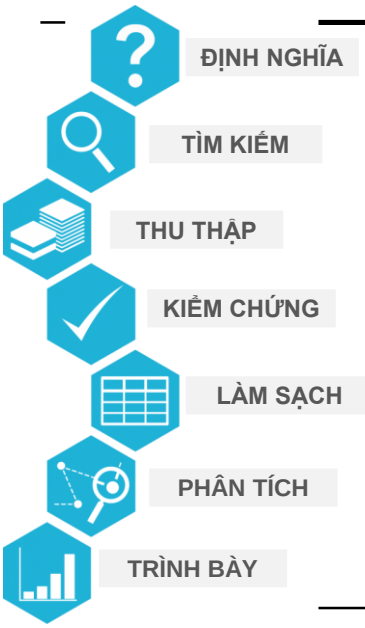
Data pipeline
Quy trình dữ liệu



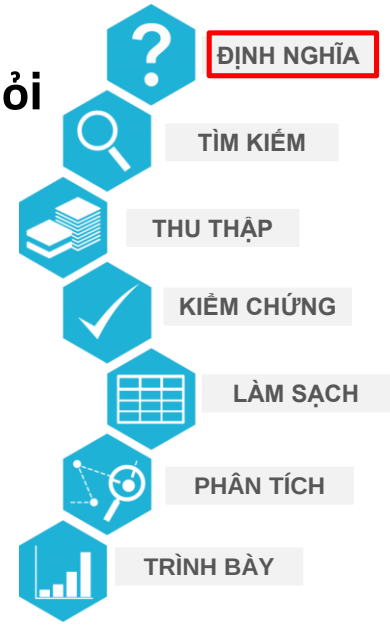
Data Pipeline – Đường ống dữ liệu

- Do School of Data phát triển
- Quy trình theo từng bước có thể áp dụng cho hầu hết các dự án về dữ liệu
- Nguồn dữ liệu từ AidData

2

Định nghĩa/Xác định các câu hỏi

- Tham khảo thông tin online
  - Các cổng dữ liệu
  - Trang web chính phủ
  - Tìm kiếm nâng cao trên Google
  - Các báo cáo

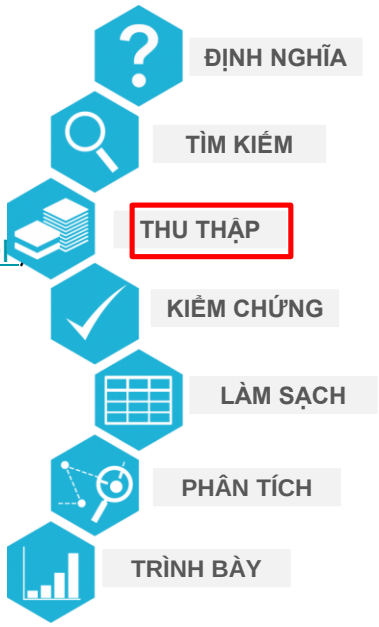


Tìm kiếm dữ liệu

- Online
  - Các cổng dữ liệu
    Trang web chính phủ
    Tìm kiếm nâng cao trên Google
    Các báo cáo
- Offline
  - Hỏi ai đó
  - Các báo cáo
  - Văn bản pháp luật/hợp đồng
  - Bản đồ

Thu thập dữ liệu:

- Tải CSV/Excel files
- Trích xuất dữ liệu từ PDF (Tabula, Small PDF)
- Scrape dữ liệu từ các trang HTML
- Sử dụng API
- Scan và OCR (i2OCR)
- Nhập liệu thủ công
- Nguồn đám đông
  - Điện thoại (Open Data Kit, Kobo toolbox)
  - Nền tảng trực tuyến(PyBossa)



Kiểm chứng dữ liệu

- Kiểm tra tính nhất quán
- Kiểm tra ngoại lai
- Tính toán thống kê cơ bản của dữ liệu (tổng, trung vị, độ lệch chuẩn)
  Kiểm tra dữ liệu meta data sẵn có
- Hỏi các chuyên gia
- Không xóa bất cứ dữ liệu gì trong giai đoạn này

Làm sạch dữ liệu:

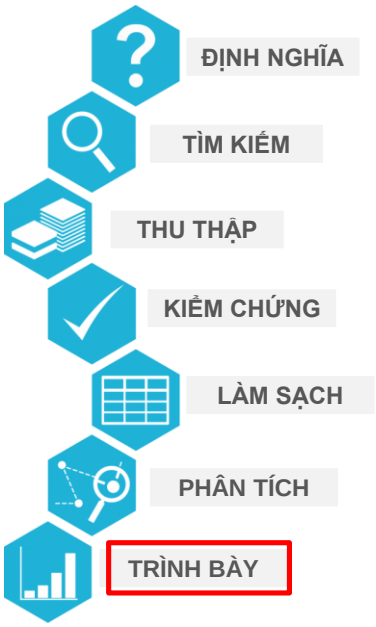
- Mục tiêu: tạo ra một bộ dữ liệu nhất quan, con người có thể hiểu được và máy có thể đọc được
- Chuẩn hóa các tên, thuật ngữ, ngày tháng, các trường
- Sửa những lỗi có thể kiểm chứng/xác minh được
- Kết nối dữ liệu với các bộ dữ liệu hữu ích khác
- Loại bỏ hoặc điền vào các giá trị rỗng

Phân tích dữ liệu:

- Phân tích thăm dò/khám phá:
  - Tìm kiếm các xu hướng thú vị trong bộ dữ liệu
- Kiểm chứng một giả thuyết:
  - Bắt đầu với một lý thuyết
  - Tìm hiểu xem dữ liệu có hỗ trợ cho lý thuyết đó không
- Các công cụ: Excel, Tableau, SPSS, QGIS, R, Stata, Python

Trình bày dữ liệu:

- Làm minh họa và kể câu chuyện từ dữ liệu
- Hiểu đối tượng tiếp nhận
- Giải thích cách dữ liệu trả lời các câu hỏi
  Công cụ: Datawrapper, Flourish, Pikochart, Canva,…

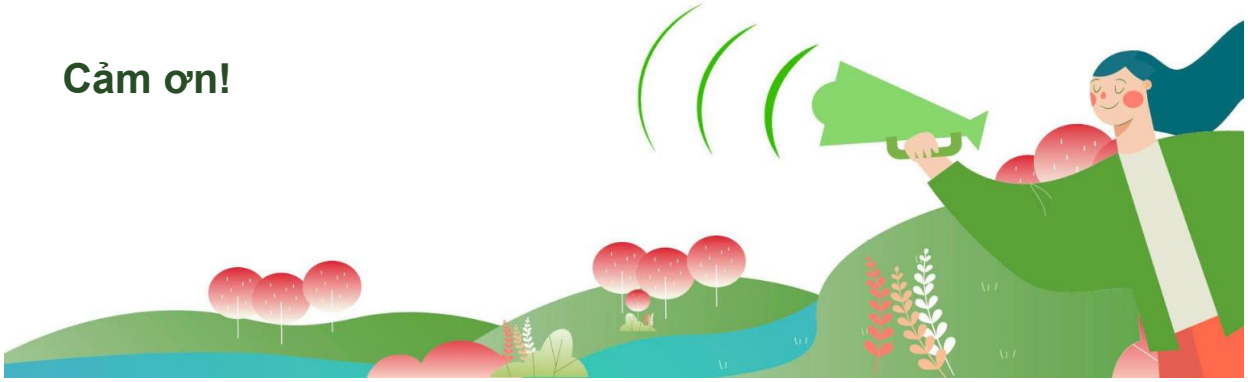


Không ai có thể "HIỂU BIẾT ĐẦY ĐỦ VỀ DỮ LIỆU"

Chúng ta sẽ luôn trong trạng thái:
có những kỹ năng đã thành thạo
có những kỹ năng có thể dùng tàm tạm
có những kỹ năng cần phải cải thiện thêm

10

Cảm ơn!



This project is funded by the European Union

V 4 M F | Voices for Mekong Forests