

Thông báo bản quyền

Các slide này được phân phối theo Giấy phép Creative Commons.

DeepLearning.AI cung cấp các slide này cho mục đích giáo dục. Bạn không được sử dụng hoặc phân phối các slide này cho mục đích thương mại. Bạn có thể tạo bản sao của các trang trình bày này và sử dụng hoặc phân phối chúng cho mục đích giáo dục miễn là bạn trích dẫn DeepLearning.AI làm nguồn của các slide.

Để biết phần còn lại của các chi tiết của giấy phép,

hãy xem <https://creativecommons.org/licenses/by-sa/2.0/legalcode>

Thu thập, dán nhãn và Xác thực dữ liệu



DeepLearning.AI

Chào mừng

Tầm quan trọng của dữ liệu

“Dữ liệu là phần khó nhất của ML và là phần quan trọng nhất để hiểu đúng...

Dữ liệu bị hỏng là nguyên nhân phổ biến nhất gây ra sự cố trong các hệ thống ML sản xuất”

- Mở rộng quy mô Machine Learning tại Uber với Michelangelo - Uber

“Không có hoạt động nào khác trong vòng đời máy học có lợi tức đầu tư cao hơn việc cải thiện dữ liệu mà một mô hình có quyền truy cập.”

- Lễ hội: Kết nối dữ liệu và mô hình ML - Gojek



DeepLearning.AI

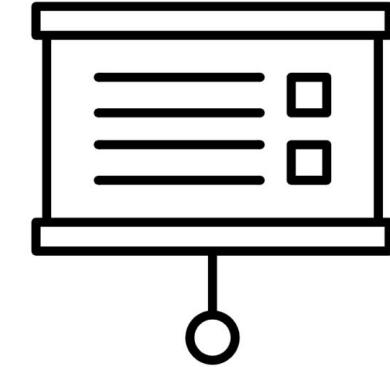
Giới thiệu về Máy

Học kỹ thuật cho sản xuất

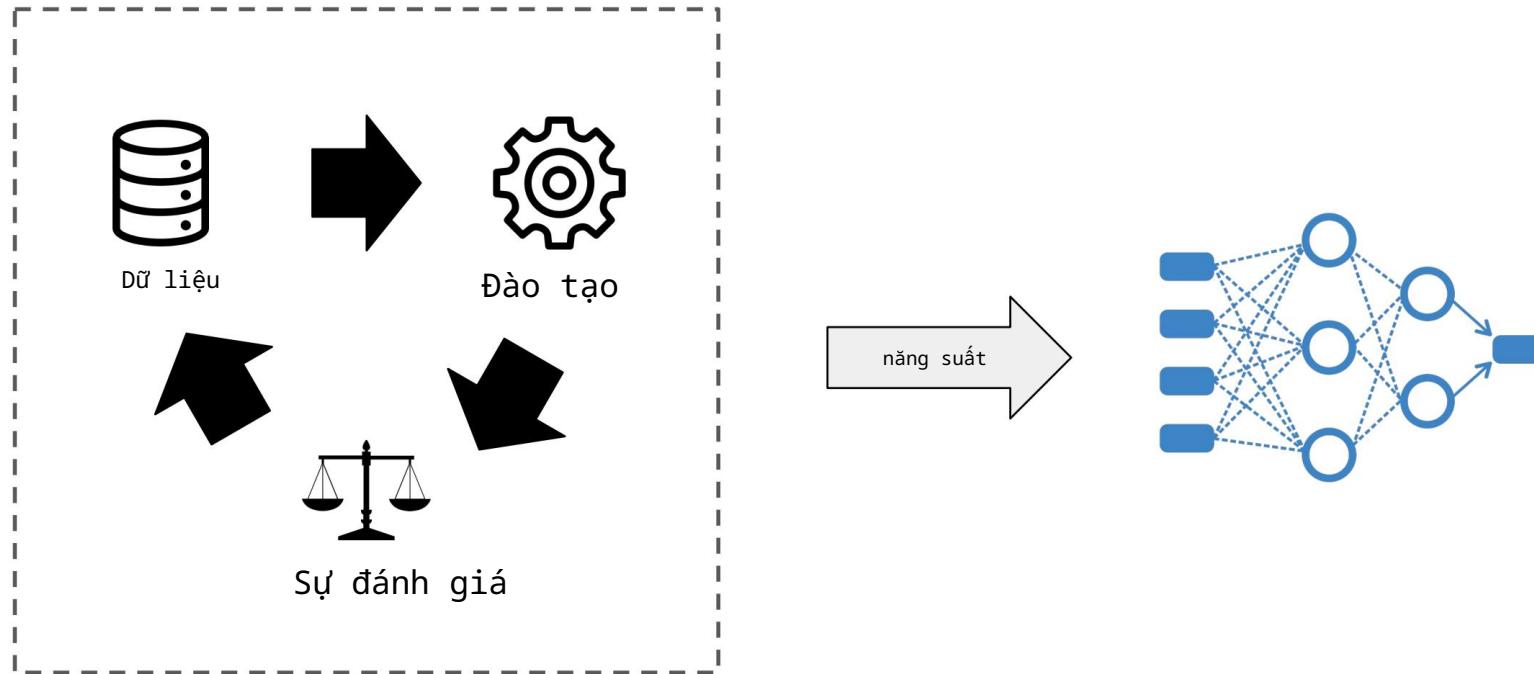
Tổng quan

Đề cương

- Kỹ thuật máy học (ML) cho sản xuất: tổng quan
- ML sản xuất = Phát triển ML + phát triển phần mềm
- Thách thức trong ML sản xuất



Mô hình ML truyền thống



Các hệ thống ML sản xuất đòi hỏi nhiều hơn thế



Lập mô hình ML so với ML sản xuất

ML Học thuật/Nghiên cứu		ML sản xuất
Dữ liệu	tĩnh	Năng động - Dịch chuyển
Ưu tiên thiết kế	Độ chính xác tổng thể cao nhất	Suy luận nhanh, khả năng diễn giải tốt sự suy luận khả năng diễn giải
đào tạo người mẫu	Điều chỉnh và đào tạo tối ưu	Liên tục đánh giá và đào tạo lại
Công bằng	Rất quan trọng	Chủ yếu chủ yếu
Thử thách	Thuật toán chính xác cao	Toàn bộ hệ thống

Học máy sản xuất

Phát triển máy học

phần mềm hiện đại
+ phát triển

Quản lý toàn bộ vòng đời của dữ liệu

- Dán nhãn
- Phạm vi không gian tính năng
- Kích thước tối thiểu
- Dữ liệu dự đoán tối đa
- Công bằng
- Điều kiện hiếm gặp

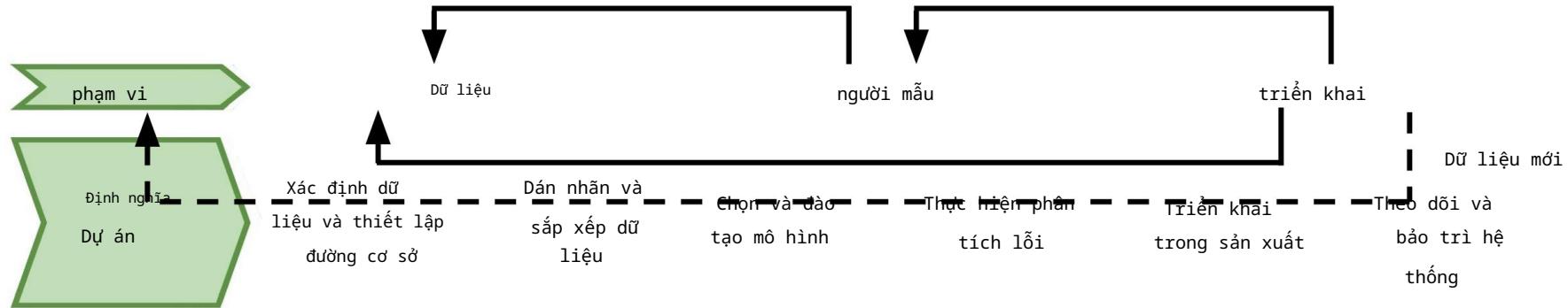
Phát triển phần mềm hiện đại

Tỉ lệ:

- Khả năng mở rộng
- Khả năng mở rộng
- Cấu hình
- Tính nhất quán & khả năng tái sản xuất
- An toàn & bảo mật
- Tính mô đun
- Khả năng kiểm tra
- Giám sát
- Thực tiễn tốt nhất



Hệ thống học máy sản xuất



Những thách thức trong ML cấp sản xuất

- Xây dựng hệ thống ML tích hợp
- Vận hành liên tục trong sản xuất
- Xử lý dữ liệu thay đổi liên tục
- Tối ưu hóa chi phí tài nguyên tính toán





DeepLearning.AI

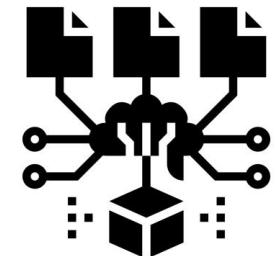
Giới thiệu về máy

Học kỹ thuật cho sản xuất

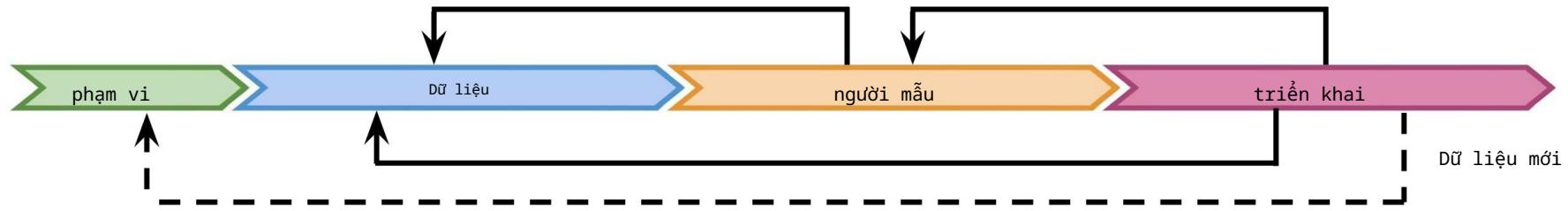
Đường ống ML

Đề cương

- Đường ống ML
- Đồ thị tuần hoàn có hướng và Khung phối hợp đường ống
- Giới thiệu về TensorFlow Extended (TFX)

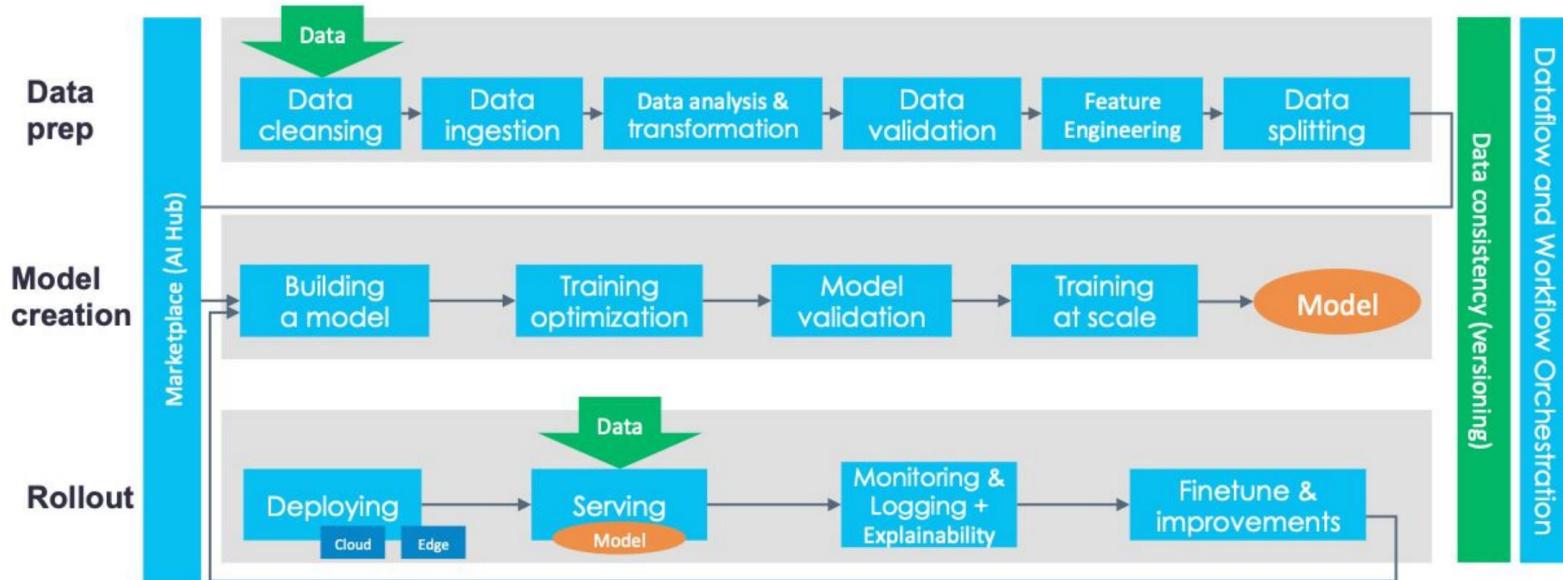


đường ống ML



Cơ sở hạ tầng ML sản xuất

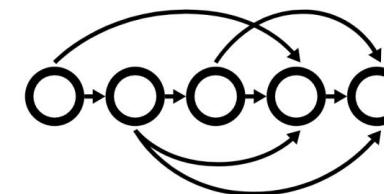
Kiến trúc tham chiếu MLops của CD Foundation



Đồ thị tuần hoàn có hướng



- Đồ thị có hướng theo chu trình (DAG) là đồ thị có hướng không có chu trình
- Quy trình công việc đường ống ML thường là DAG
- DAG xác định trình tự của các tác vụ sẽ được thực hiện, dựa trên các mối quan hệ và sự phụ thuộc.



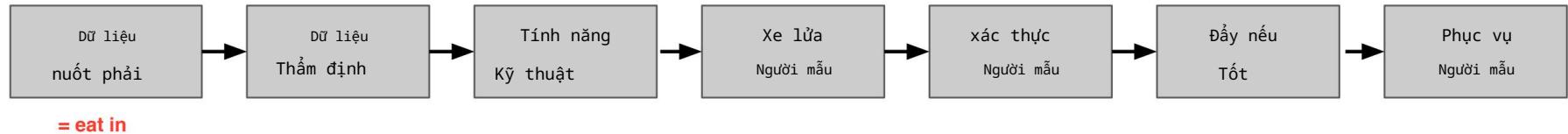
Khung điều phối đường ống



- Chịu trách nhiệm lên lịch cho các thành phần khác nhau trong một phụ thuộc DAG đường dẫn ML
- Trợ giúp về tự động hóa đường ống
- Ví dụ: Airflow, Argo, Celery, Luigi, Kubeflow

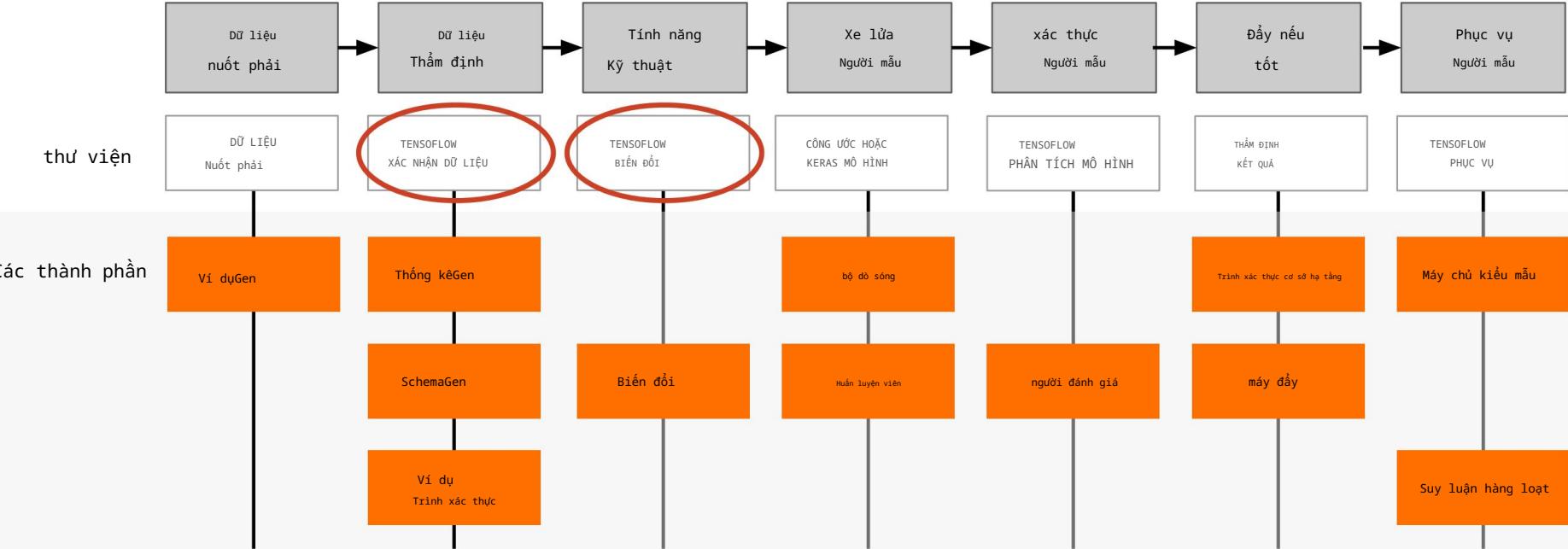
TensorFlow Extended (TFX)

Nền tảng đầu cuối để triển khai các quy trình ML sản xuất

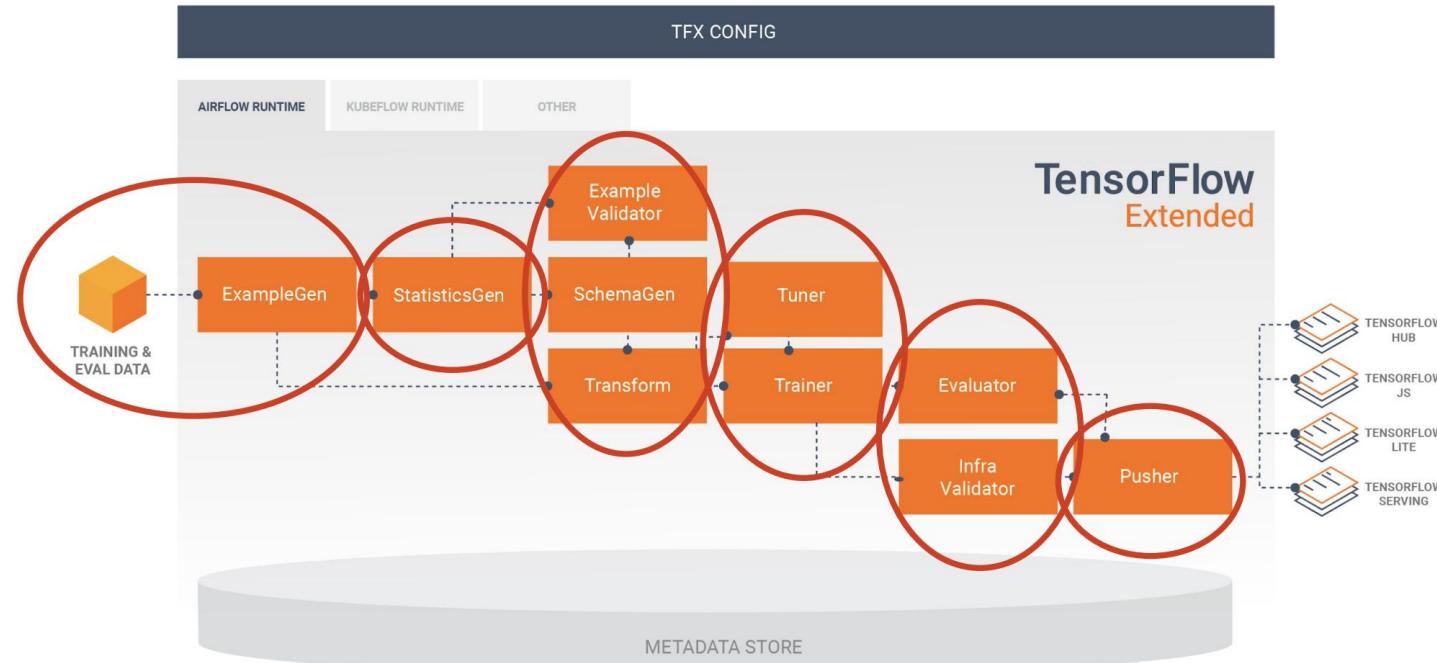


Chuỗi các thành phần được thiết kế cho các tác vụ học máy hiệu suất cao, có thể mở rộng

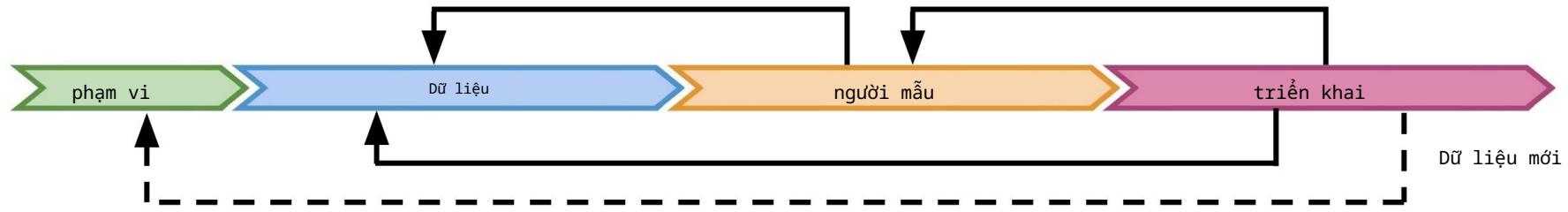
Linh kiện sản xuất TFX



TFX xin chào thế giới



Những điểm chính



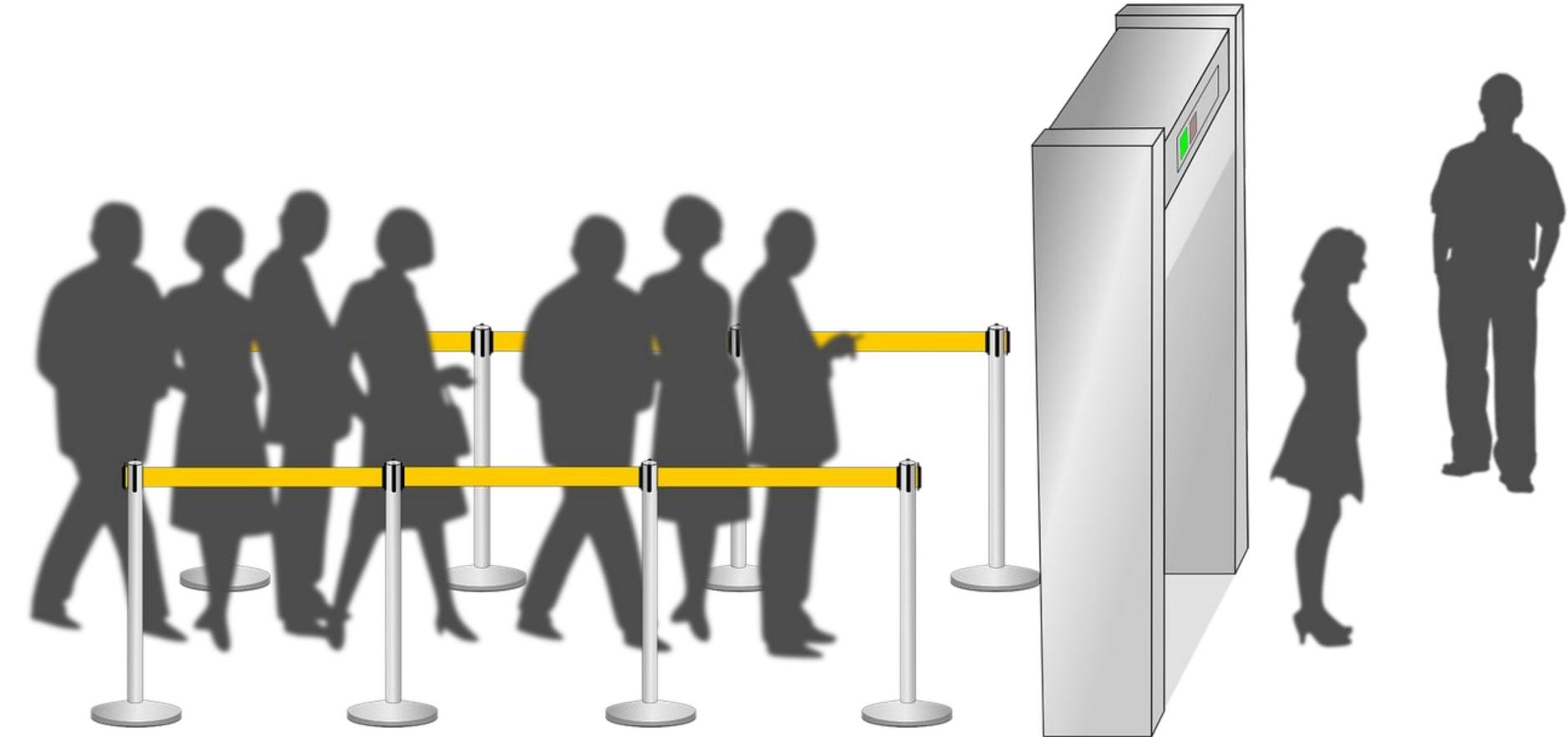
- Quy trình ML sản xuất: tự động hóa, giám sát và duy trì các quy trình từ đầu đến cuối
- ML sản xuất không chỉ là mã ML
 - Phát triển ML + phát triển phần mềm
- TFX là một nền tảng ML đầu cuối mã nguồn mở



DeepLearning.AI

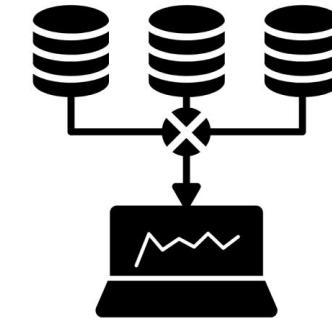
Thu thập dữ liệu

Tầm quan trọng của dữ liệu



Đề cương

- Tầm quan trọng của chất lượng dữ liệu
- Đường ống dữ liệu: thu thập, nhập và chuẩn bị dữ liệu
- Thu thập và giám sát dữ liệu



Tầm quan trọng của dữ liệu

"Dữ liệu là phần khó nhất của ML và là phần quan trọng nhất để hiểu đúng... Dữ liệu bị hỏng là phần lớn nhất nguyên nhân phổ biến của các vấn đề trong hệ thống ML sản xuất"

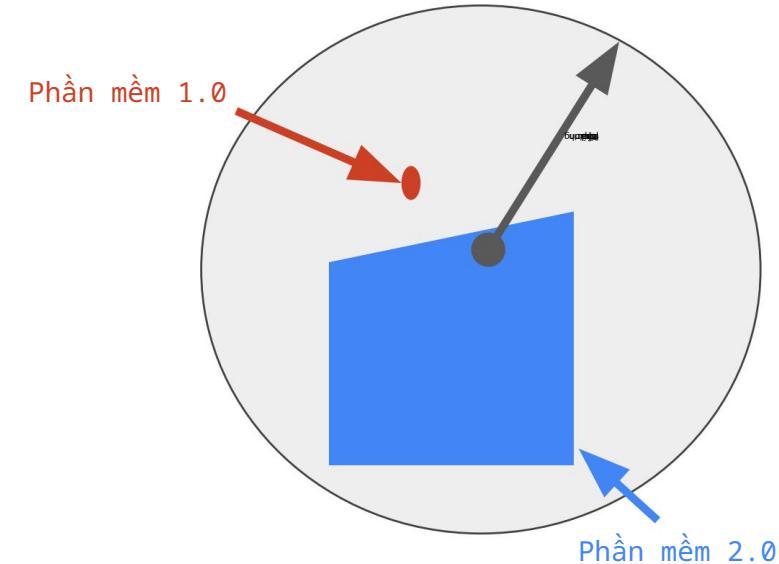
- Mở rộng quy mô Machine Learning tại Uber với Michelangelo - Uber

"Không có hoạt động nào khác trong vòng đời máy học có lợi tức đầu tư cao hơn việc cải thiện dữ liệu mà mô hình có quyền truy cập."

- Lễ hội: Kết nối dữ liệu và mô hình ML - Gojek

ML: Dữ liệu là công dân hạng nhất

- Phần mềm 1.0
 - Hướng dẫn rõ ràng cho máy tính
- Phần mềm 2.0
 - Chỉ định một số mục tiêu về hành vi của chương trình
 - Tìm giải pháp bằng kỹ thuật tối ưu hóa
 - Dữ liệu tốt là chìa khóa thành công
 - Mã trong Phần mềm = Dữ liệu trong ML



Mỗi thứ bắt đầu với dữ liệu

- Người mẫu không phải là ma thuật
 - Dữ liệu có ý nghĩa:
 - tối đa hóa nội dung dự đoán
 - xóa dữ liệu không mang tính thông tin
 - tính năng bảo hiểm không gian



Rác vào, rác ra

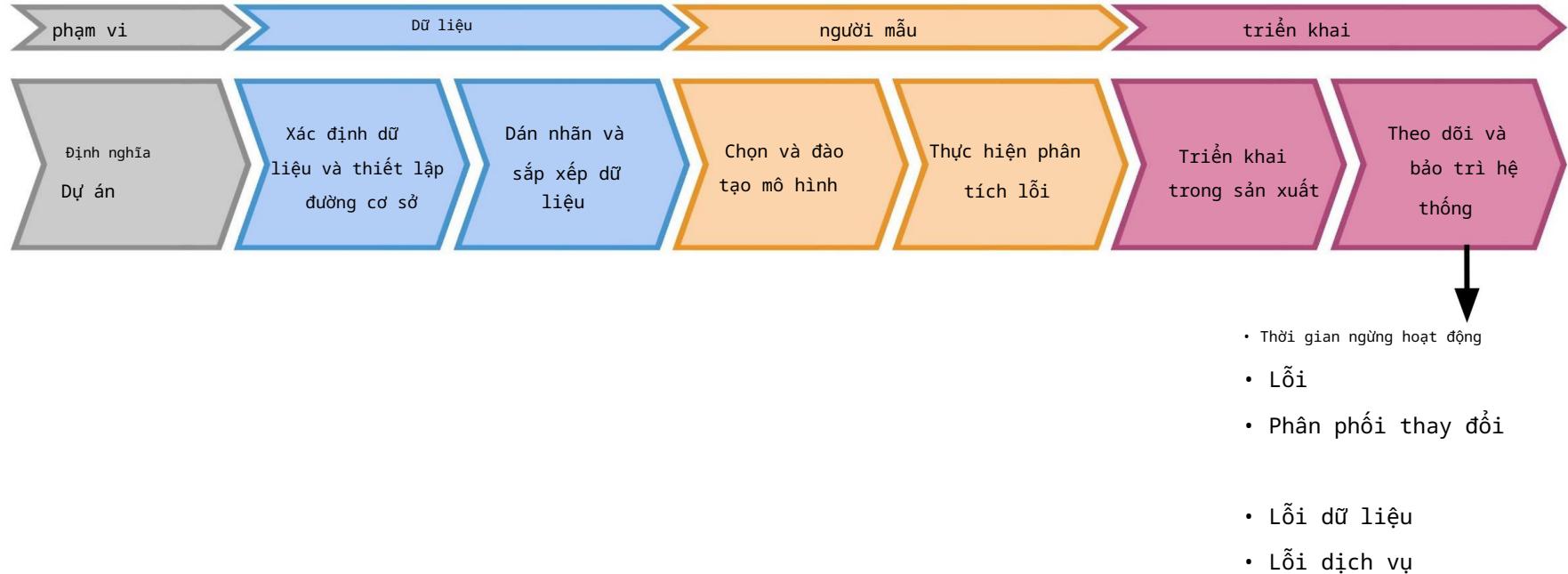
$$f(\text{trash}) = \text{trash}$$

đường dẫn dữ liệu



- Thu thập dữ liệu
- Nhập dữ liệu
- Định dạng dữ liệu
- Kỹ thuật tính năng
- Khai thác tính năng

Thu thập và giám sát dữ liệu



Những điểm chính

- Thấu hiểu người dùng, biến nhu cầu của người dùng thành bài toán dữ liệu
- Đảm bảo vùng phủ sóng dữ liệu và tín hiệu dự đoán cao
- Nguồn, lưu trữ và giám sát dữ liệu chất lượng một cách có trách nhiệm



DeepLearning.AI

Thu thập dữ liệu

Ứng dụng ví dụ:
đề xuất chạy



Ứng dụng ví dụ: Đề xuất chạy

người dùng	người chạy
Người dùng cần	Chạy thường xuyên hơn
Hành động của người dùng	Hoàn thành chạy bằng ứng dụng
Đầu ra hệ thống ML	<ul style="list-style-type: none"> • Những tuyến đường để đề xuất • Khi nào nên đề xuất chúng
Học hệ thống ML	<ul style="list-style-type: none"> • Các mẫu hành vi xung quanh việc chấp nhận lời nhắc chạy • Hoàn thành chạy • Cải thiện tính nhất quán

cân nhắc chính

- Tính khả dụng và thu thập dữ liệu
 - Có bao nhiêu/loại dữ liệu nào?
 - Dữ liệu mới đến thường xuyên như thế nào?
 - Có chú thích không?
 - Nếu không, việc dán nhãn khó/tốn kém như thế nào?
- Chuyển nhu cầu của người dùng thành nhu cầu dữ liệu
 - Dữ liệu cần thiết
 - Các tính năng cần thiết
 - Nhãn cần thiết



tập dữ liệu mẫu

ĐẶC TRƯNG

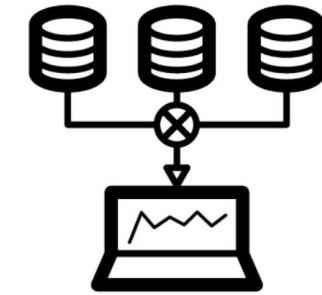
Chạy ID người chạy		Độ cao thời gian của người chạy	Vui vẻ	
AV3DE	cuộc chạy đua marathon Boston	03:40:32	1.300 ft	Thấp
X8KGF	Lễ hội tháng mười Seattle 5k	00:35:40	0 ft	Trung bình
BH9IU	Houston Nửa marathon	02:01:18	200 ft	Trung bình

N
A
H
N
N



Nhận biết dữ liệu của bạn

- Xác định nguồn dữ liệu
- Kiểm tra xem chúng có được làm mới không
- Nhắc quan về giá trị, đơn vị & kiểu dữ liệu
- Theo dõi các giá trị ngoại lệ và lỗi



Các vấn đề về bộ dữ liệu

- Định dạng không nhất quán
 - Là số 0 “0”, “0,0” hoặc chỉ báo thiếu phép đo
- Tổng hợp lỗi từ các Mô hình ML khác
- Giám sát các nguồn dữ liệu cho các sự cố hệ thống và sự cố ngừng hoạt động

Đo lường hiệu quả dữ liệu

- Trực giác về giá trị dữ liệu có thể gây hiểu nhầm
 - Tính năng nào có giá trị tiên đoán và tính năng nào có giá trị không?
- Kỹ thuật tính năng giúp tối đa hóa các tín hiệu dự đoán
- Lựa chọn tính năng giúp đo các tín hiệu dự đoán

Dịch nhu cầu của người dùng thành nhu cầu dữ liệu

dữ liệu cần thiết

- Chạy dữ liệu từ ứng dụng
- Dữ liệu nhân khẩu học
- Dữ liệu địa lý cục bộ

Dịch nhu cầu của người dùng thành nhu cầu dữ liệu

Tính năng cần thiết

- Nhân khẩu học của người chạy
- Thời gian trong ngày
- Tỷ lệ hoàn thành chạy
- Tốc độ
- Quãng đường đã chạy
- Độ cao đạt được
- Nhịp tim

Dịch nhu cầu của người dùng thành nhu cầu dữ liệu

nhận cần thiết

- Người chạy chấp nhận hoặc từ chối đề xuất ứng dụng
- Phản hồi do người dùng tạo về lý do đề xuất vật bị loại bỏ
- Đánh giá của người dùng về sự thú vị của các lần chạy được đề xuất

Những điểm chính

- Hiểu người dùng của bạn, biến nhu cầu của họ thành vấn đề về dữ liệu
 - Loại/bao nhiêu dữ liệu có sẵn
 - Các chi tiết và vấn đề về dữ liệu của bạn là gì
 - Các tính năng dự đoán của bạn là gì
 - Nhận bạn đang theo dõi là gì
 - Số liệu của bạn là gì





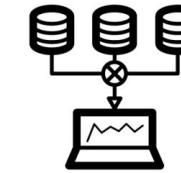
DeepLearning.AI

Thu thập dữ liệu

Dữ liệu chịu trách nhiệm:
Bảo mật, Quyền riêng tư &
Công bằng

Đề cương

- Tìm nguồn dữ liệu
- Bảo mật dữ liệu và quyền riêng tư của người dùng
- Thiên vị và Công bằng

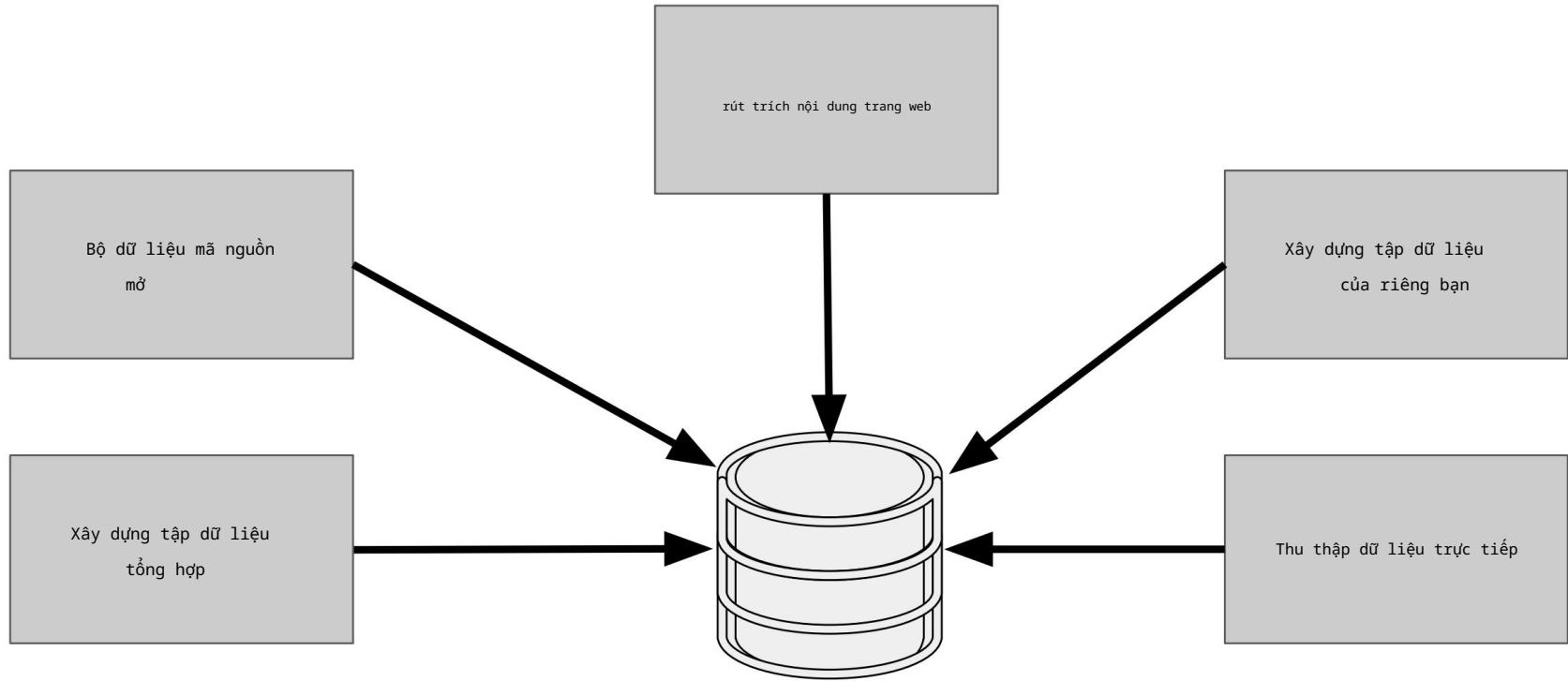


Tránh những sai lệch có vấn đề trong bộ dữ liệu

Ví dụ: trình phân loại được đào tạo trên bộ dữ liệu Open Images



Nguồn dữ liệu có trách nhiệm



Bảo mật dữ liệu và quyền riêng tư

- Thu thập và quản lý dữ liệu không chỉ là về người mẫu
 - Cung cấp cho người dùng quyền kiểm soát dữ liệu nào có thể được thu thập
 - Có nguy cơ vô tình tiết lộ dữ liệu người dùng không?
- Tuân thủ các quy định và chính sách (ví dụ: GDPR)

Quyền riêng tư của người dùng

- Bảo vệ thông tin nhận dạng cá nhân
 - Tổng hợp - thay thế các giá trị duy nhất bằng tóm tắt giá trị
 - Biên tập lại - loại bỏ một số dữ liệu để tạo ít dữ liệu hơn hoàn thành bức tranh

Hệ thống ML có thể khiến người dùng thất bại như thế nào



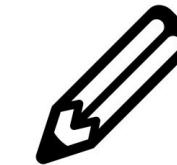
hội chợ



chiu trách nhiệm

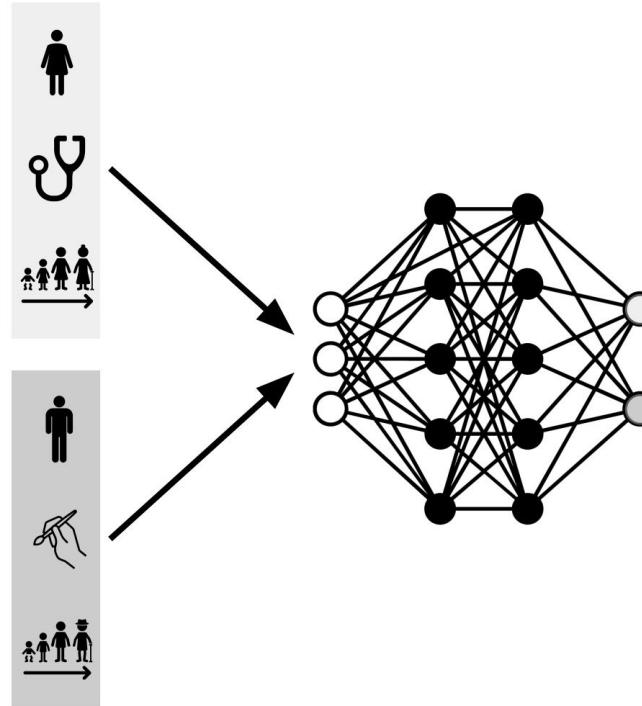


mình bạch có thể giải thích



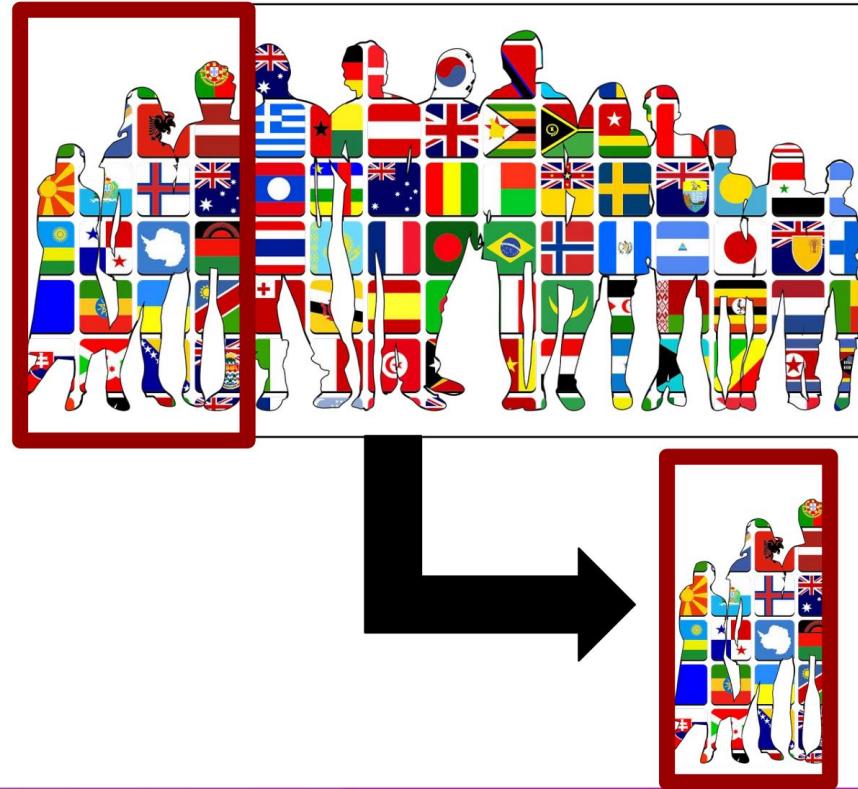
- Tác hại của đại diện
- Từ chối cơ hội
- Lỗi sản phẩm không tương ứng
- Bị hại bởi bất lợi

Cam kết công bằng



- Đảm bảo các mô hình của bạn công bằng
 - Nhóm công bằng, chính xác như nhau
- Xu hướng trong con người được dán nhãn và/hoặc thu thập dữ liệu
- Mô hình ML có thể khuếch đại các thành kiến

Biểu diễn dữ liệu thiên vị



Giảm sự thiên vị: Thiết kế hệ thống ghi nhãn công bằng

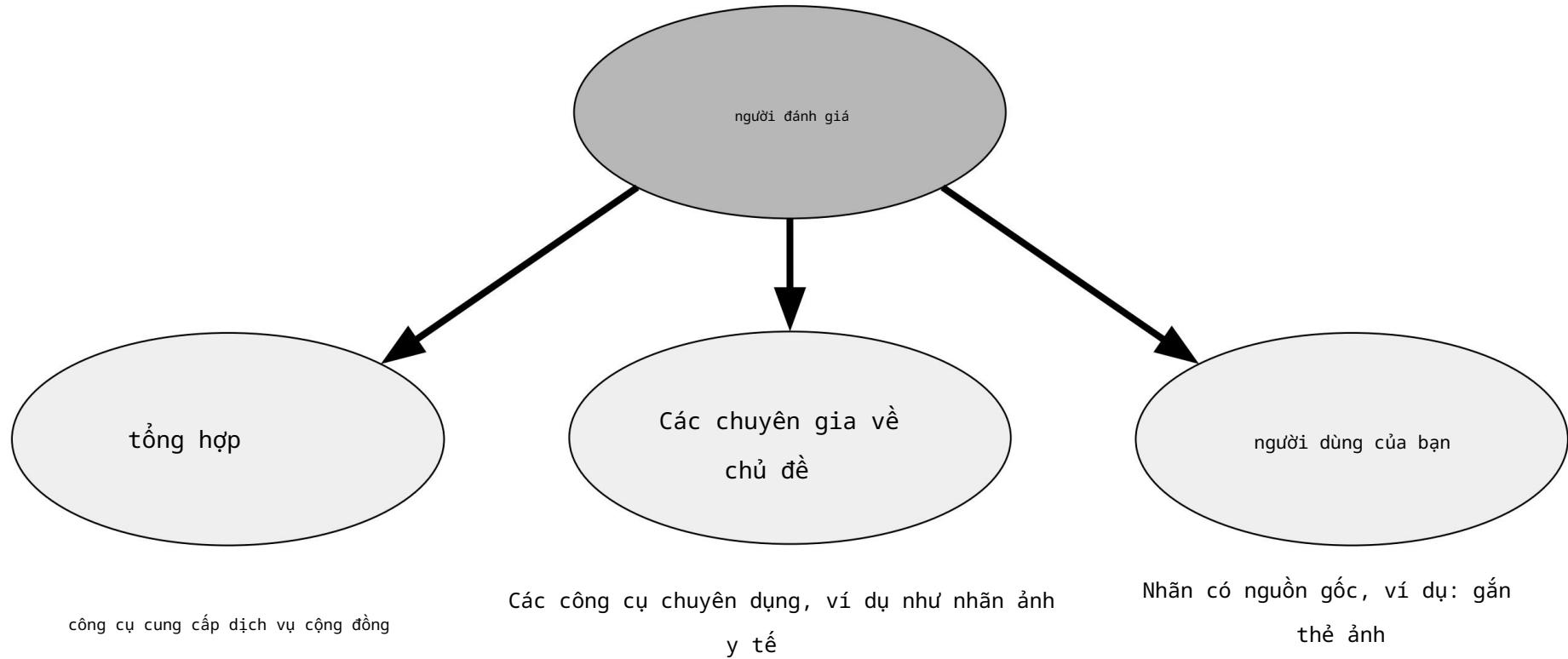
- Nhãn chính xác là cần thiết cho việc học có giám sát

- Việc dán nhãn có thể được thực hiện bằng cách:

- Tự động hóa (ghi nhật ký hoặc giám sát yếu)
- Con người (hay còn gọi là "Người đánh giá", thường được giám sát bán phần)



Các loại người đánh giá con người



công cụ cung cấp dịch vụ cộng đồng

Các công cụ chuyên dụng, ví dụ như nhãn ảnh
y tế

Nhãn có nguồn gốc, ví dụ: gắn
thẻ ảnh

Những điểm chính

- Đảm bảo tính đa dạng của nhóm người xếp hạng
- Điều tra bối cảnh và các ưu đãi của người đánh giá
- Đánh giá các công cụ đánh giá
- Quản lý chi phí
- Xác định yêu cầu về độ tươi



DeepLearning.AI

Ghi nhãn dữ liệu

Nghiên cứu điển hình: Suy thoái
Hiệu suất người mẫu

Bạn là một nhà bán lẻ trực
tuyến bán giày . . .

Mô hình của bạn dự
đoán tỷ lệ nhấp
(CTR), giúp bạn quyết định
số lượng hàng tồn kho cần
đặt hàng



Khi đột nhiên

AUC và độ chính xác dự đoán của
bạn đã giảm trên giày công sở nam!







Làm thế nào để chúng ta biết rằng
chúng ta có một vấn đề?





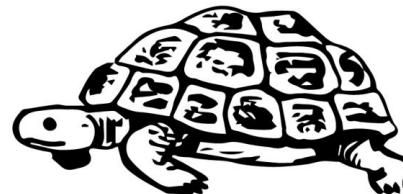
Nghiên cứu điển hình: hành động

- Làm thế nào để phát hiện vấn đề sớm?
- Nguyên nhân có thể là gì? • Có
thể làm gì để giải quyết những vấn đề này?

Điều gì gây ra vấn đề?

Các loại vấn đề:

- Chậm - ví dụ: trôi
- Nhanh - ví dụ: cảm biến kém, cập nhật phần mềm kém



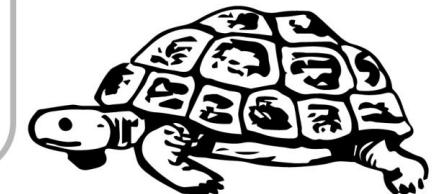
vấn đề dần dần

Thay đổi dữ liệu

- Xu hướng và tính thời vụ • Phân phối các thay đổi về tính năng • Tầm quan trọng tương đối của các thay đổi về tính năng

thế giới thay đổi

- Phong cách thay đổi • Phạm vi và quy trình thay đổi • Đổi thủ cạnh tranh thay đổi
- Kinh doanh mở rộng sang các khu vực địa lý khác



vấn đề đột ngột

Vấn đề thu thập dữ liệu

- Cảm biến/máy ảnh kém • Dữ liệu nhạt ký không hợp lệ • Cảm biến/máy ảnh bị vô hiệu hóa hoặc di chuyển

Sự cố hệ thống

- Cập nhật phần mềm kém • Mất kết nối mạng • Hệ thống ngừng hoạt động • Thông tin xác thực không hợp lệ



Tại sao “Hiểu” mô hình?

- Dự đoán sai không có chi phí thông nhất cho doanh nghiệp của bạn
- Dữ liệu bạn có hiếm khi là dữ liệu bạn mong muốn
- Mục tiêu mô hình gần như luôn là đại diện cho doanh nghiệp của bạn
mục tiêu
- Một số phần trăm khách hàng của bạn có thể có trải nghiệm không tốt

Thế giới thực không đứng yên!



DeepLearning.AI

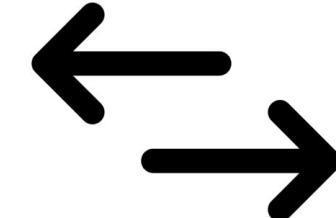
Ghi nhãn dữ liệu

Dữ liệu và Khái niệm

Thay đổi trong
ML sản xuất

Đề cương

- Phát hiện sự cố với các mô hình đã triển khai
 - Thay đổi dữ liệu và khái niệm
- Thay đổi sự thật cơ bản
 - Các vấn đề dễ dàng
 - Bài toán khó hơn
 - Bài toán thực sự khó



Phát hiện sự cố với các mô hình đã triển khai

- Dữ liệu và phạm vi thay đổi
- Giám sát các mô hình và xác thực dữ liệu để phát hiện sớm các sự cố
- Thay đổi sự thật cơ bản: gắn nhãn dữ liệu đào tạo mới

vấn đề dễ dàng

- Chân đất thay đổi chậm (tháng, năm)
- Đào tạo lại mô hình do:
 - Cải tiến mô hình, dữ liệu tốt hơn
 - Thay đổi về phần mềm và/hoặc hệ thống
 - Dán nhãn
 - Bộ dữ liệu được tuyển chọn
 - Dựa trên đám đông



vấn đề khó khăn hơn

- Sự thật cơ bản thay đổi nhanh hơn (tuần)
- Đào tạo lại mô hình do:
 - Giảm hiệu suất của mô hình
 - Cải thiện mô hình, dữ liệu tốt hơn
 - Thay đổi trong phần mềm và/hoặc hệ thống
- Dán nhãn
 - Phản hồi trực tiếp
 - Dựa trên đám đông



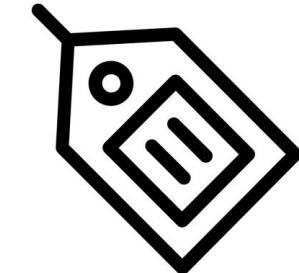
vấn đề thực sự khó khăn

- Sự thật mặt đất thay đổi rất nhanh (ngày, giờ, phút)
- Đào tạo lại mô hình do:
 - Giảm hiệu suất của mô hình
 - Cải thiện mô hình, dữ liệu tốt hơn
 - Thay đổi trong phần mềm và/hoặc hệ thống
- Dán nhãn
 - Phản hồi trực tiếp
 - Giám sát yếu



Những điểm chính

- Hiệu suất của mô hình suy giảm theo thời gian
 - Trôi dạt dữ liệu và khái niệm
- Đào tạo lại mô hình giúp cải thiện hiệu suất
 - Ghi nhãn dữ liệu để thay đổi sự thật cơ bản và nhãn khan hiếm





DeepLearning.AI

Ghi nhãn dữ liệu

Quá trình phản hồi và
dán nhãn con người

ghi nhãn dữ liệu

Nhiều phương pháp

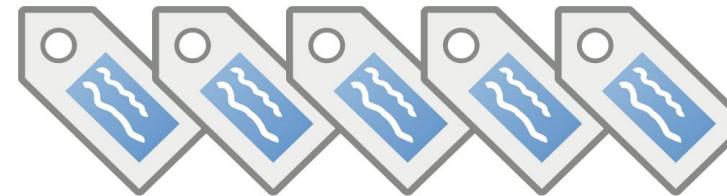
- Phản hồi quy trình (Ghi nhãn trực tiếp)
- Dán nhãn cho con người

~~nhéibáhágiaónsgám sát • Ghi~~
~~Hộ họchủ động~~ →
~~sátigusát yểu • Giám~~



Thực hành sau như phương
pháp dán nhãn nâng cao

ghi nhãn dữ liệu



Quy trình phản hồi

Ví dụ: Số lần nhấp thực tế so với số lần nhấp được dự đoán

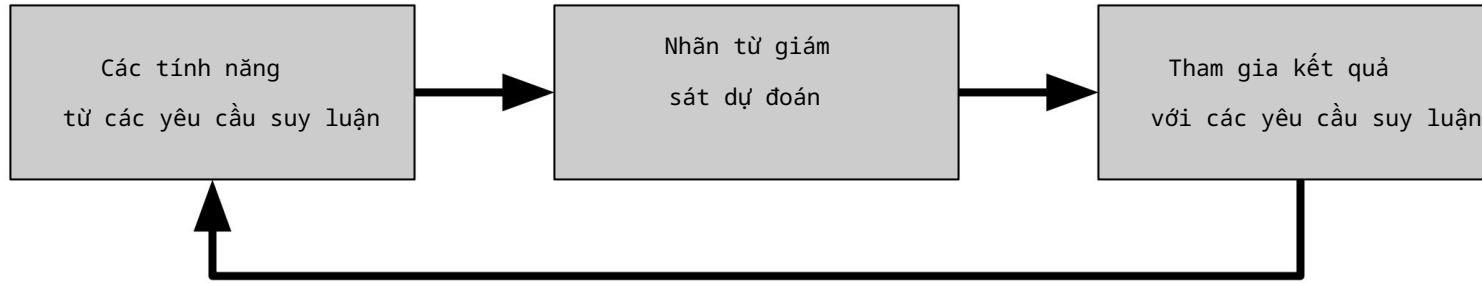
dán nhãn con người

Ví dụ: Bác sĩ tim mạch dán nhãn hình ảnh MRI

Tại sao ghi nhãn lại quan trọng trong ML sản xuất?

- Sử dụng dữ liệu có sẵn của doanh nghiệp/tổ chức
- Đào tạo lại mô hình thường xuyên
- Gắn nhãn cho quy trình quan trọng và đang diễn ra
- Tạo tập dữ liệu đào tạo cần có nhãn

Ghi nhãn trực tiếp: liên tục tạo tập dữ liệu huấn luyện



Tương tự với phần thuởng học tăng cường

Phản hồi quy trình - lợi thế

- Tạo tập dữ liệu huấn luyện liên tục •

Nhãn phát triển nhanh

chóng • Thu thập tín hiệu nhãn mạnh

Quá trình phản hồi - nhược điểm

- Bị cản trở bởi bản chất cố hữu của vấn đề •
Không nắm bắt được sự thật cơ bản
- Thiết kế bespoke chủ yếu

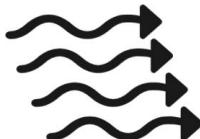
Phản hồi quy trình - Công cụ phân tích nhật ký mã nguồn mở



nhật ký

Quy trình xử lý dữ liệu mã nguồn mở và miễn phí

- Nhập dữ liệu từ vô số nguồn
- Biến đổi nó
- Gửi nó vào "kho" yêu thích của bạn.



lưu loát

Trình thu thập dữ liệu mã nguồn mở

Thông nhất việc thu thập và sử dụng dữ liệu

Phản hồi về quy trình - Phân tích nhật ký đám mây



Phân tích nhật ký đám mây

Nhật ký đám mây của Google

- Dữ liệu và sự kiện từ Google Cloud và AWS
- Mật phỏng ràng buộc. Ghi nhật ký: các thành phần ứng dụng, hệ thống đám mây tại chỗ và lai
- Gửi nó đến "kho" yêu thích của bạn

Tìm kiếm đòn hồi AWS

Màn hình Azure

ghi nhãn con người

Mọi người ("người đánh giá") để kiểm tra dữ liệu và gán nhãn theo cách thủ công



Dữ liệu thô

Dữ liệu không được gắn nhãn và mơ hồ được
gửi đến người đánh giá để chú thích

Tập dữ liệu huấn luyện đã
sẵn sàng để sử dụng

Dán nhãn con người - Phương pháp luận



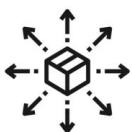
Dữ liệu chưa được gắn nhãn được thu thập



"Người đánh giá" con người được tuyển dụng



Hướng dẫn để hướng dẫn người đánh giá được tạo ra



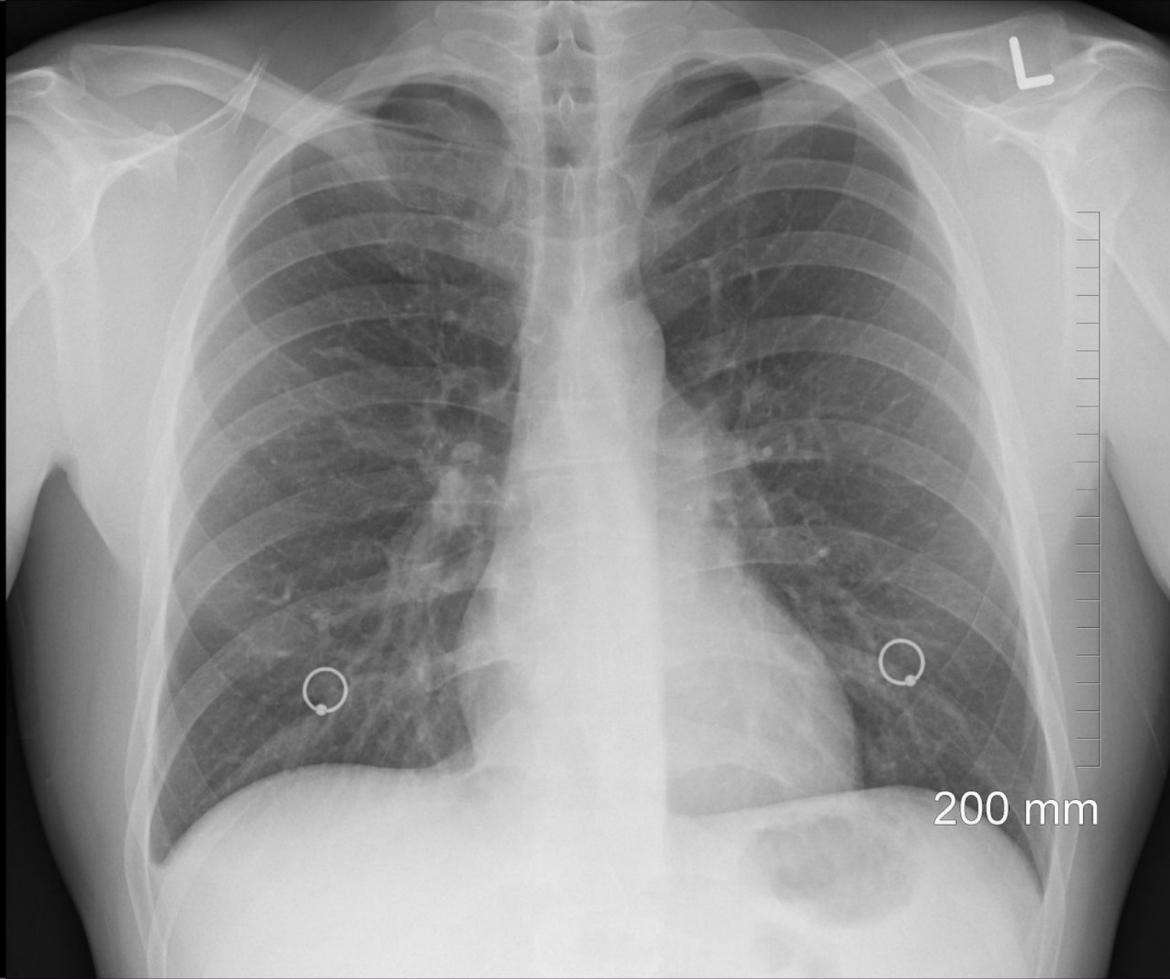
Dữ liệu được phân chia và gán cho người đánh giá



Nhãn được thu thập và xung đột được giải quyết

Nhân nhẫn - lợi thế

- Nhiều nhẫn hơn
- Học có giám sát thuận túy



Nhân nhẫn - Nhược điểm



Tính nhất quán về chất lượng: Nhiều bộ dữ liệu khó dán nhãn cho con người



Chậm

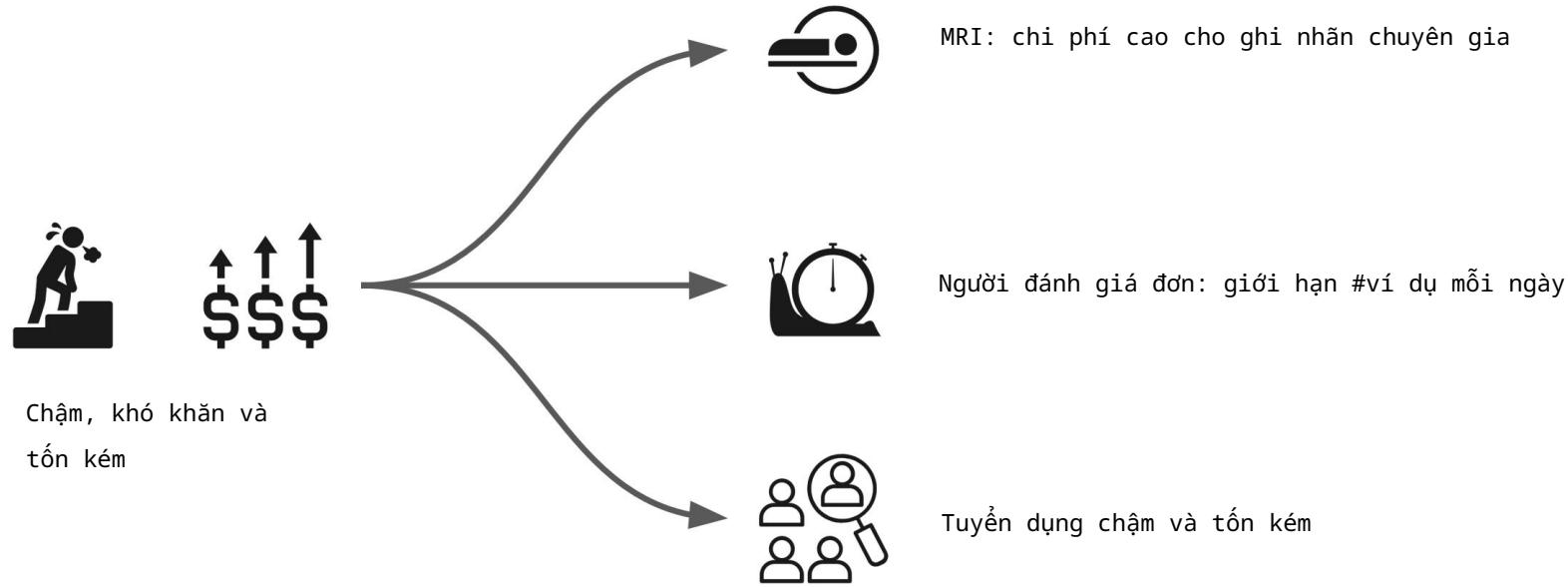


Đắt



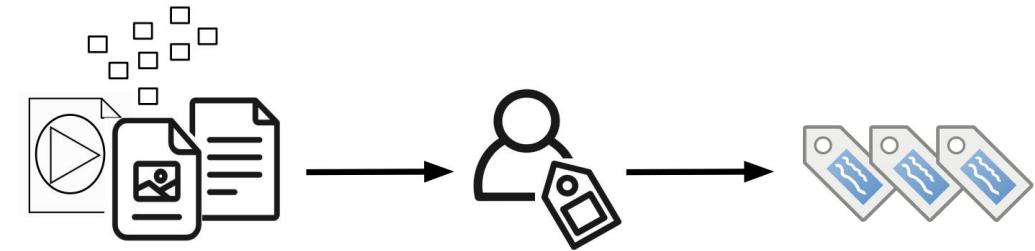
Quản lý tập dữ liệu nhỏ

Tại sao dán nhãn của con người là một vấn đề?



Những điểm chính

- Nhiều phương pháp ghi nhãn dữ liệu
 - Quy trình phản hồi
 - Dán nhãn cho con người
- Ưu và nhược điểm của cả hai





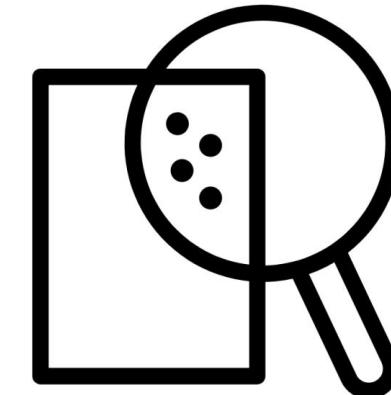
DeepLearning.AI

Xác thực dữ liệu

Phát hiện các vấn đề về dữ liệu

Đề cương

- Vấn đề về dữ liệu
 - Trôi và nghiêng
 - Dữ liệu và khái niệm Trôi dạt
 - Lược đồ nghiêng
 - Độ lệch phân phối
 - Phát hiện các vấn đề về dữ liệu



Trôi và nghiêng

Trôi

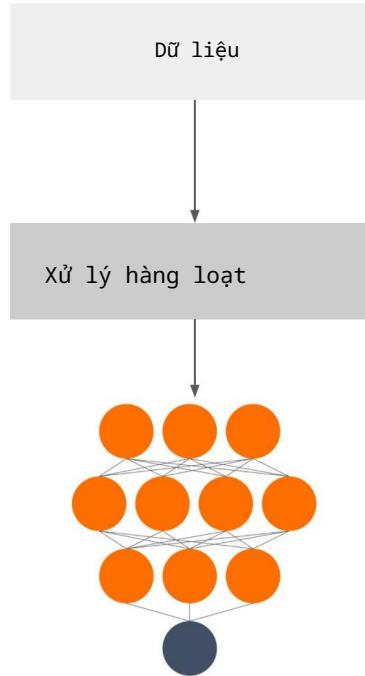
Thay đổi dữ liệu theo thời gian, chẳng hạn như dữ liệu được thu thập mỗi ngày
một lần

Nghiêng

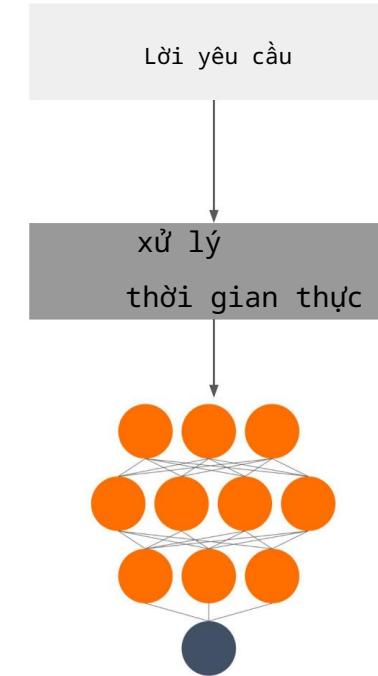
Sự khác biệt giữa hai phiên bản tĩnh hoặc các nguồn khác nhau,
chẳng hạn như tập huấn luyện và tập phục vụ

Đường ống ML điển hình

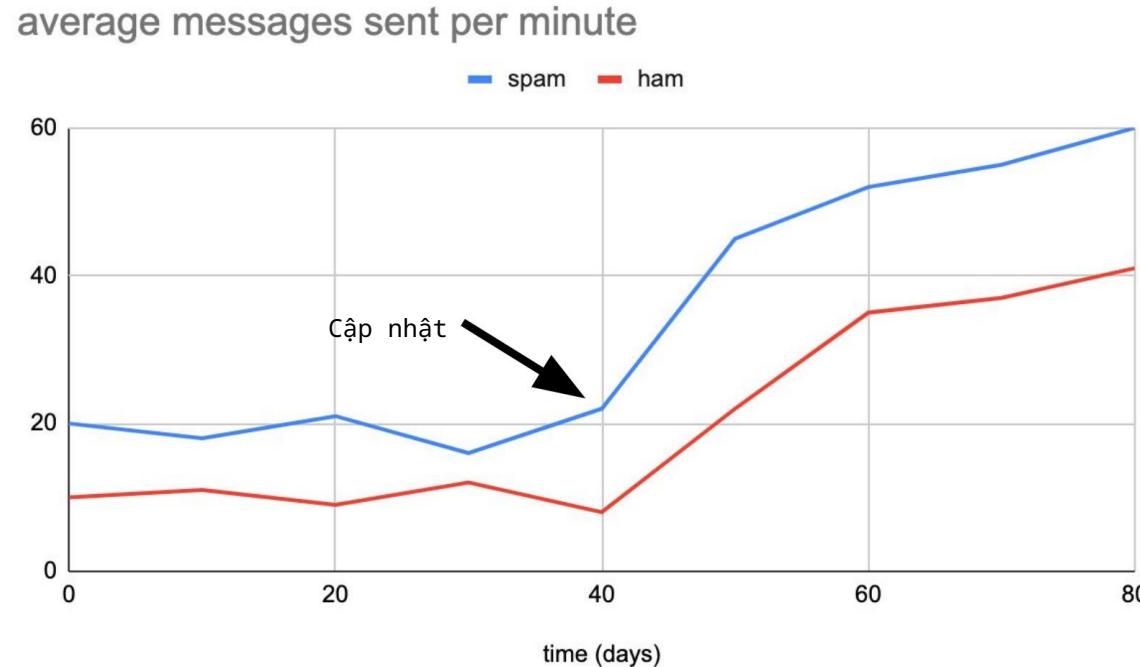
Trong quá trình đào tạo



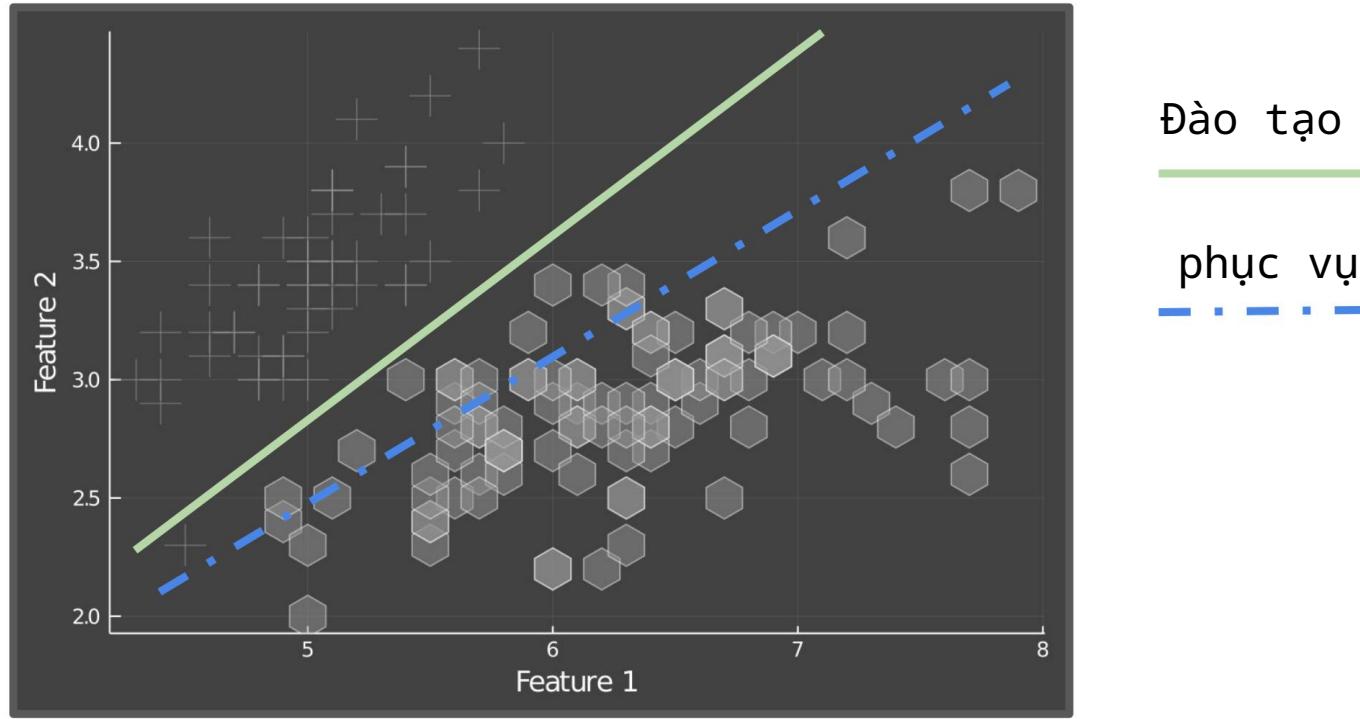
Trong khi phục vụ



Phân rã mô hình: Dữ liệu trôi dạt



Suy giảm hiệu suất : Khái niệm trôi dạt



Phát hiện các vấn đề về dữ liệu

- Phát hiện lệch giản đồ
 - Dữ liệu huấn luyện và phục vụ không giống nhau lược đồ
- Phát hiện độ lệch phân phối
 - Thay đổi tập dữ liệu thay đổi đồng biến hoặc khái niệm
- Yêu cầu đánh giá liên tục

Phát hiện lệch phân phối

	Đào tạo	phục vụ
Chung	$P_{\text{train}}(y, x)$	$P_{\text{serve}}(y, x)$
có điều kiện	$P_{\text{train}}(y x)$	$P_{\text{serve}}(y x)$
cận biên	$P_{\text{train}}(x)$	$P_{\text{serve}}(x)$

thay đổi tập dữ liệu

$$P_{\text{train}}(y, x) \neq P_{\text{serve}}(y, x)$$

dịch chuyển đồng biến

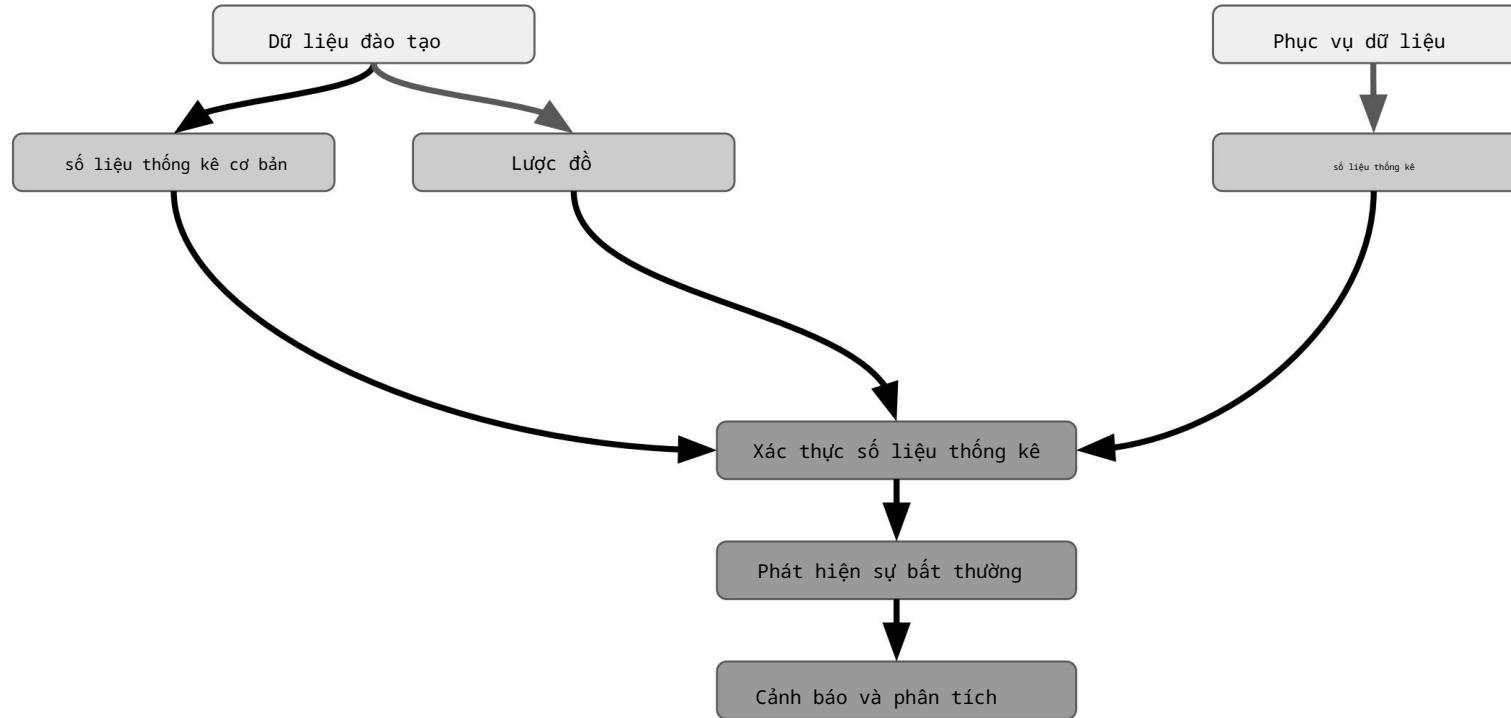
$$P_{\text{train}}(y|x) = P_{\text{serve}}(y|x)$$

thay đổi khái niệm

$$P_{\text{train}}(y|x) \neq P_{\text{serve}}(y|x)$$

$$P_{\text{train}}(x) = P_{\text{serve}}(x)$$

Luồng công việc phát hiện xiên





DeepLearning.AI

Xác thực dữ liệu

TenorFlow

Xác nhận dữ liệu

Xác thực dữ liệu TensorFlow (TFDV)



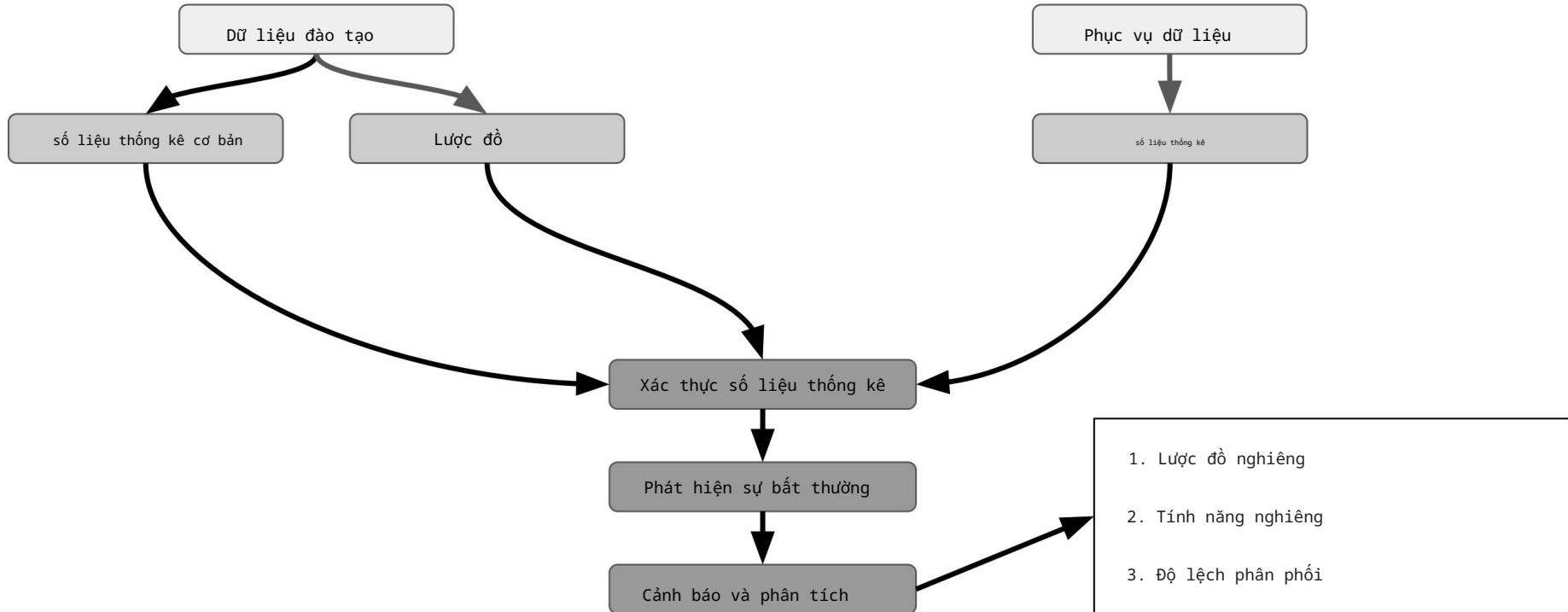
- Hiểu, xác thực và giám sát dữ liệu ML tại tỉ lệ
- Được sử dụng để phân tích và xác thực hàng petabyte dữ liệu tại Google mỗi ngày
- Thành tích đã được chứng minh trong việc giúp người dùng TFX duy trì sức khỏe của đường ống ML của họ

khả năng TFDV

- Tạo thống kê dữ liệu và trực quan hóa trình duyệt
- Suy ra lược đồ dữ liệu
- Thực hiện kiểm tra tính hợp lệ đối với lược đồ •

Phát hiện sai lệch đào tạo/phục vụ

Phát hiện nghiêng - TFDV

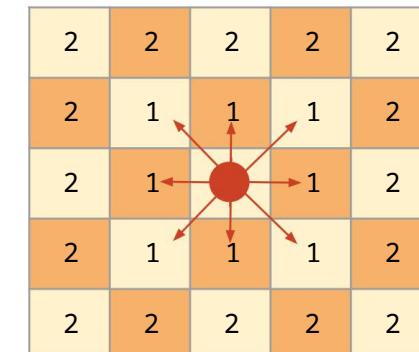


Xiên Xiên - TFDV

- Được hỗ trợ cho các tính năng phân loại
- Biểu thị bằng khoảng cách vô cực L (Khoảng cách Chebyshev) :

$$D_{\text{Chebyshev}}(x, y) = \max_i(|x_i - y_i|)$$

- Đặt ngưỡng nhận cảnh báo



lược đồ nghiêng

Dữ liệu cung cấp và đào tạo không tuân theo cùng một lược đồ:

- Ví dụ: int != float

tính năng nghiêng

Giá trị tính năng đào tạo khác với giá trị tính năng phục vụ:

- Giá trị tính năng được sửa đổi giữa đào tạo và phục vụ thời gian
- Chuyển đổi chỉ được áp dụng ở một trong hai trường hợp

lệch phân phối

Phân phối tập dữ liệu phục vụ và đào tạo khác nhau đáng kể:

- Phương pháp lấy mẫu sai trong quá trình đào tạo
- Các nguồn dữ liệu khác nhau để đào tạo và phục vụ dữ liệu •
Xu hướng, tính thời vụ, thay đổi của dữ liệu theo thời gian

Những điểm chính

- TFDV: Thông kê mô tả theo tỷ lệ với trực quan hóa các khía cạnh được nhúng
- Nó cung cấp cái nhìn sâu sắc về:
 - Số liệu thống kê cơ bản về dữ liệu của bạn là gì
 - Số liệu thống kê về tập dữ liệu đào tạo, đánh giá và phục vụ của bạn so sánh như thế nào
 - Làm cách nào bạn có thể phát hiện và khắc phục sự bất thường của dữ liệu

Gói (lại

- Sự khác biệt giữa lập mô hình ML và hệ thống ML sản xuất
- Thu thập dữ liệu có trách nhiệm để xây dựng hệ thống ML sản xuất công bằng
- Quá trình phản hồi và ghi nhận con người
- Phát hiện các vấn đề về dữ liệu

Thực hành xác thực dữ liệu với TFDV trong vở bài tập tuần này

Kiểm tra kỹ năng của bạn với bài tập lập trình