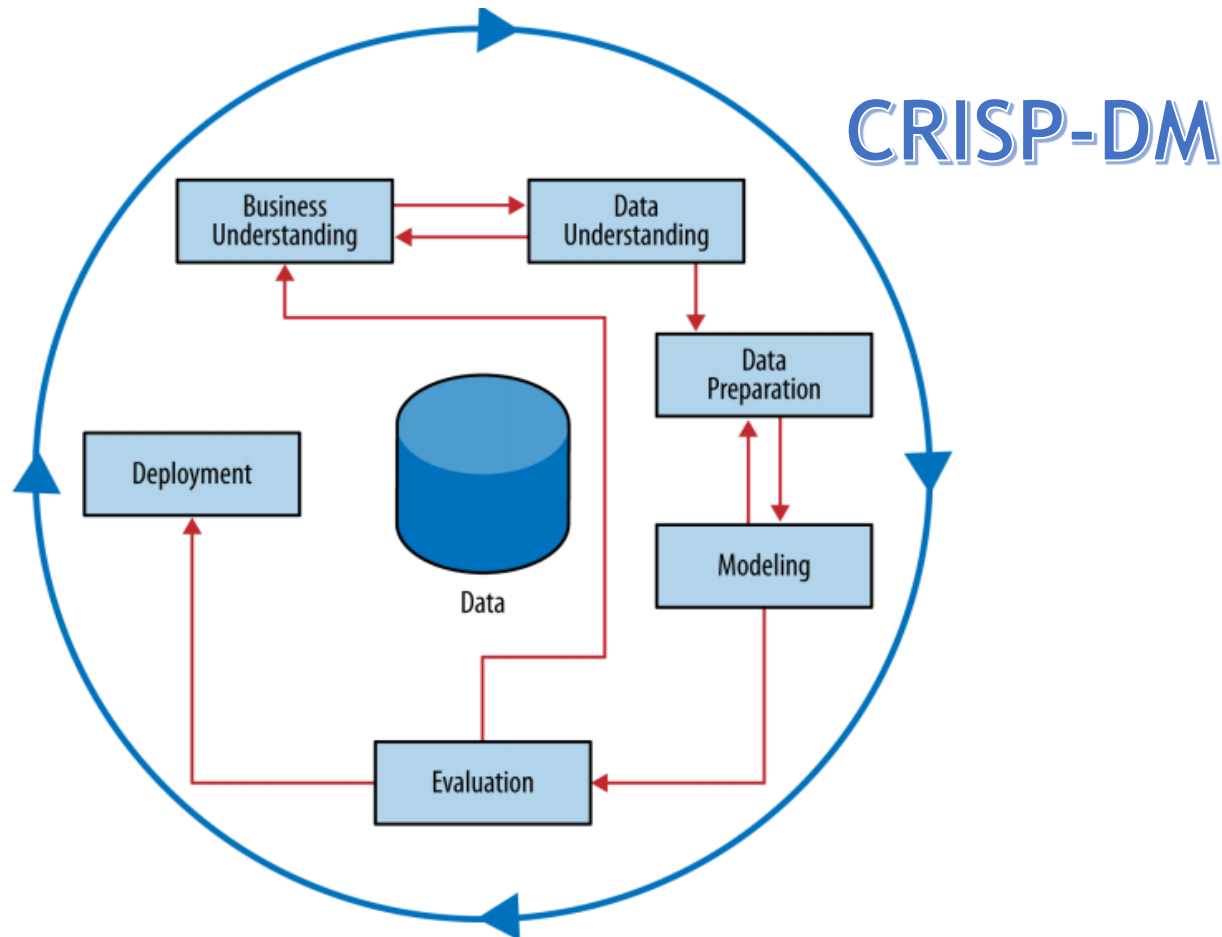
Ứng dụng

- Trong kinh tế: phân tích tình hình tài chính, dự đoán được giá cổ phiếu, phát hiện gian lận, tiếp thị, dự đoán xu hướng thị trường...
- Thống kê, phân tích dữ liệu và hỗ trợ ra quyết định.
- Y học: dựa vào mối liên hệ giữa các triệu chứng để chuẩn đoán bệnh và hướng điều trị.
- Mạng viễn thông: phân tích các cuộc gọi điện thoại để dự đoán hành vi người dùng nhằm nâng cao chất lượng dịch vụ.

NỘI DUNG

- **GIỚI THIỆU KHAI THÁC DỮ LIỆU**
 - Khái niệm
 - Ứng dụng
- **CÁC BƯỚC KHAI THÁC DỮ LIỆU**
 - Các bước khai thác dữ liệu
 - Các kỹ thuật khai thác dữ liệu
- **TIỀN XỬ LÝ DỮ LIỆU**
 - Khái niệm
 - Các bước tiền xử lý dữ liệu
- **Thực hành bằng công cụ orange**

QUY TRÌNH KHAI THÁC DỮ LIỆU



Cross Industry Standard Process for Data Mining (CRISP-DM)

(Chapman et al., 2000)

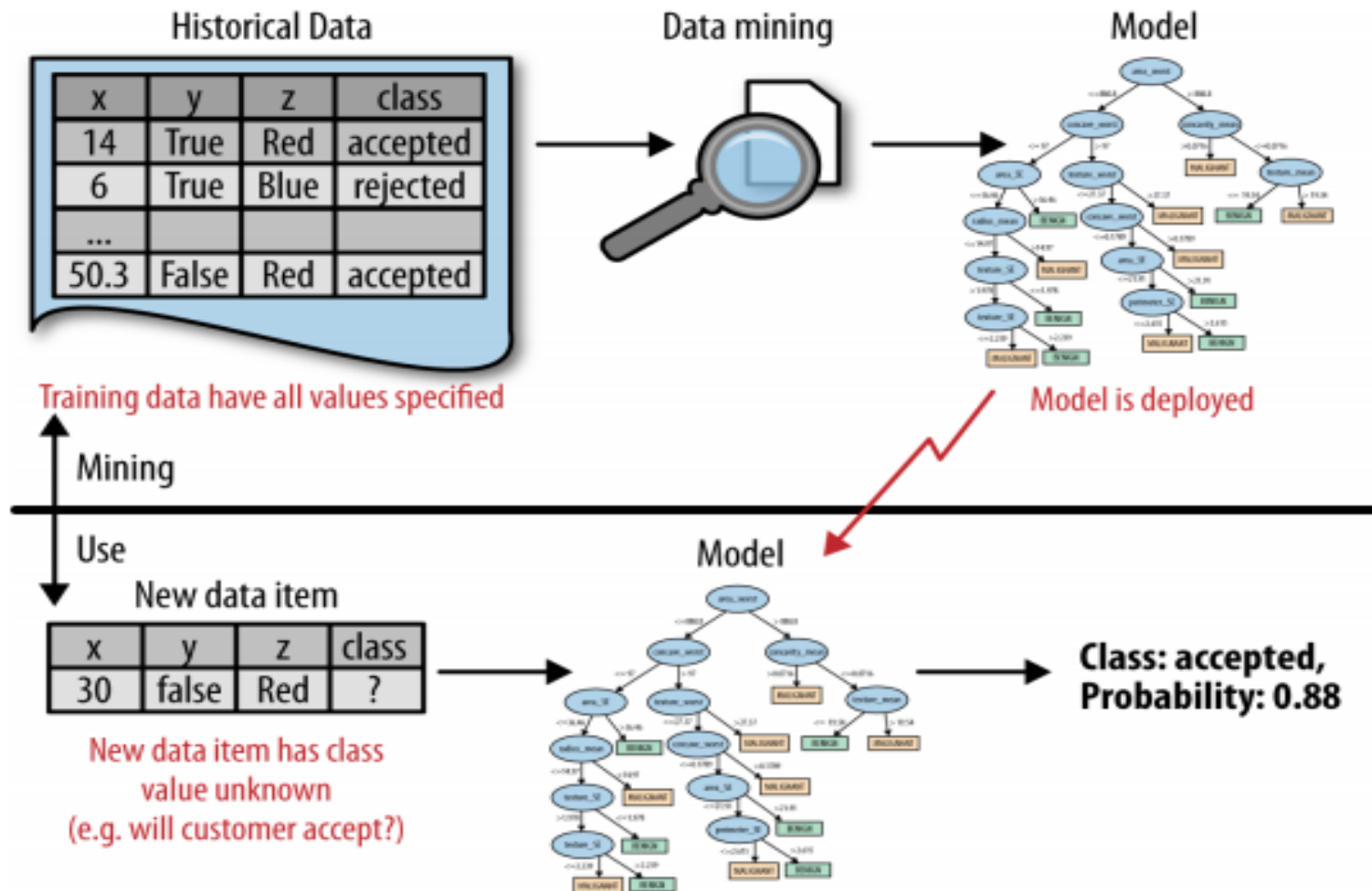
CÁC BƯỚC KHAI THÁC DỮ LIỆU

Các bước tiếp cận vấn đề khai thác dữ liệu trong kinh doanh

1. **Business Understanding:** sử dụng mục tiêu kinh doanh và bối cảnh hiện tại để xác định mục tiêu khai thác dữ liệu
2. **Data Understanding:** kiểm tra tình trạng dữ liệu để xác định dữ liệu đang có liệu có phù hợp với mục tiêu khai thác hay không
3. **Data preparation:** thực hiện các bước tiền xử lý để chuẩn hóa dữ liệu sẵn sàng cho các giai đoạn tiếp theo. Giai đoạn này thường chiếm đến 90% thời gian của cả quy trình.
4. **Mô hình hóa:** sử dụng các mô hình thống kê, máy học để xác định các mẫu/ quy luật của dữ liệu.
5. **Đánh giá:** kiểm tra tính hiệu quả của mô hình có đáp ứng với mục tiêu kinh doanh hay không, có đủ tin cậy hay không
6. **Triển khai:** đưa mô hình giải pháp vào ứng dụng trong các hoạt động của doanh nghiệp



Khai thác dữ liệu và sử dụng kết quả



MỘT SỐ KỸ THUẬT KHAI THÁC DỮ LIỆU

Data mining techniques

Classification

Clustering

Regression

Outer

CÁC DẠNG DỮ LIỆU TRONG THỰC TẾ

► Các tập tin:

- ❑ Text: các loại file văn bản
- ❑ Web data: dữ liệu web
- ❑ Hình ảnh/video
- ❑ Âm thanh

► CSDL quan hệ/không quan hệ

- ❑ SQL
- ❑ MySQL

► Kho dữ liệu

CÁC LOẠI DỮ LIỆU TRONG KHAI THÁC DỮ LIỆU

► Dữ liệu được phân làm hai loại:

□ Định tính (**qualitative/categorical**): Mô tả bằng chuỗi (string)

- Định danh (nominal): dùng mô tả nhãn để phân loại đối tượng
- Nhị phân: nam/nữ, âm tính/dương tính
- Thứ tự: giỏi/khá/trung bình

□ Định lượng (**quantitative/numeric**): Mô tả bằng số

- Rời rạc (Discrete): Nhận những giá trị chắc chắn, rời rạc (có thể đếm được)
- Liên tục (Continuous): Có thể nhận các giá trị bất kỳ trong một khoảng xác định

Ví dụ: Dữ liệu mô tả về hồ ly:

- + Màu: đen, trắng, đỏ...(định tính)
- + Số đuôi: 1, 3, 7, 9 (rời rạc)
- + Cân nặng: 2kg, 2.5kg...(liên tục)

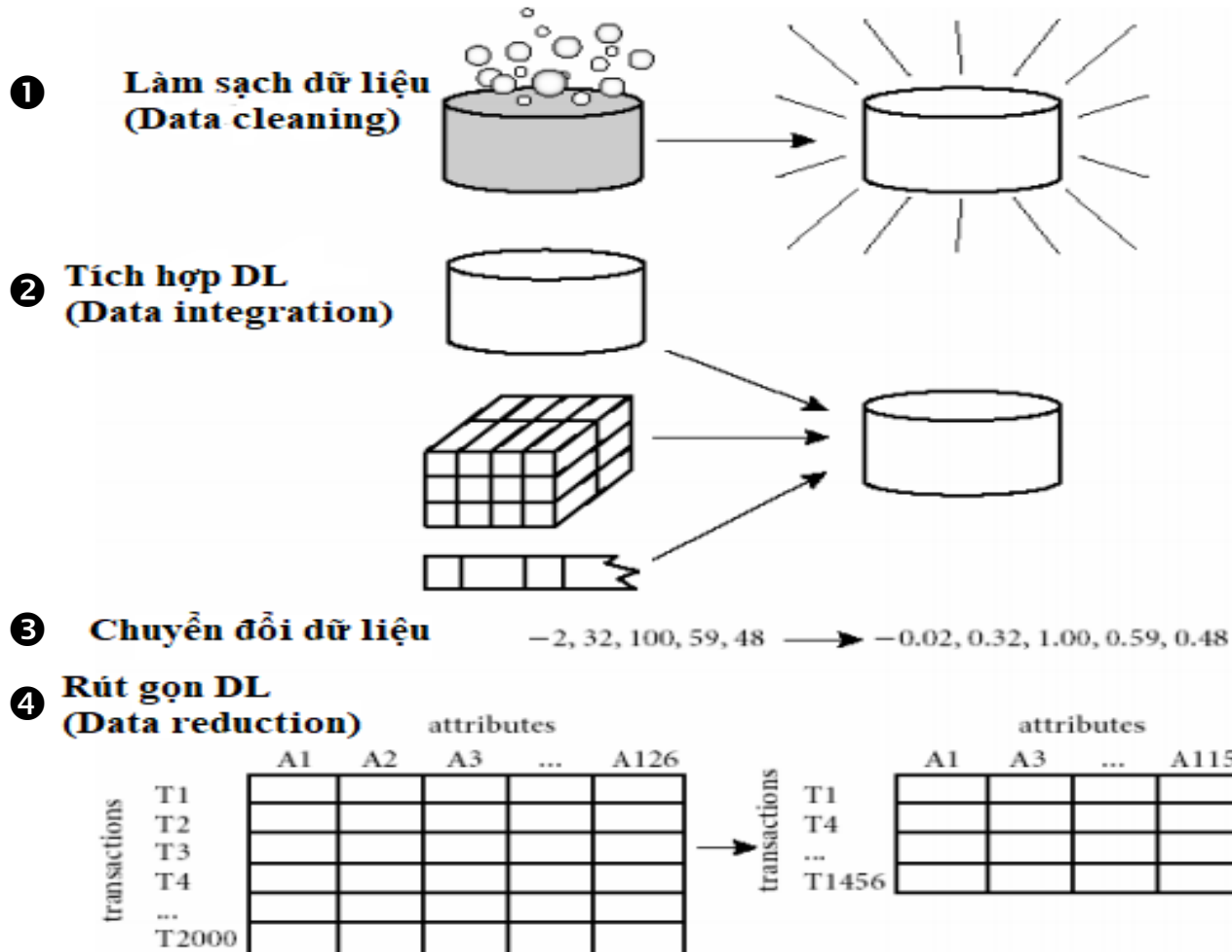
NỘI DUNG

- **GIỚI THIỆU KHAI THÁC DỮ LIỆU**
 - Khái niệm
 - Ứng dụng
- **CÁC BƯỚC KHAI THÁC DỮ LIỆU**
 - Các bước khai thác dữ liệu
 - Các kỹ thuật khai thác dữ liệu
- **TIỀN XỬ LÝ DỮ LIỆU**
 - Khái niệm
 - Các bước tiền xử lý dữ liệu
- **Thực hành bằng công cụ orange**

TIỀN XỬ LÝ DỮ LIỆU

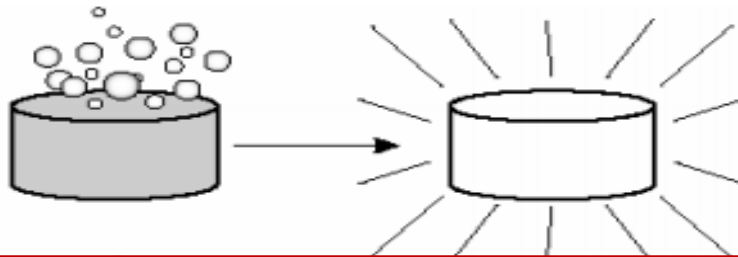
- ▶ Khái niệm: Là quá trình xử lý dữ liệu thô/gốc (raw/original data) nhằm cải thiện chất lượng dữ liệu (quality of the data) và do đó, cải thiện chất lượng của kết quả khai phá.
 - ❑ Dữ liệu thô/gốc: có thể có cấu trúc hoặc không có cấu trúc; nằm ở nhiều định dạng khác nhau (tập tin hoặc CSDL)
 - ❑ Chất lượng dữ liệu (data quality): tính chính xác, tính hiện hành, tính toàn vẹn, tính nhất quán

CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU

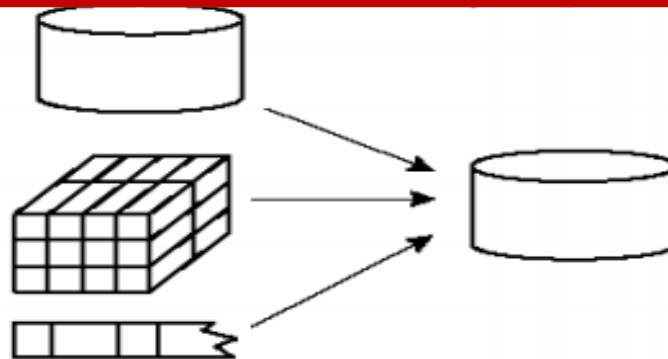


CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU

① Làm sạch dữ liệu (Data cleaning)



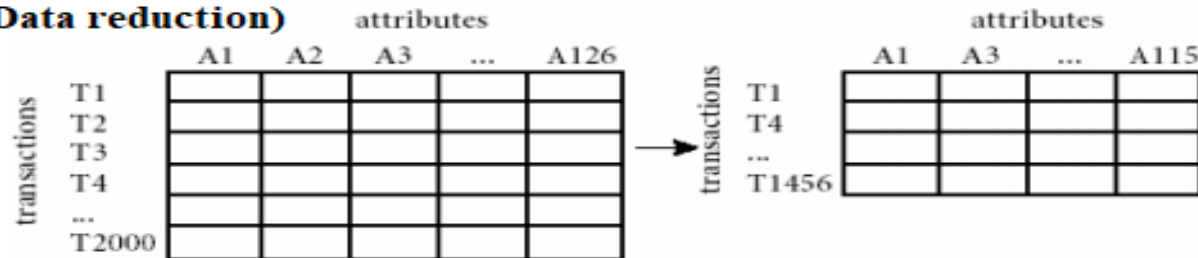
Tích hợp DL (Data integration)



Chuyển đổi dữ liệu

→ -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Rút gọn DL (Data reduction)



LÀM SẠCH DỮ LIỆU

- ▶ Làm sạch dữ liệu (data cleaning/cleansing): loại bỏ nhiễu (remove noise), hiệu chỉnh những phần dữ liệu không nhất quán (correct data inconsistencies)
- ▶ Bao gồm:
 - ❑ Xử lý dữ liệu bị thiếu (missing data)
 - ❑ Xử lý dữ liệu bị nhiễu (noisy data)
 - ❑ Xử lý dữ liệu không nhất quán
 - ❑ Tóm tắt hóa dữ liệu

XỬ LÝ DỮ LIỆU BỊ THIẾU (missing data)

- ▶ Là dữ liệu không có sẵn, không đủ khi cần sử dụng
- ▶ Nguyên nhân:
 - ❑ Khách quan (không tồn tại lúc được nhập liệu, sự cố, ...)
 - ❑ Chủ quan (tác nhân con người)
- ▶ Giải pháp cho dữ liệu bị thiếu
 - Bỏ qua
 - Xử lý tay (không tự động, bán tự động)
 - Dùng giá trị thay thế (tự động): hằng số toàn cục, trị phổ biến nhất, trung bình toàn cục, trung bình cục bộ, trị dự đoán, ...
 - Ngăn chặn dữ liệu bị thiếu: thiết kế tốt CSDL và các thủ tục nhập liệu (các ràng buộc dữ liệu)

XỬ LÝ DỮ LIỆU NHIỀU

- ▶ Bao gồm: nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
- ▶ Định nghĩa:
 - ❑ Outliers: những dữ liệu (đối tượng) không tuân theo đặc tính/hành vi chung của tập dữ liệu (đối tượng).
 - ❑ Noisy data: bị loại bỏ (rejected/discarded outliers) như là những trường hợp ngoại lệ (exceptions).
- ▶ Nguyên nhân:
 - ❑ Khách quan (công cụ thu thập dữ liệu, lỗi trên đường truyền, giới hạn công nghệ, ...)
 - ❑ Chủ quan (tác nhân con người)

XỬ LÝ DỮ LIỆU NHIỀU

- ▶ Giải pháp nhận diện phần tử biên
 - ❑ Dựa trên phân bố thống kê (statistical distribution-based)
 - ❑ Dựa trên khoảng cách (distance-based)
 - ❑ Dựa trên mật độ (density-based)
 - ❑ Dựa trên độ lệch (deviation-based)
- ▶ Giải pháp giảm thiểu nhiễu
 - ❑ Phân giỏ (binning)
 - ❑ Hồi quy (regression)
 - ❑ Phân tích cụm (cluster analysis)

XỬ LÝ DỮ LIỆU KHÔNG NHẤT QUÁN

► Định nghĩa:

- ❑ Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể. VD:
- ❑ Dữ liệu được ghi nhận không phản ánh đúng ngữ nghĩa cho các đối tượng/thực thể: VD: ràng buộc khóa ngoại

► Nguyên nhân:

- ❑ Sự không nhất quán trong các qui ước đặt tên hay mã dữ liệu
- ❑ Định dạng không nhất quán của các vùng nhập liệu.
- ❑ Thiết bị ghi nhận dữ liệu

XỬ LÝ DỮ LIỆU KHÔNG NHẤT QUÁN

► Giải pháp:

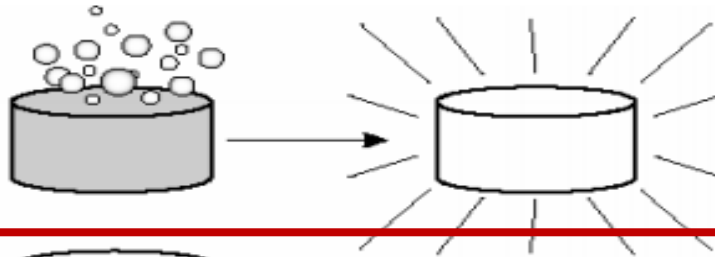
- ❑ Tận dụng siêu dữ liệu, ràng buộc dữ liệu, sự kiểm tra của nhà phân tích dữ liệu cho việc nhận diện
- ❑ Điều chỉnh dữ liệu không nhất quán bằng tay.
- ❑ Các giải pháp biến đổi/chuẩn hóa dữ liệu tự động

TÓM TẮT HÓA DỮ LIỆU (nghiên cứu thêm)

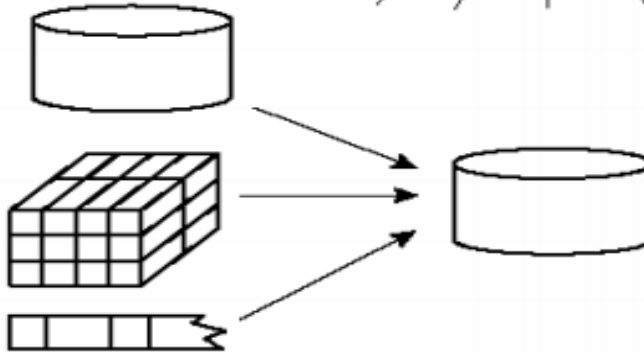
- ▶ Xác định các thuộc tính (properties) tiêu biểu của dữ liệu về xu hướng chính (central tendency) và sự phân tán (dispersion) của dữ liệu.
 - Các độ đo về xu hướng chính: mean, median, mode, midrange...
 - Các độ đo về sự phân tán: quartiles, interquartile range (IQR), variance
- ▶ Nhận diện dữ liệu nổi bật/hiếm: nhiễu (noise) hoặc phần tử biên (outliers), cung cấp cái nhìn tổng quan về dữ liệu

CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU

**Làm sạch dữ liệu
(Data cleaning)**



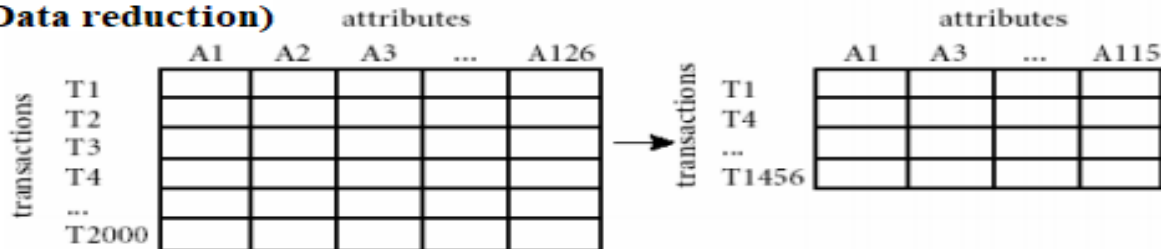
**② Tích hợp DL
(Data integration)**



Chuyển đổi dữ liệu

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

**Rút gọn DL
(Data reduction)**



TÍCH HỢP DỮ LIỆU

- ▶ Tích hợp dữ liệu (data integration): trộn dữ liệu (merge data) từ nhiều nguồn khác nhau vào một kho dữ liệu
- ▶ Bao gồm:
 - Vấn đề nhận dạng thực thể
 - Tích hợp lược đồ (schema integration)
 - So trùng đối tượng (object matching)
 - Vấn đề dư thừa (redundancy)
 - Phát hiện và xử lý mâu thuẫn giá trị dữ liệu (detection and resolution of data value conflicts)

VẤN ĐỀ NHẬN DẠNG THỰC THỂ

- ▶ Các thực thể (object/entity/attribute) đến từ nhiều nguồn dữ liệu.
 - ▶ Hai hay nhiều thực thể khác nhau diễn tả cùng một thực thể thực.
 - ❑ Ví dụ ở mức lược đồ (schema): customer_id trong nguồn S1 và cust_number trong nguồn S2.
 - ❑ Ví dụ ở mức thể hiện (instance): “R & D” trong nguồn S1 và “Research & Development” trong nguồn S2. “Male” và “Female” trong nguồn S1 và “Nam” và “Nữ” trong nguồn S2.
- Vai trò của siêu dữ liệu (metadata)

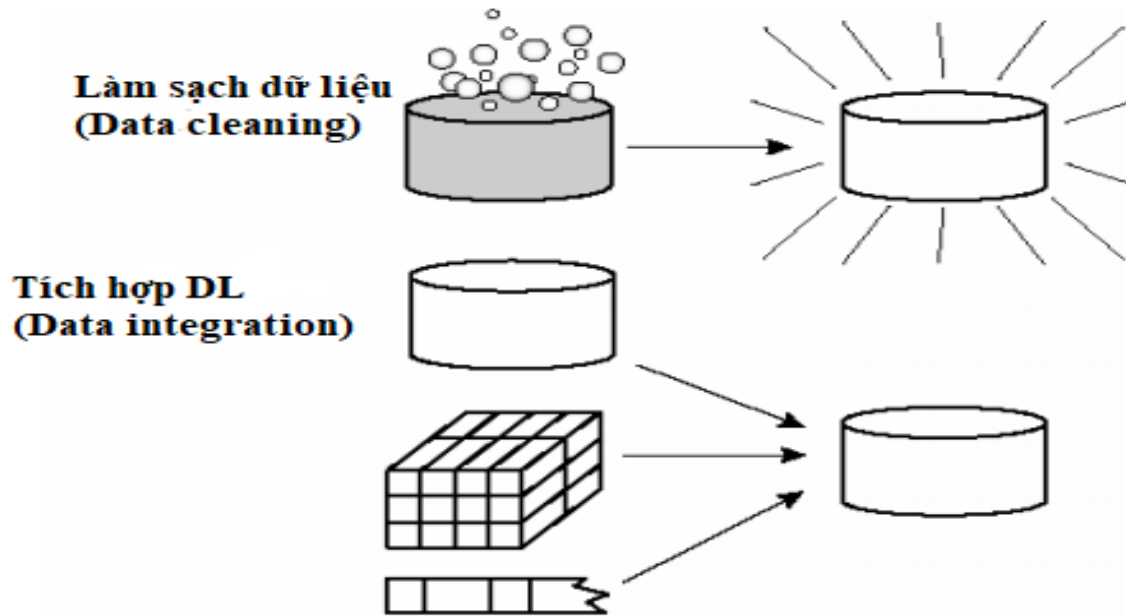
VẤN ĐỀ DƯ THỪA

- ▶ Hiện tượng: giá trị của một thuộc tính có thể được dẫn ra/tính từ một/nhiều thuộc tính khác, vấn đề trùng lặp dữ liệu (duplication).
- ▶ Nguyên nhân: tổ chức dữ liệu kém, không nhất quán trong việc đặt tên chiều/thuộc tính.
- ▶ Phát hiện dư thừa: phân tích tương quan (correlation analysis)
 - ❑ Dựa trên dữ liệu hiện có, kiểm tra khả năng dẫn ra một thuộc tính B từ thuộc tính A.
 - ❑ Đối với các thuộc tính số (numerical attributes), đánh giá tương quan giữa hai thuộc tính với các hệ số tương quan (correlation coefficient, aka Pearson's product moment coefficient).
 - ❑ Đối với các thuộc tính rời rạc (categorical/discrete attributes), đánh giá tương quan giữa hai thuộc tính với phép kiểm thử (chi bình phương)

VẤN ĐỀ MÂU THUẦN GIÁ TRỊ DỮ LIỆU

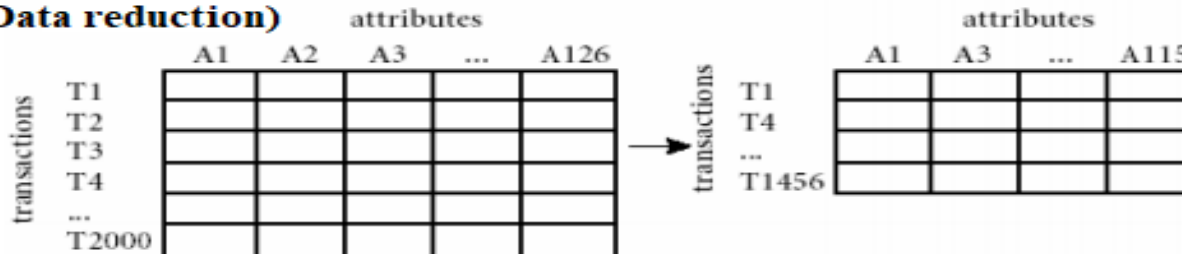
- ▶ Cho cùng một thực thể thật, các giá trị thuộc tính đến từ các nguồn dữ liệu khác nhau có thể khác nhau về cách biểu diễn (representation), đo lường (scaling), và mã hóa (encoding).
 - ❑ Mâu thuẫn do định dạng: dd/mm/yyyy vs. mm/dd/yyyy
 - ❑ Mâu thuẫn do đơn vị: gram vs. kg
 - ❑ Mâu thuẫn do mã hóa: “yes” và “no” với “1” và “0”.

CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU



③ **Chuyển đổi dữ liệu** $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Rút gọn DL (Data reduction)



CHUYỂN ĐỔI DỮ LIỆU

- ▶ Chuyển đổi dữ liệu (data transformation): chuẩn hoá dữ liệu (data normalization)
- ▶ Bao gồm:
 - ❑ Làm trơn dữ liệu (smoothing)
 - ❑ Kết hợp dữ liệu (aggregation)
 - ❑ Tổng quát hóa dữ liệu (generalization)
 - ❑ Chuẩn hóa dữ liệu (normalization)
 - ❑ Xây dựng thuộc tính (attribute/feature construction)

LÀM TRƠN DỮ LIỆU

- ▶ Các phương pháp binning (bin means, bin medians, bin boundaries)
- ▶ Hồi quy
- ▶ Các kỹ thuật gom cụm (phân tích phần tử biên)
- ▶ Các phương pháp rời rạc hóa dữ liệu

KẾT HỢP DỮ LIỆU

- ▶ Các tác vụ kết hợp/tóm tắt dữ liệu
- ▶ Chuyển dữ liệu ở mức chi tiết này sang dữ liệu ở mức kém chi tiết hơn
- ▶ Hỗ trợ việc phân tích dữ liệu ở nhiều độ mịn thời gian khác nhau

TỔNG QUÁT HÓA

- ▶ Chuyển đổi dữ liệu cấp thấp/nguyên tố/thô sang các khái niệm ở mức cao hơn thông qua các phân cấp ý niệm

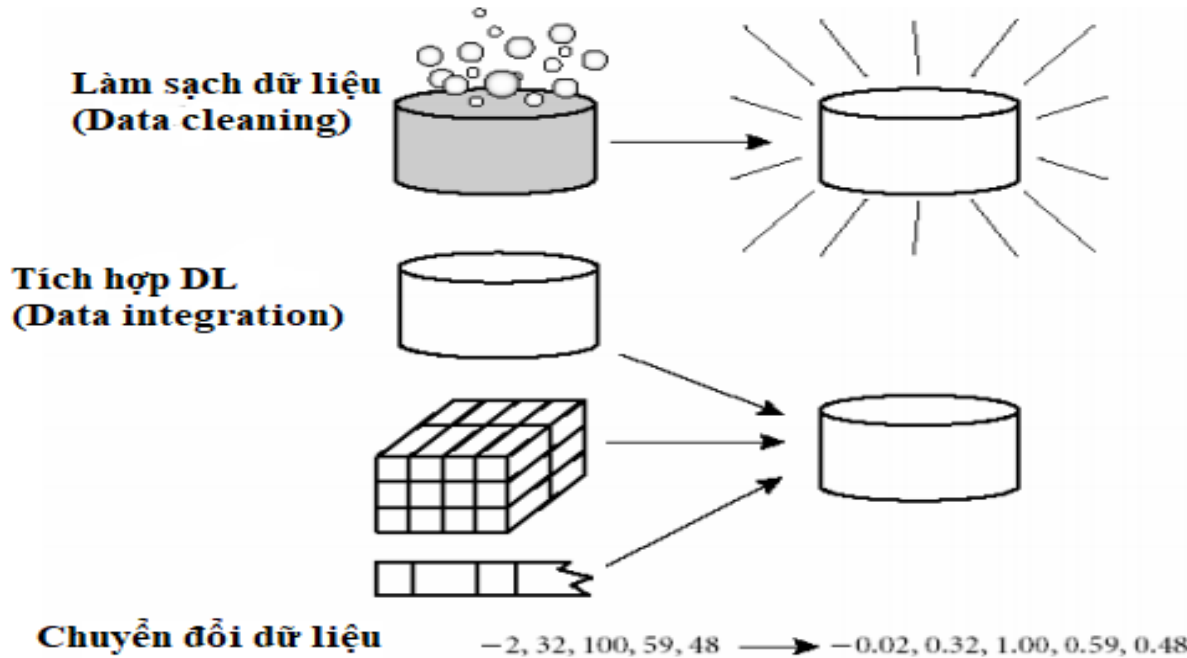
CHUẨN HÓA DỮ LIỆU

- ▶ min-max normalization
 - ▶ z-score normalization
 - ▶ Normalization by decimal scaling
- Các giá trị thuộc tính được chuyển đổi vào một miền trị nhất định được định nghĩa trước

XÂY DỰNG THUỘC TÍNH

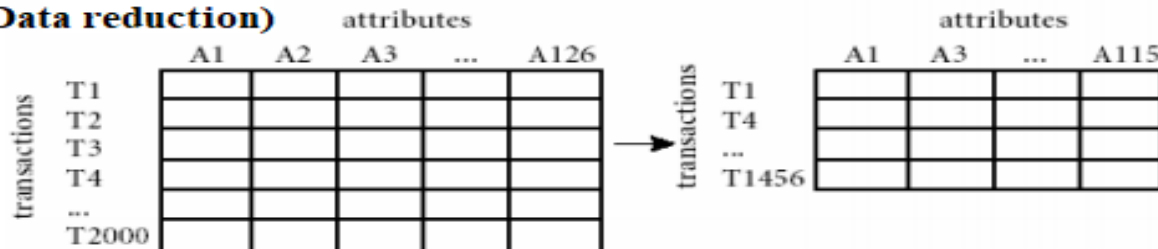
- ▶ Các thuộc tính mới được xây dựng và thêm vào từ tập các thuộc tính sẵn có.
- ▶ Hỗ trợ kiểm tra tính chính xác và giúp hiểu cấu trúc của dữ liệu nhiều chiều.
- ▶ Hỗ trợ phát hiện thông tin thiếu sót về các mối quan hệ giữa các thuộc tính dữ liệu.

CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU



④

Rút gọn DL (Data reduction)



RÚT GỌN DỮ LIỆU

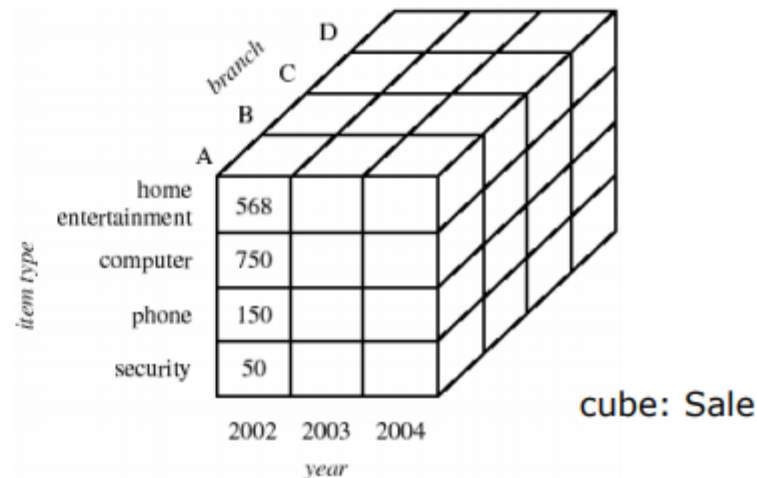
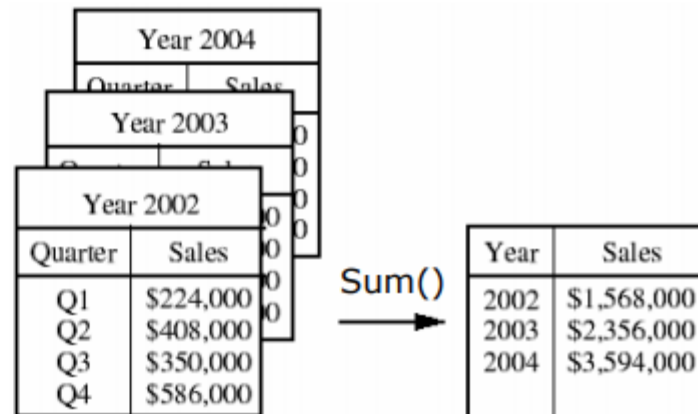
- ▶ Rút gọn dữ liệu (data reduction): thu giảm kích thước dữ liệu (nghĩa là giảm số phần tử) bằng kết hợp dữ liệu (data aggregation), loại bỏ các đặc điểm dư thừa (redundant features) (nghĩa là giảm số chiều/thuộc tính dữ liệu), gom cụm dữ liệu
- ▶ Bao gồm:
 - ❑ Kết hợp khối dữ liệu (data cube aggregation)
 - ❑ Chọn tập con các thuộc tính (attribute subset selection)
 - ❑ Thu giảm chiều (dimensionality reduction)
 - ❑ Thu giảm lượng (numerosity reduction)
 - ❑ Tạo phân cấp ý niệm (concept hierarchy generation)
 - ❑ Rời rạc hóa (discretization)

KẾT HỢP KHỐI DỮ LIỆU

❑ Kết hợp khối dữ liệu

(data cube aggregation)

- Dạng dữ liệu: additive, semi-additive (numerical)
- Kết hợp dữ liệu bằng các hàm nhóm: average, min, max, sum, count, ...
- Dữ liệu ở các mức trừu tượng khác nhau.
- Mức trừu tượng càng cao giúp thu giảm lượng dữ liệu càng nhiều.



CHỌN TẬP CON THUỘC TÍNH

- ▶ Giảm kích thước tập dữ liệu bằng việc loại bỏ những thuộc tính/chiều/đặc trưng dư thừa/không thích hợp (redundant/irrelevant).
- ▶ Mục tiêu: tập ít các thuộc tính nhất vẫn đảm bảo phân bố xác suất (probability distribution) của các lớp dữ liệu đạt được gần với phân bố xác suất ban đầu với tất cả các thuộc tính

THU GIẢM CHIỀU

- ▶ Biến đổi wavelet (wavelet transforms)
- ▶ Phân tích nhân tố chính (principal component analysis)

THU GIẢM LƯỢNG (numerosity reduction)

- ▶ Các kỹ thuật giảm lượng dữ liệu bằng các dạng biểu diễn dữ liệu thay thế.
- ▶ Các phương pháp có thông số (parametric): mô hình ước lượng dữ liệu
 - Hồi quy
- ▶ Các phương pháp phi thông số (nonparametric): lưu trữ các biểu diễn thu giảm của dữ liệu
 - Histogram, Clustering, Sampling

RỜI RẠC HÓA DỮ LIỆU

- ▶ Giảm số lượng giá trị của một thuộc tính liên tục (continuous attribute) bằng các chia miền trị thuộc tính thành các khoảng (intervals)
- ▶ Các nhãn (labels) được gán cho các khoảng (intervals) này và được dùng thay giá trị thực của thuộc tính
- ▶ Các trị thuộc tính có thể được phân hoạch theo một phân cấp (hierarchical) hay ở nhiều mức phân giải khác nhau (multiresolution)

RỜI RẠC HÓA DỮ LIỆU

- ▶ Thu giảm số trị của một thuộc tính liên tục (continuous attribute) bằng cách chia miền trị thành các khoảng (interval) có dán nhãn. Các nhãn này được dùng thay cho các giá trị thực.
- ▶ Tiến hành theo hai cách: trên xuống (top down) và dưới lên (bottom up), có giám sát (supervised) và không có giám sát (unsupervised).
- ▶ Tạo phân hoạch phân cấp/đa phân giải (multiresolution) trên các trị thuộc tính

TẠO CÂY PHÂN CẤP KHÁI NIỆM

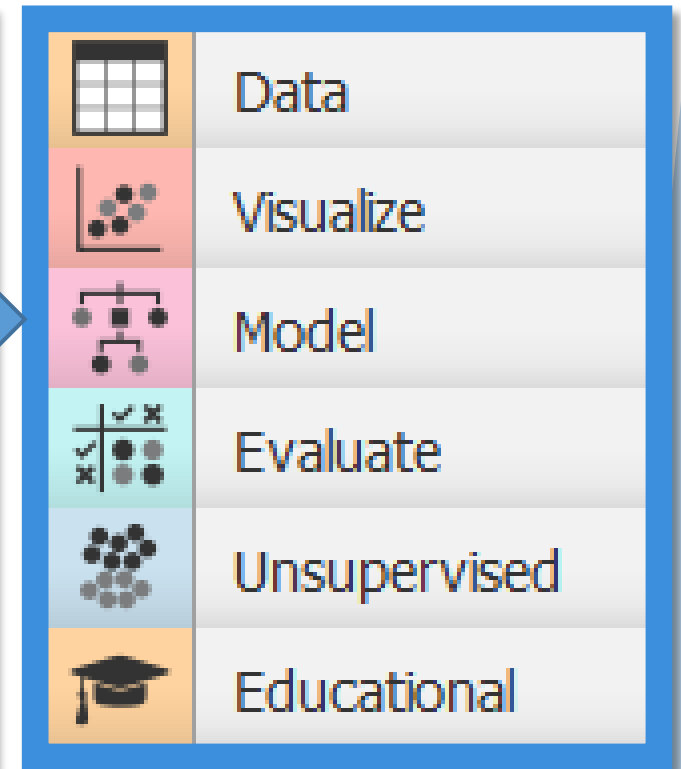
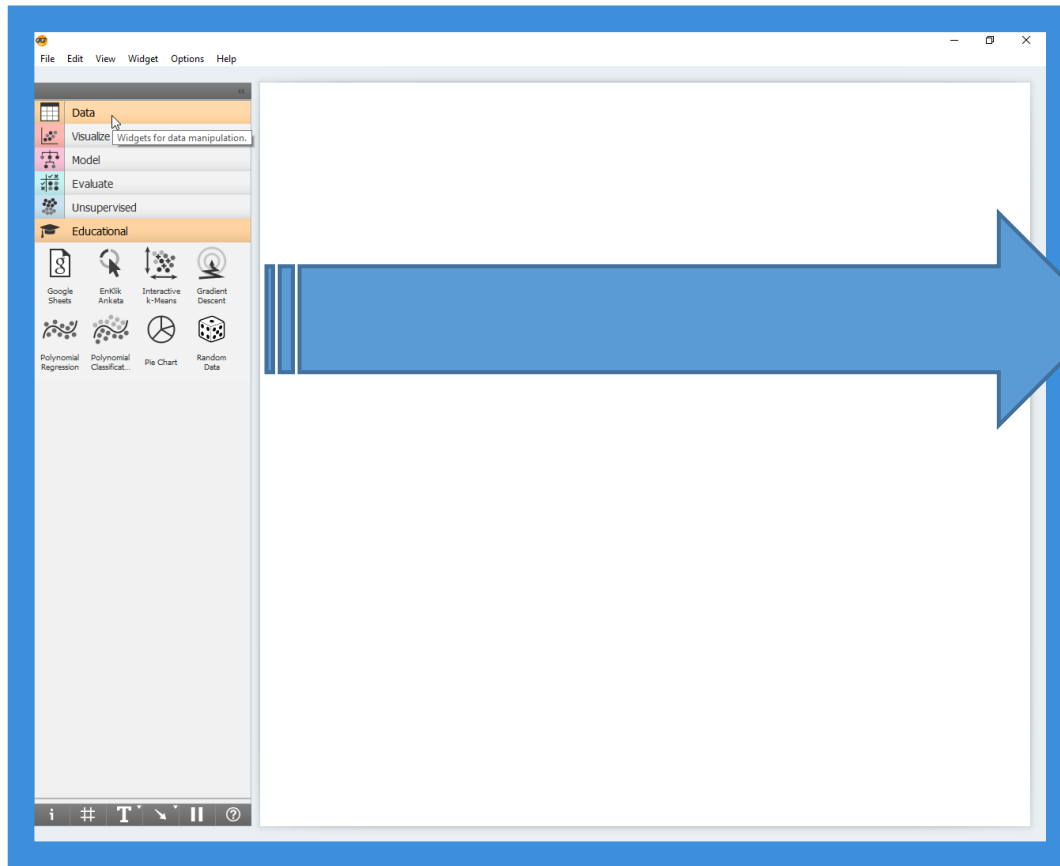
- ▶ Hỗ trợ khai phá dữ liệu ở nhiều mức trừu tượng
- ▶ Cho thuộc tính số (numerical attributes): binning, histogram analysis, entropy-based discretization, χ^2 -merging, cluster analysis, discretization by intuitive partitioning
- ▶ Cho thuộc tính phân loại/rời rạc (categorical/discrete attributes): chỉ định tường minh bởi người sử dụng hay chuyên gia, nhóm dữ liệu tường minh, dựa trên số lượng trị phân biệt (khác nhau) của mỗi thuộc tính

THỰC HÀNH TRÊN ORANGE

- ▶ Theo dõi clip hướng dẫn thực hành trên LMS
- ▶ Nạp dữ liệu
- ▶ Quan sát bảng dữ liệu.
- ▶ Visualize dữ liệu.
- ▶ Chuẩn hoá dữ liệu và xử lý dữ liệu bị thiếu.
- ▶ Lưu dữ liệu đã xử lý.

Khởi động phần mềm Orange

► Chọn vào biểu tượng



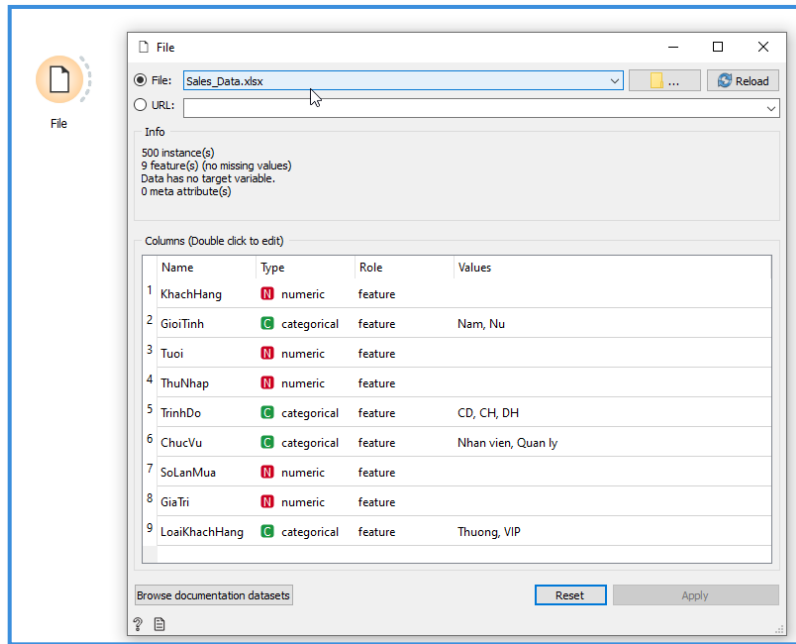
Ví dụ minh họa

► Nạp dữ liệu: có 2 cách

❑ Lưu trữ trên Máy tính



File

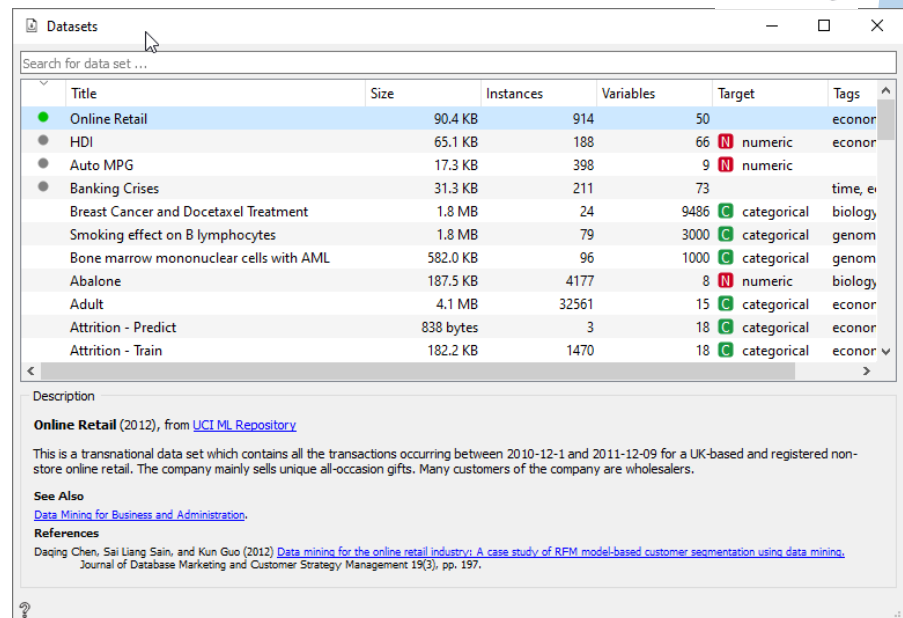


Sales_Data.xlsx

❑ Lưu trữ trên Internet



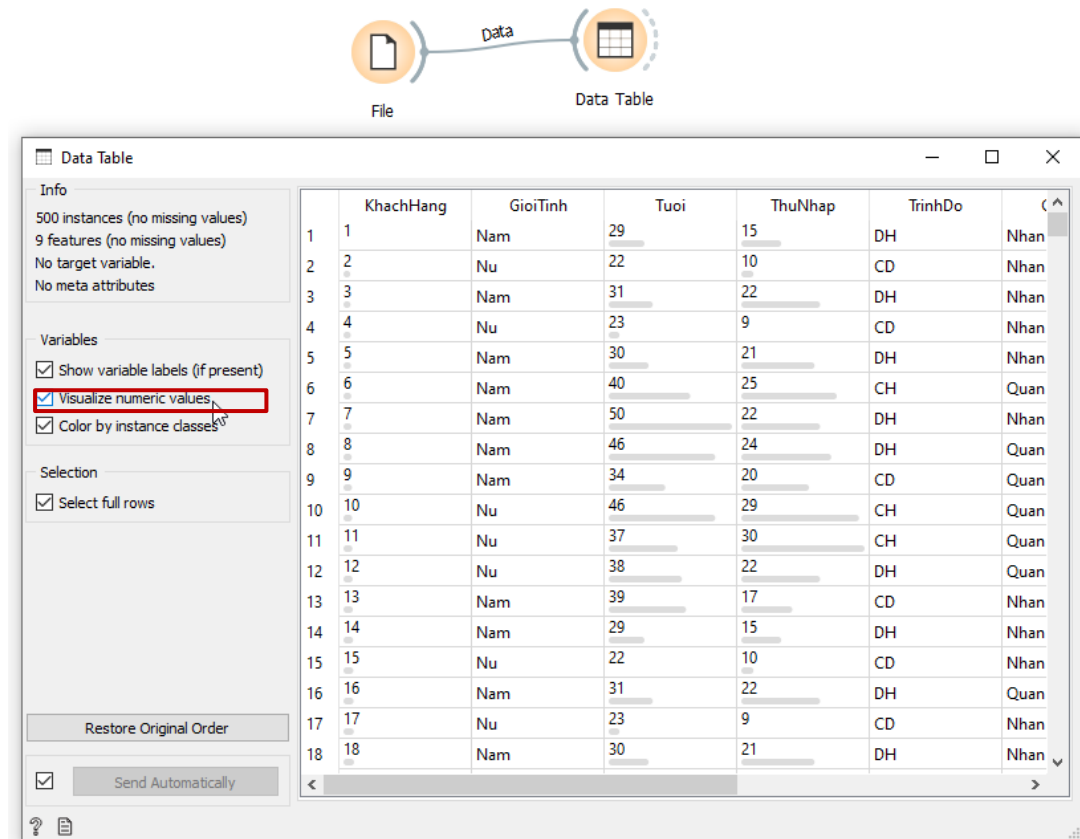
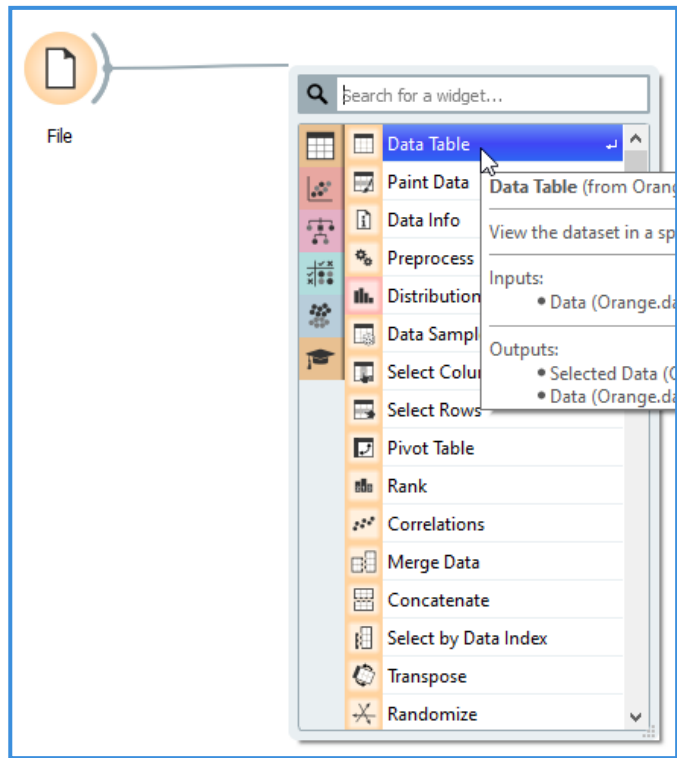
Datasets



Online Retail

Xem thông tin dữ liệu

► Kéo và Chọn chức năng **Data Table**: 

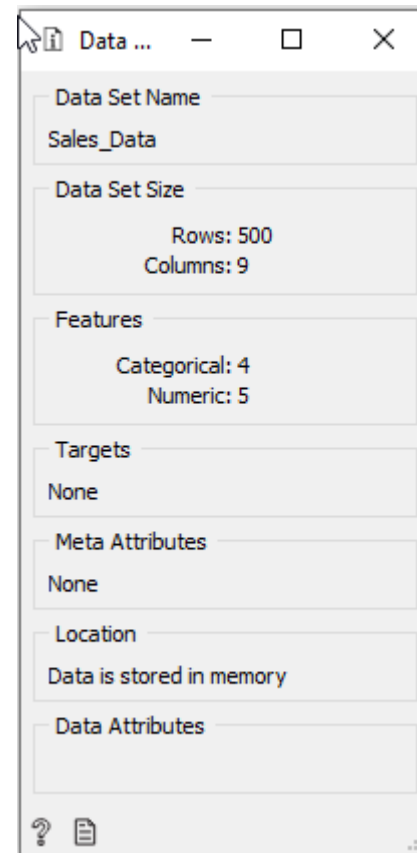
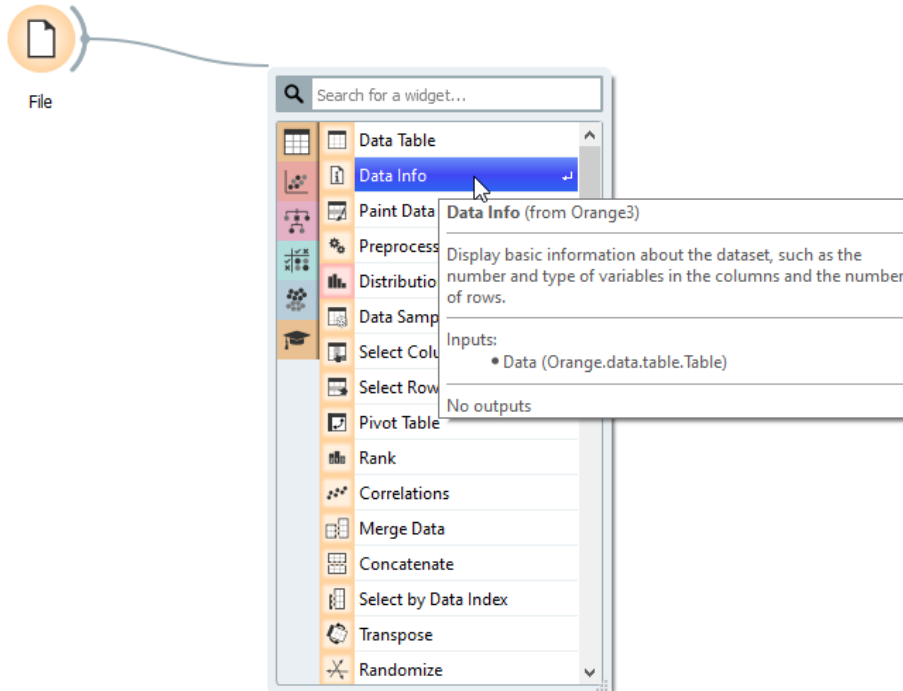


The screenshot shows the Orange3 Data Table widget interface. The 'Info' panel on the left displays the dataset details: 500 instances (no missing values), 9 features (no missing values), No target variable, and No meta attributes. The 'Variables' panel shows checkboxes for 'Show variable labels (if present)', 'Visualize numeric values' (checked), and 'Color by instance classes'. The 'Selection' panel shows a checkbox for 'Select full rows' (checked). The 'Send Automatically' checkbox is also checked. The main table displays the data with columns: KháchHang, GioiTinh, Tuoi, ThuNhap, TrinhDo, and a partially visible column labeled 'Nhan'.

	KhachHang	GioiTinh	Tuoi	ThuNhap	TrinhDo	Nhan
1	1	Nam	29	15	DH	Nhan
2	2	Nu	22	10	CD	Nhan
3	3	Nam	31	22	DH	Nhan
4	4	Nu	23	9	CD	Nhan
5	5	Nam	30	21	DH	Nhan
6	6	Nam	40	25	CH	Quan
7	7	Nam	50	22	DH	Nhan
8	8	Nam	46	24	DH	Quan
9	9	Nam	34	20	CD	Quan
10	10	Nu	46	29	CH	Quan
11	11	Nu	37	30	CH	Quan
12	12	Nu	38	22	DH	Quan
13	13	Nam	39	17	CD	Nhan
14	14	Nam	29	15	DH	Nhan
15	15	Nu	22	10	CD	Nhan
16	16	Nam	31	22	DH	Quan
17	17	Nu	23	9	CD	Nhan
18	18	Nam	30	21	DH	Nhan

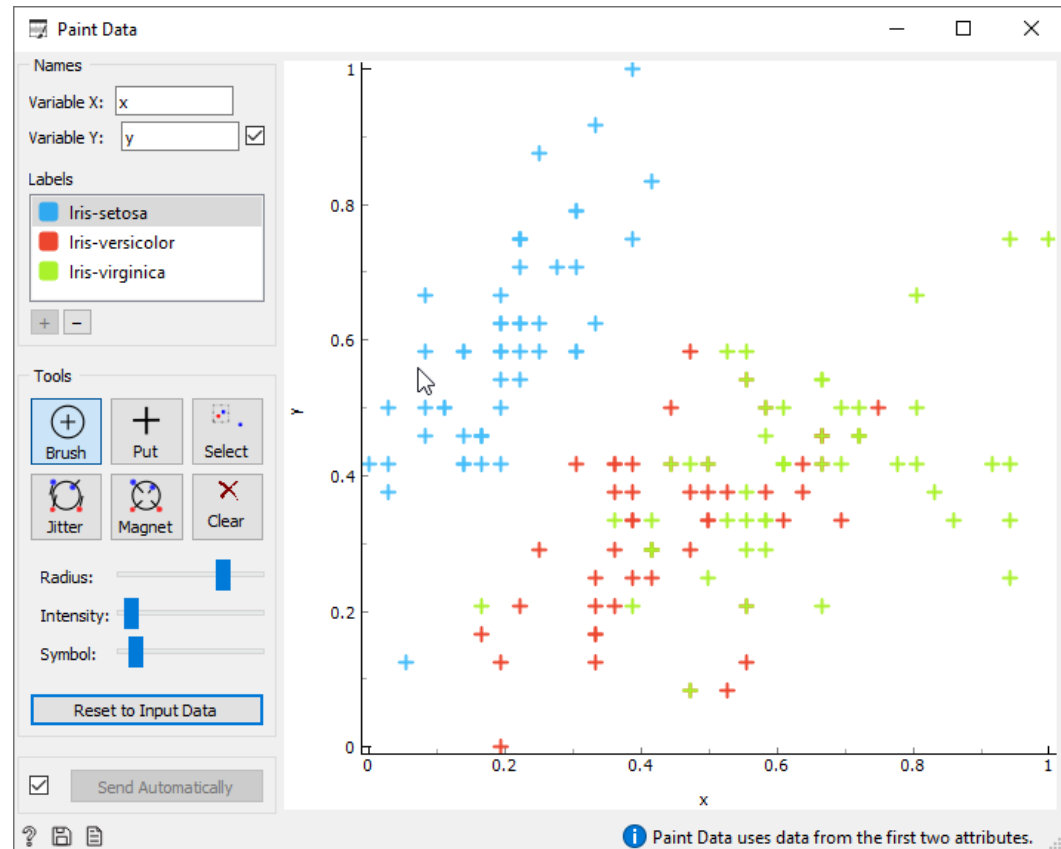
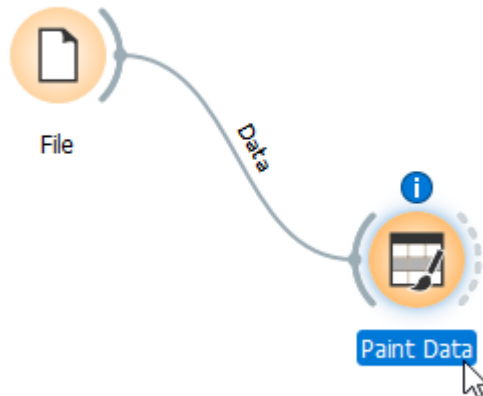
Xem thông tin về dữ liệu

► Kéo và Chọn chức năng **Data Info**:



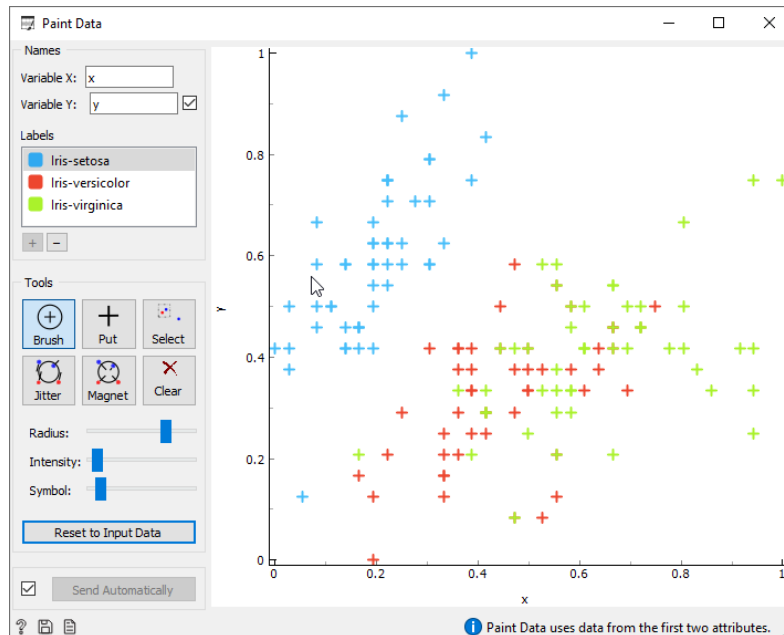
Vẽ biểu đồ về dữ liệu

► Kéo và Chọn chức năng **Paint Data**:

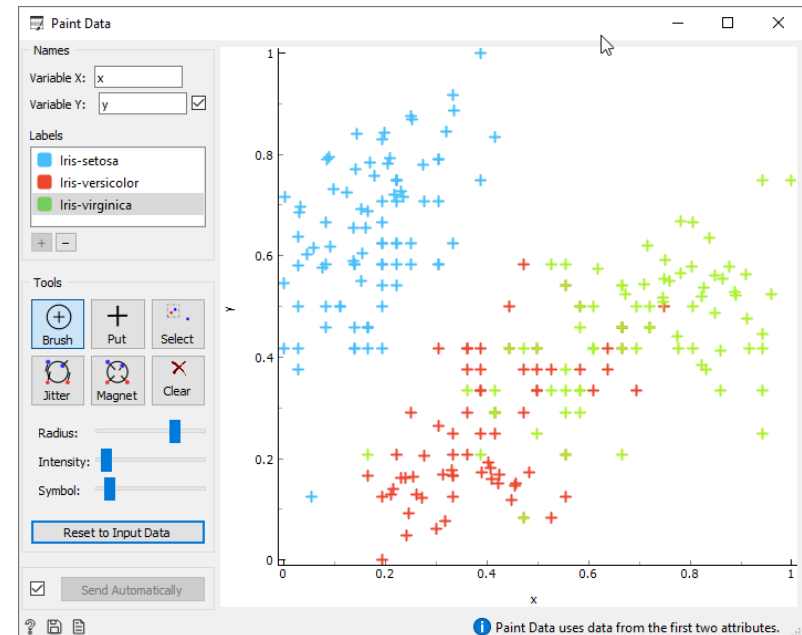


Thêm dữ liệu từ biểu đồ

- Chọn các nhãn, kích chuột lên vùng dữ liệu còn thiếu



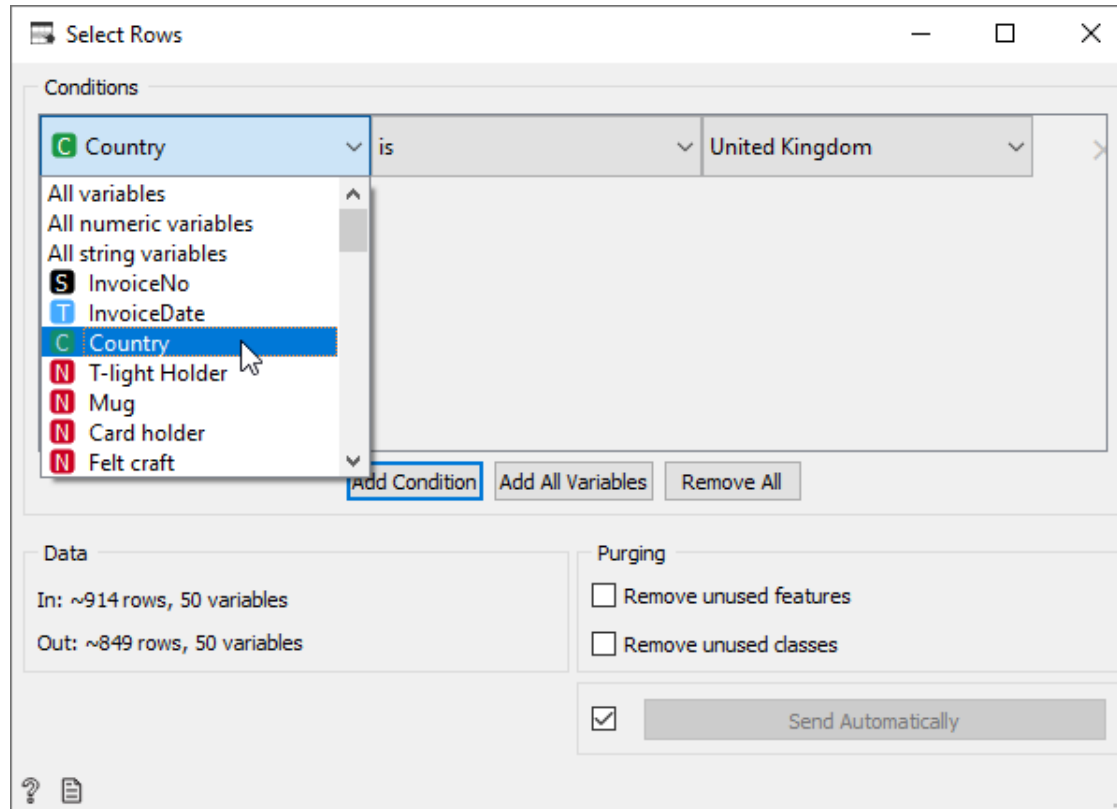
Dữ liệu gốc



Dữ liệu đã bổ sung

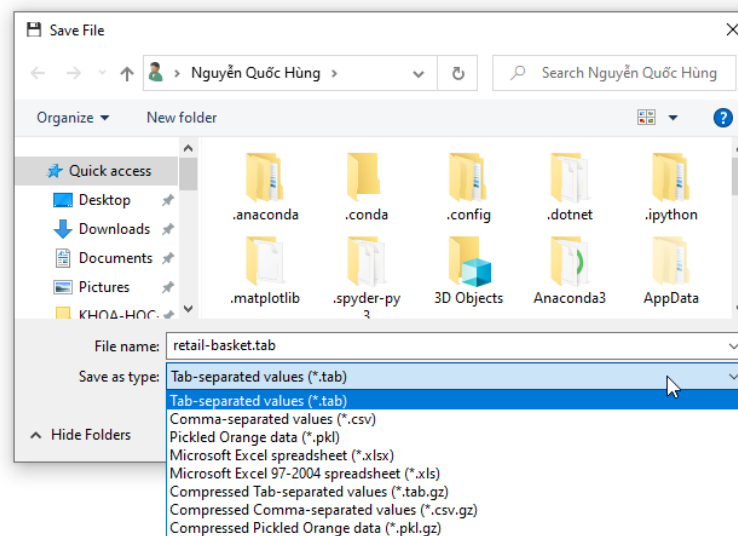
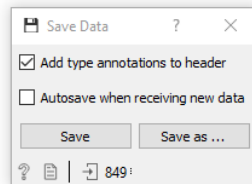
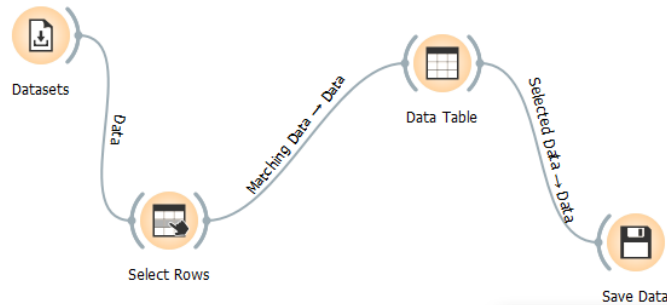
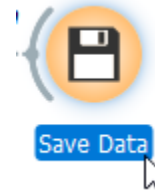
Lựa chọn dữ liệu theo điều kiện

► Chọn chức năng Select Rows



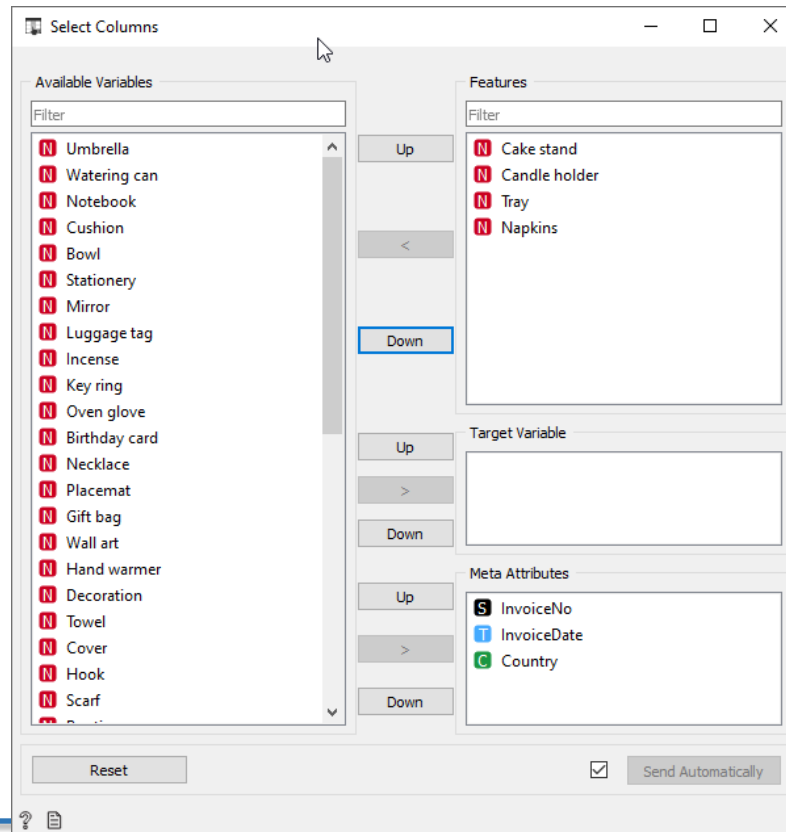
Lưu dữ liệu đã chọn

► Chọn chức năng Save Data:



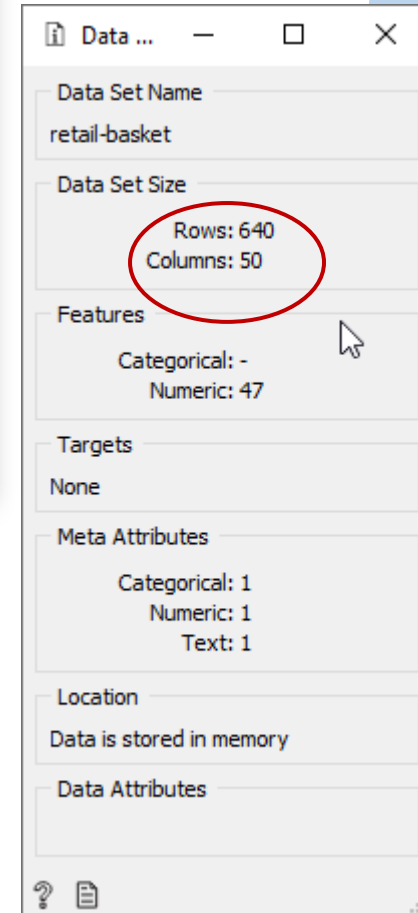
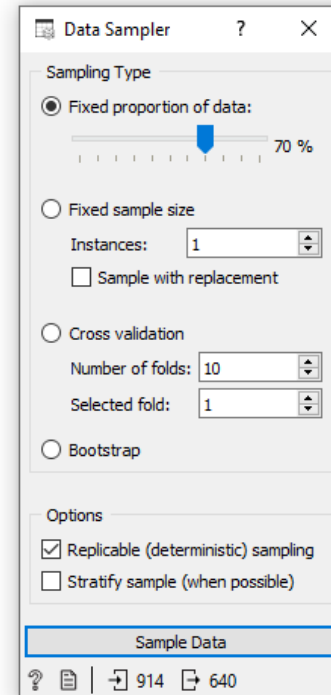
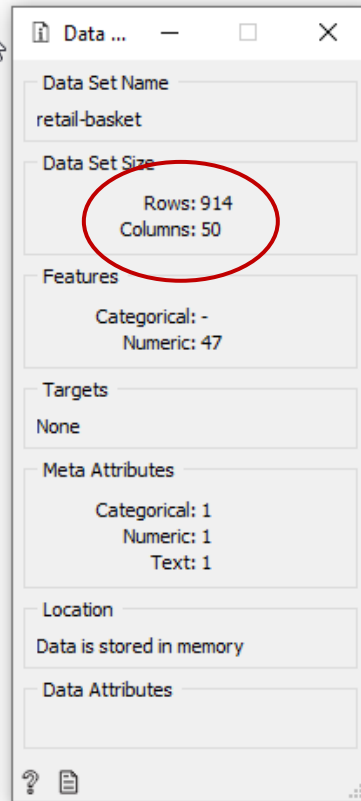
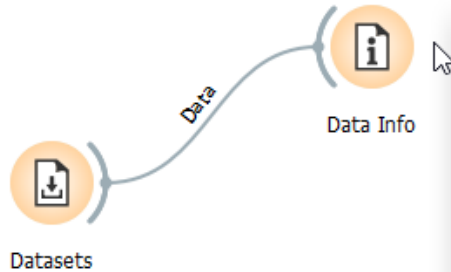
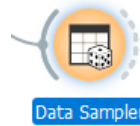
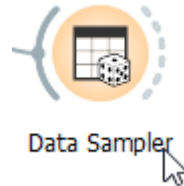
Lựa chọn một số cột dữ liệu

► Chọn chức năng Select columns



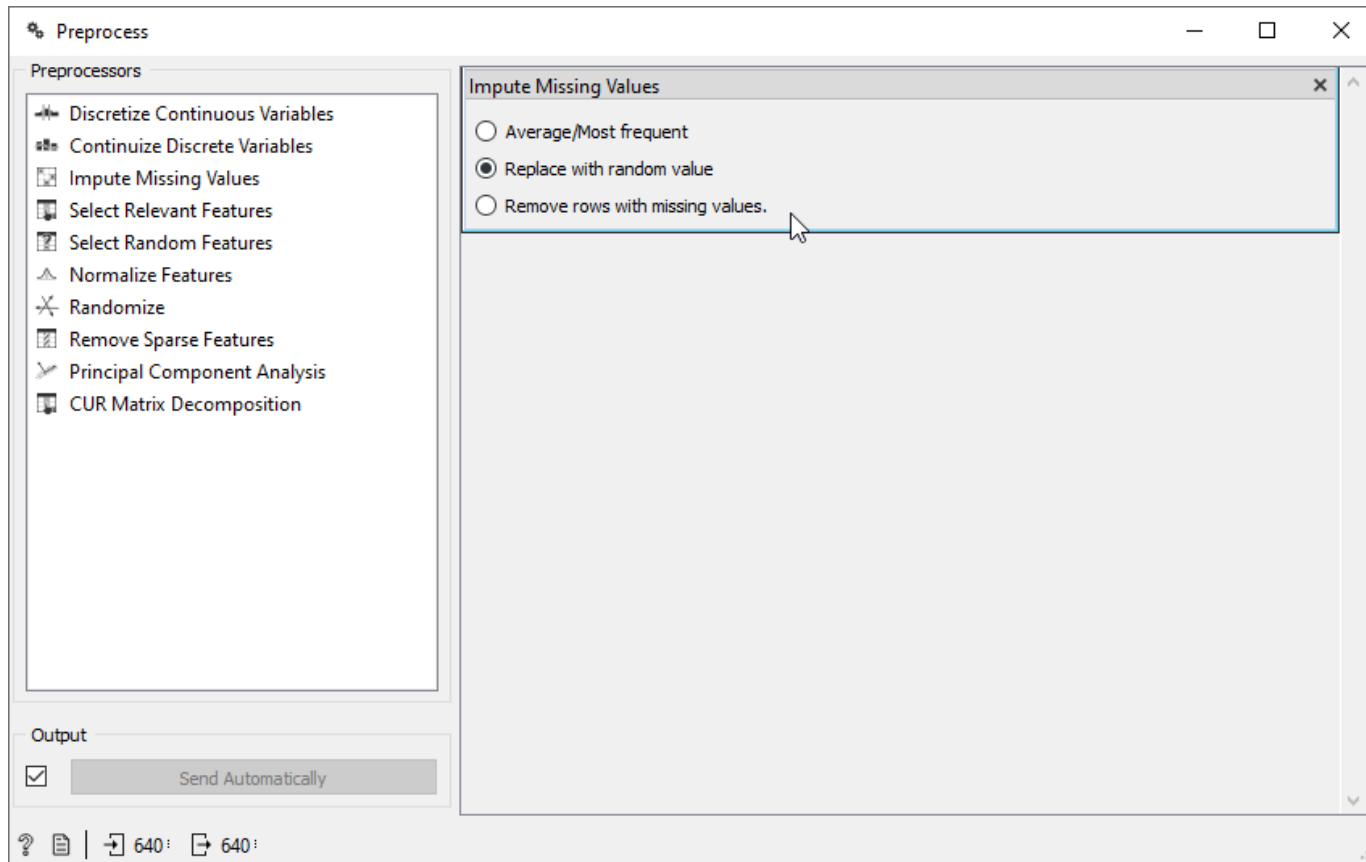
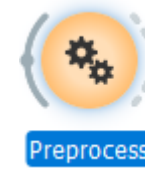
Lấy ngẫu nhiên dữ liệu

► Chọn chức năng:

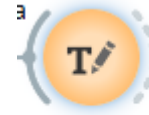


Tiền xử lý dữ liệu

► Chọn vào chức năng: Preprocess



Đổi tên trường dữ liệu



► Chọn chức năng Edit Domain

Edit Domain

Edit Domain

Variables

- T-light Holder
- Mug
- Card holder
- Felt craft
- Frame
- Purse
- Cake stand
- Candle holder
- Tray
- Napkins
- Rain poncho
- Alarm clock
- Magnets
- Umbrella
- Watering can
- Notebook
- Cushion
- Bowl
- Stationery
- Mirror
- Luggage tag
- Incense
- Key ring
- Oven glove
- Birthday card

Edit

Name: Quốc gia

Type: ☒ Categorical ☐ Ordered

Values: Australia → Australia
Belgium → Belgium
Channel Islands → Channel Islands
Denmark → Denmark
EIRE → EIRE
France → France
Germany → Germany
Iceland → Iceland
Italy → Italy
Japan → Japan
Lithuania → Lithuania

Labels: Key Value

Reset Selected Reset All Apply

	InvoiceNo	InvoiceDate	Quốc gia	T-light Holder
1	537044	2010-05-12 10:5...	United Kingdom	1.0
2	537898	2010-09-12 10:4...	United Kingdom	1.0
3	536991	2010-03-12 15:1...	United Kingdom	1.0
4	536993	2010-03-12 15:1...	United Kingdom	2.0
5	C536855	2010-03-12 10:1...	United Kingdom	2.0
6	537243	2010-06-12 10:1...	United Kingdom	1.0
7	537685	2010-08-12 10:2...	United Kingdom	1.0
8	537463	2010-07-12 10:0...	France	1.0
9	C537333	2010-06-12 12:0...	Germany	1.0
10	536398	2010-01-12 10:5...	United Kingdom	2.0
11	537458	2010-07-12 09:5...	United Kingdom	2.0
12	537369	2010-06-12 12:4...	United Kingdom	1.0
13	536755	2010-02-12 14:1...	United Kingdom	1.0
14	536520	2010-01-12 12:4...	United Kingdom	1.0
15	536883	2010-03-12 11:4...	United Kingdom	1.0
16	536407	2010-01-12 11:3...	United Kingdom	1.0
17	536789	2010-02-12 15:2...	United Kingdom	5.0
18	537981	2010-09-12 11:3...	United Kingdom	2.0
19	537158	2010-05-12 13:1...	United Kingdom	2.0