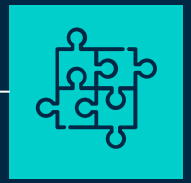
# DIFFERENTIAL PRIVACY using K-ANONYMITY

Toán ứng dụng

# THÔNG TIN NHÓM

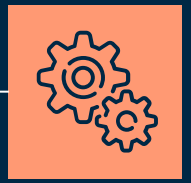
| MSSV | HỌ TÊN | % CÔNG VIỆC | % HOÀN THÀNH |
|---|---|---|---|
| 18120019 | Nguyễn Hoàng Dũng | 25% | 100% |
| 18120040 | Nguyễn Đăng Khoa | 25% | 100% |
| 18120043 | Phạm Minh Khôi | 25% | 100% |
| 18120051 | Nguyễn Hoàng Lân | 25% | 100% |

# MỤC LỤC



01

GIỚI THIỆU DIFFERENTIAL PRIVACY



02

BÀI TOÁN K-ANONYMITY

03

HƯỚNG TIẾP CẬN VÀ THÍ NGHIỆM

01

# GIỚI THIỆU DIFFERENTIAL PRIVACY

# DIFFERENTIAL PRIVACY

Một tiêu chuẩn trong việc bảo vệ thông tin riêng tư, dựa trên các định nghĩa toán học về tính bảo mật đối với kết quả của thuật toán phân tích thống kê và máy học

Một hàm ngẫu nhiên K đạt DP mức độ ϵ cho mọi bộ dữ liệu D1 và D2 khác nhau nhiều nhất một phần tử $S \subseteq \text{Range}(K)$ sao cho:

$$\Pr[K(D1) \in S] \leq \exp(\epsilon) \times \Pr[K(D2) \in S]$$

# CÁC LOẠI DIFFERENTIAL PRIVACY

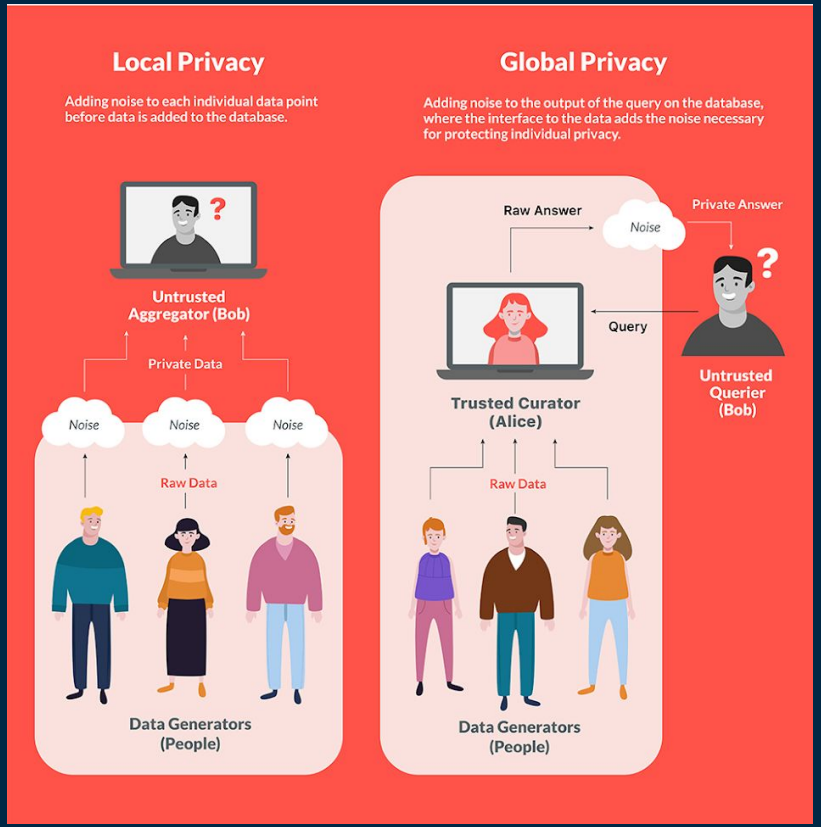
Local Differential Privacy: Các giá trị nhiễu được thêm vào mỗi điểm dữ liệu

Global Differential Privacy: Các giá trị nhiễu được thêm vào thông tin trả về cho các câu truy vấn đến cơ sở dữ liệu

Global DP có thể dẫn đến kết quả truy vấn chính xác hơn, trong khi giữ độ bảo mật tương đương Local DP

Sự hiệu quả đó được đánh đổi bằng niềm tin của cá nhân cung cấp dữ liệu với tổ chức quản lý dữ liệu.

# LOCAL PRIVACY - GLOBAL PRIVACY



# MỘT SỐ THUẬT TOÁN LIÊN QUAN

1. Flip a Coin
2. Phân phối Laplace
3. Cơ chế RAPTOR (2014, Google)
4. Thuật toán RNM (2014, D. Work)
5. Phương pháp PATE (2017, Google)
6. Các thuật toán k-Anonymity
7. Federated Learning

# BÀI TOÁN k-ANONYMITY

02

# GIỚI THIỆU K-ANONYMITY

Khái niệm k-Anonymity lần đầu tiên được đưa ra bởi Latanya Sweeney và Pierangela Samarati trong một bài báo xuất bản năm 1998 là một nỗ lực để giải quyết vấn đề:

> *"Với dữ liệu có cấu trúc trường dành riêng cho từng người, ta tạo ra một mẫu dữ liệu đảm bảo rằng những cá nhân có trong mẫu dữ liệu không thể bị xác định ngược lại đồng thời dữ liệu vẫn còn ý nghĩa trên thực tế."*

*Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression* - Latanya Sweeney và Pierangela, 1998

# ĐỊNH NGHĨA K-ANONYMITY

- k-Anonymity là một thuộc tính được định nghĩa cho các bộ dữ liệu ẩn danh.
- Một mẫu dữ liệu được cho là có thuộc tính k-Anonymity nếu thông tin cho mỗi cá nhân có trong mẫu đó không thể phân biệt với ít nhất k-1 người khác trong mẫu.

*Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression* - Latanya Sweeney và Pierangela, 1998

# VÍ DỤ DỮ LIỆU ĐẠT K-ANONYMITY

| id | age | city_birth | zip_code | disease |
|---|---|---|---|---|
| 1 | 55 | San Giovanni | 17049 | Cancer |
| 2 | 23 | Comina | 26151 | AIDS |
| 3 | 17 | Marina Di Camerota | 73015 | AIDS |
| 4 | 62 | San Giovanni | 17028 | Autism |
| 5 | 40 | Marina Di Camerota | 58014 | Autism |

| id | age | city_birth | zip_code | disease |
|---|---|---|---|---|
| 1 | 50~100 | San Giovanni | 17*** | Cancer |
| 2 | 0~50 | Italy | ***** | AIDS |
| 3 | 0~50 | Italy | ***** | AIDS |
| 4 | 50~100 | San Giovanni | 17*** | Autism |
| 5 | 0~50 | Italy | ***** | Autism |

Dữ liệu trước và sau khi sử dụng Thuật toán Ẩn danh để đạt được 2-Anonymity

# MỘT SỐ ĐỊNH NGHĨA

| | |
|---|---|
| **Quasi-Identifier** (QIDs) | Một nhóm các thuộc tính trong bảng dữ liệu, khi được kết hợp với nhau, thì tồn tại mẫu dữ liệu có thể định danh một cá nhân nào đó trên thực tế. |
| **Sensitive Attributes** (SA) | Các thuộc tính chứa các thông tin đặc biệt nhạy cảm của một cá nhân. Một cá nhân chắc chắn sẽ không muốn bị lộ thông tin này. |
| **Non-Sensitive Attributes** (non-SA) | Những thuộc tính còn lại không có tính chất đặc biệt. |
| **Equivalence Class** (EQ) | Một bộ thực thể có giá trị các thuộc tính QID giống nhau |

# CÁC ĐỊNH NGHĨA CHẶT CHẼ HƠN

ℓ-Diversity: Đảm bảo bằng mỗi nhóm đã đạt k-Anonymity tồn tại thêm ít nhất l giá trị SA khác nhau.

*l-Diversity: Privacy beyond k-Anonymity - Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, 2007.*

# CÁC ĐỊNH NGHĨA CHẶT CHẼ HƠN

- t-Closeness: Mở rộng ra từ l-Diversity kết hợp thêm thông tin phân bố của dữ liệu, đảm bảo rằng phân bố xác suất của các SA trong một nhóm k-ẩn danh sẽ tương đồng với phân bố của thuộc tính đó trong toàn bộ dữ liệu

*t-Closeness: Privacy beyond k-Anonymity and l-Diversity: Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, 2007*

# CÁC PHƯƠNG PHÁP CHUNG

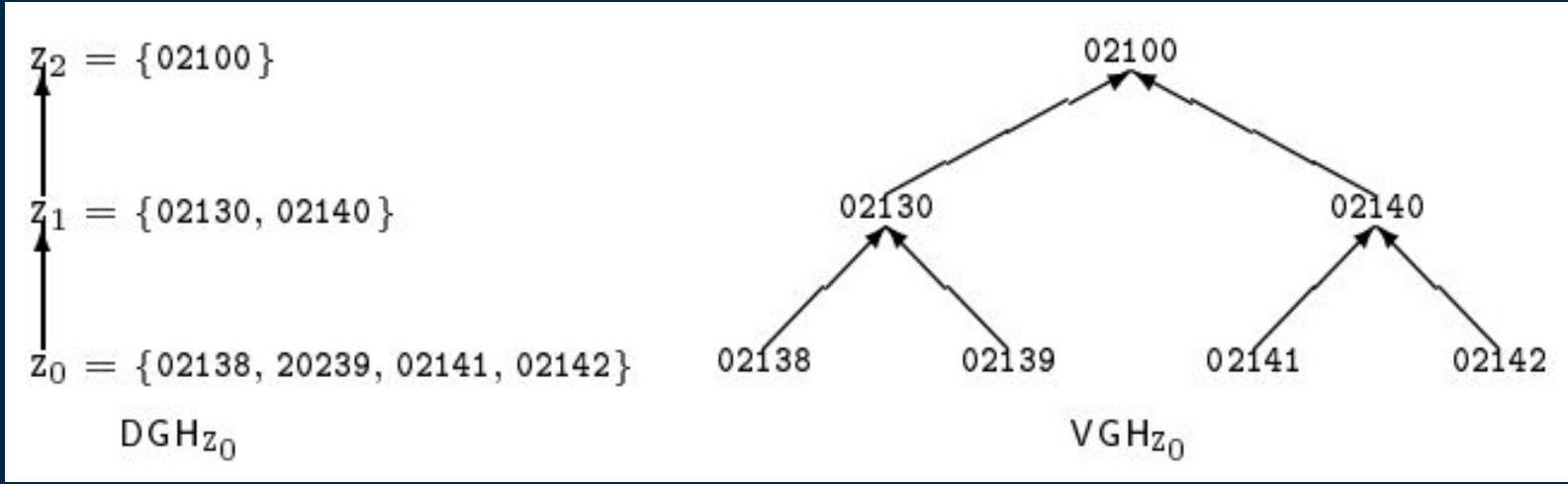
Tổng quát hóa (Generalization):

- Thay thế giá trị của thuộc tính bằng giá trị tổng quát hơn
- Tận dụng phân cấp tổng quát theo miền giá trị của thuộc tính (domain/value generalization hierarchies)

Áp chế (Suppression):

- Loại bỏ đi mẫu dữ liệu không thể thỏa ràng buộc

# PHƯƠNG PHÁP TỔNG QUÁT HÓA




Ví dụ: Phương pháp tổng quát hóa

# PHƯƠNG PHÁP TỔNG QUÁT HÓA

Tổng quát hóa được chia thành các loại:

- Toàn cục:
  - Đơn chiều
  - Đa chiều
- Địa phương

# PHƯƠNG PHÁP TỔNG QUÁT HÓA

| Row-id | Age | Zipcode |
|---|---|---|
| $R1$ | 24 | 53712 |
| $R2$ | 25 | 53711 |
| $R3$ | 30 | 53711 |
| $R4$ | 30 | 53711 |
| $R5$ | 32 | 53712 |
| $R6$ | 32 | 53713 |

Dữ liệu gốc

| Row-id | Age | Zipcode |
|---|---|---|
| $R1$ | [24-30] | [53711-53712] |
| $R2$ | [24-30] | [53711-53712] |
| $R3$ | [24-30] | [53711-53712] |
| $R4$ | [30-32] | [53711-53713] |
| $R5$ | [30-32] | [53711-53713] |
| $R6$ | [30-32] | [53711-53713] |

Tổng quát hóa toàn cục

| Row-id | Age | Zipcode |
|---|---|---|
| $R1$ | [24-32] | [53712-53713] |
| $R2$ | [25-30] | 53711 |
| $R3$ | [25-30] | 53711 |
| $R4$ | [25-30] | 53711 |
| $R5$ | [24-32] | [53712-53713] |
| $R6$ | [24-32] | [53712-53713] |

Tổng quát hóa địa phương

# PHƯƠNG PHÁP TỔNG QUÁT HÓA - TOÀN CỤC

Dữ liệu gốc

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| 25 | Male | 53711 | Flu |
| 25 | Female | 53712 | Hepatitis |
| 26 | Male | 53711 | Brochitis |
| 27 | Male | 53710 | Broken Arm |
| 27 | Female | 53712 | AIDS |
| 28 | Male | 53711 | Hang Nail |

Tổng quát hóa đơn chiều

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| [25-28] | Male | [53710-53711] | Flu |
| [25-28] | Female | 53712 | Hepatitis |
| [25-28] | Male | [53710-53711] | Brochitis |
| [25-28] | Male | [53710-53711] | Broken Arm |
| [25-28] | Female | 53712 | AIDS |
| [25-28] | Male | [53710-53711] | Hang Nail |

Tổng quát hóa đa chiều

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| [25-26] | Male | 53711 | Flu |
| [25-27] | Female | 53712 | Hepatitis |
| [25-26] | Male | 53711 | Brochitis |
| [27-28] | Male | [53710-53711] | Broken Arm |
| [25-27] | Female | 53712 | AIDS |
| [27-28] | Male | [53710-53711] | Hang Nail |

# CÁC THUẬT TOÁN K-ANONYMITY

Nhóm xem xét 5 thuật toán k-Anonymity

# MỘT SỐ THUẬT TOÁN K-ANONYMITY

- Thuật toán Datafly (*1997, Latanya Arvette Sweeney*)
- Thuật toán Incognito (*2005, Kristen LeFevre*)
- Thuật toán Top-Down Greedy (*2006, J. Xu*)
- Thuật toán dựa trên KNN (*2008, Jun-Lin Lin và Meng-Cheng We*)
- Thuật toán Mondrian (*2006, Kristen LeFevre*)

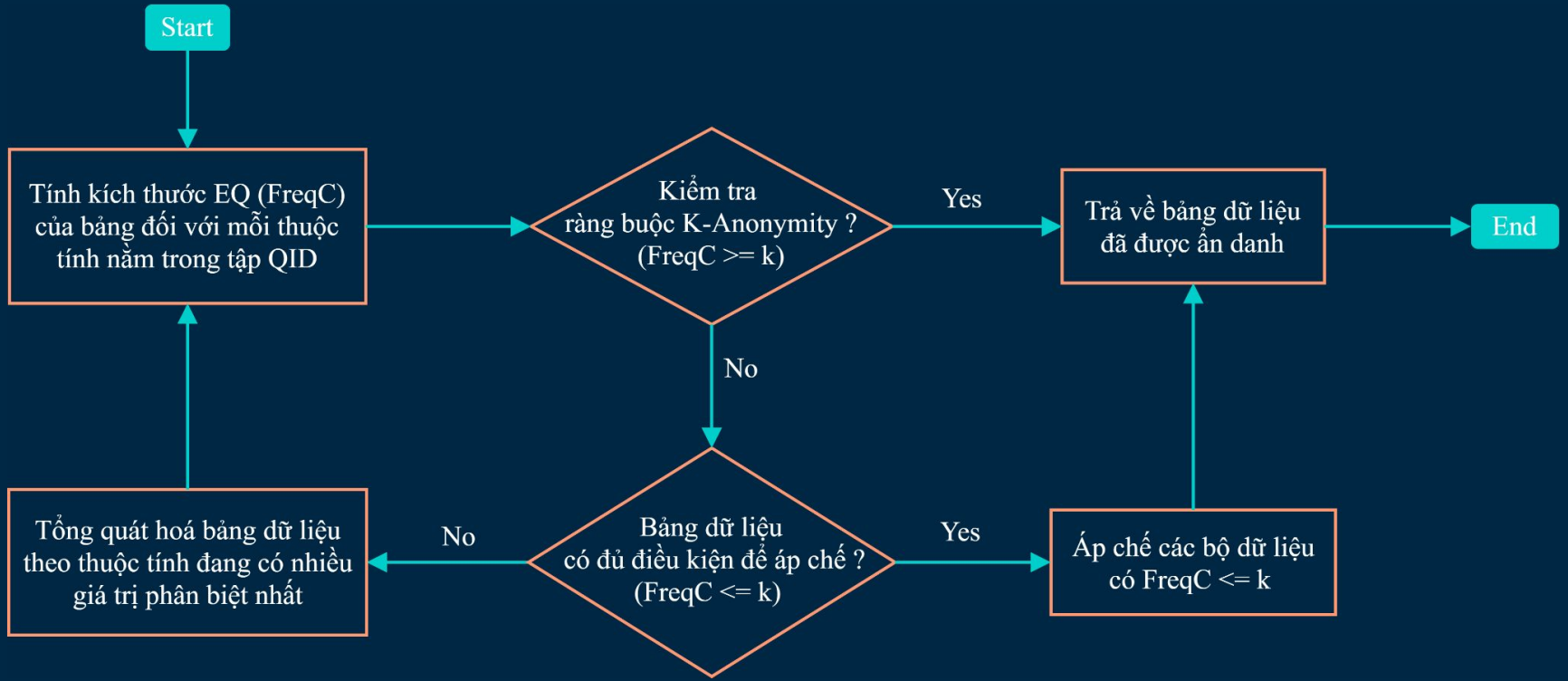
# MỘT SỐ THUẬT TOÁN K-ANONYMITY

- Thuật toán Datafly (*1997, Latanya Arvette Sweeney*)
- Thuật toán Incognito (*2005, Kristen LeFevre*)
- Thuật toán Top-Down Greedy (*2006, J. Xu*)
- Thuật toán dựa trên KNN (*2008, Jun-Lin Lin và Meng-Cheng We*)
- Thuật toán Mondrian (*2006, Kristen LeFevre*)

# THUẬT TOÁN DATAFLY

- Thuật toán lựa chọn các thuộc tính có nhiều giá trị khác biệt nhất và thực hiện các việc tổng quát hóa, thay thế, thêm, xóa, sửa.
- Thuật toán hoạt động dựa trên toàn bộ cây phân tầng và tiếp cận theo phương pháp tham lam kết hợp heuristic. Do đó, thuật toán chỉ duyệt qua một số đỉnh để đưa ra đáp án
- Thuật toán này rất hiệu quả về mặt về thời gian, tuy nhiên gặp phải hạn chế là có thể mắc kẹt ở cực tiểu địa phương.

# THUẬT TOÁN DATAFLY



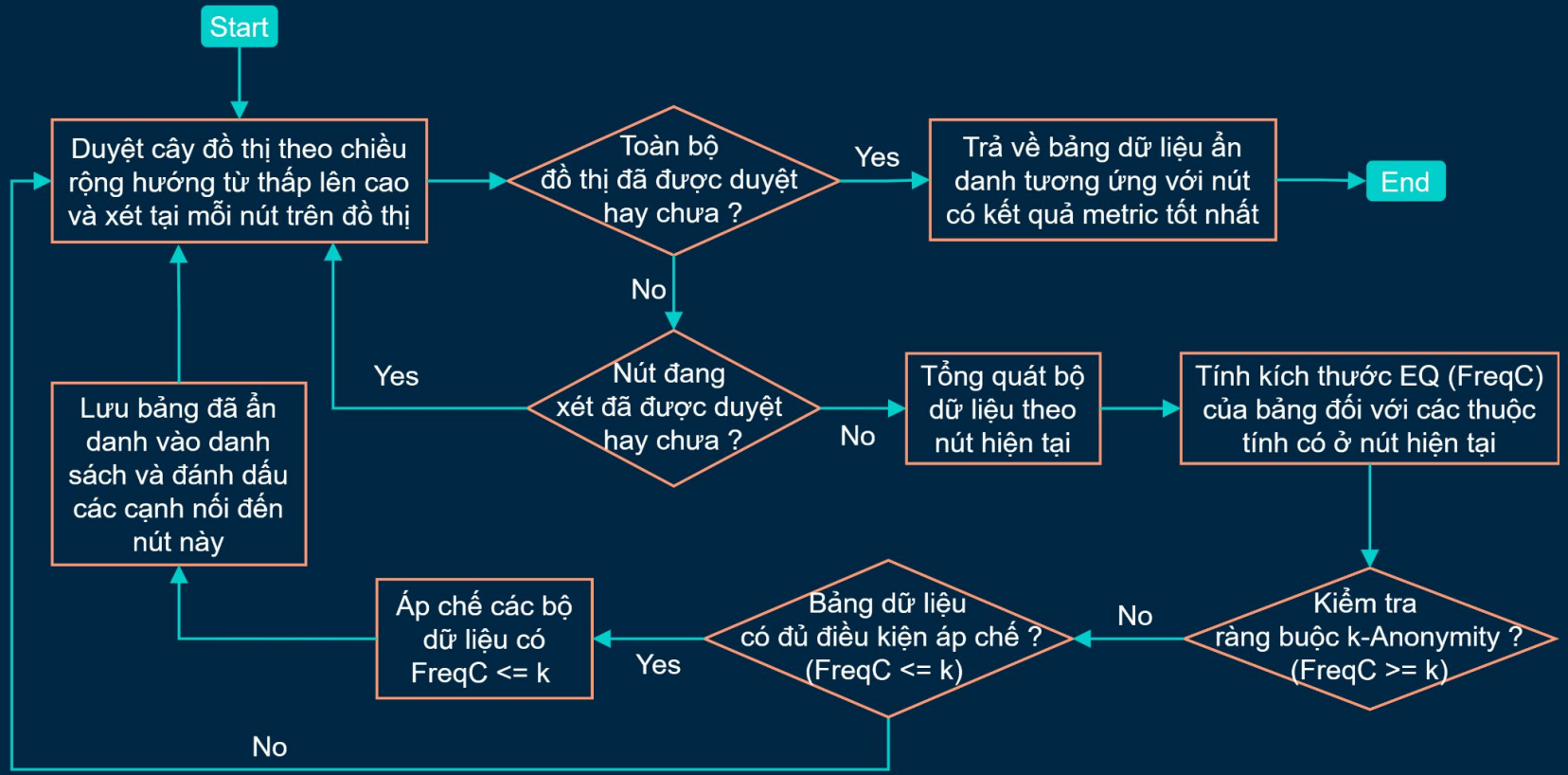
# MỘT SỐ THUẬT TOÁN K-ANONYMITY

- Thuật toán Datafly (*1997, Latanya Arvette Sweeney*)
- Thuật toán Incognito (*2005, Kristen LeFevre*)
- Thuật toán Top-Down Greedy (*2006, J. Xu*)
- Thuật toán dựa trên KNN (*2008, Jun-Lin Lin và Meng-Cheng We*)
- Thuật toán Mondrian (*2006, Kristen LeFevre*)

# THUẬT TOÁN INCOGNITO

- Thuật xây dựng một đồ thị dựa trên các mức độ tổng quát của thuộc tính và duyệt qua đồ thị đó bằng phương pháp duyệt theo chiều rộng (breadth-first search) theo hướng bottom-up. Sau khi duyệt qua toàn bộ đồ thị sẽ tìm được một bảng dữ liệu ẩn danh tương ứng theo từng node đồ thị.
- Incognito có thể tìm được nhiều bảng dữ liệu thỏa k-anonymity, sau đó áp dụng các hàm đánh giá để tìm ra bảng có điểm tốt nhất

# THUẬT TOÁN INCOGNITO



# MỘT SỐ THUẬT TOÁN K-ANONYMITY

- Thuật toán Datafly (*1997, Latanya Arvette Sweeney*)
- Thuật toán Incognito (*2005, Kristen LeFevre*)
- Thuật toán Top-Down Greedy (*2006, J. Xu*)
- Thuật toán dựa trên KNN (*2008, Jun-Lin Lin và Meng-Cheng We*)
- Thuật toán Mondrian (*2006, Kristen LeFevre*)

# THUẬT TOÁN TOP-DOWN GREEDY

- Thuật toán phân vùng nhị phân đệ quy dữ liệu thành các lớp tương đương càng lúc càng cục bộ hơn.
- Sau khi phân vùng hoàn tất, tất cả các lớp tương đương chứa ít hơn k phần tử được hậu xử lý để đạt được k-anonymity.
- Đối với mỗi lớp tương đương G như vậy, hai bước tiếp theo được áp dụng: Bước đầu tiên là tìm kiếm trong tất cả các lớp tương đương có kích thước ít nhất là 2k−|G|, và tập con Gs các bộ có kích thước k−|G| với độ lỗi thấp nhất (G ∪ Gs).
- Giải pháp có độ lỗi tổng thể nhỏ nhất sẽ được áp dụng và lặp lại

# MỘT SỐ THUẬT TOÁN K-ANONYMITY

- Thuật toán Datafly (*1997, Latanya Arvette Sweeney*)
- Thuật toán Incognito (*2005, Kristen LeFevre*)
- Thuật toán Top-Down Greedy (*2006, J. Xu*)
- Thuật toán dựa trên KNN (*2008, Jun-Lin Lin và Meng-Cheng We*)
- Thuật toán Mondrian (*2006, Kristen LeFevre*)

# THUẬT TOÁN DỰA TRÊN KNN

- Ở mỗi vòng lặp, một bộ dữ liệu sẽ được chọn ngẫu nhiên và k−1 bộ gần nhất dựa trên một hàm khoảng cách sẽ được chọn cùng. Những bộ này sẽ được gán vào một lớp tương đương và xóa bỏ khỏi dữ liệu gốc.
- Quy trình này được lặp lại tới khi tất cả bộ đã được xử lý và thuộc về một lớp.

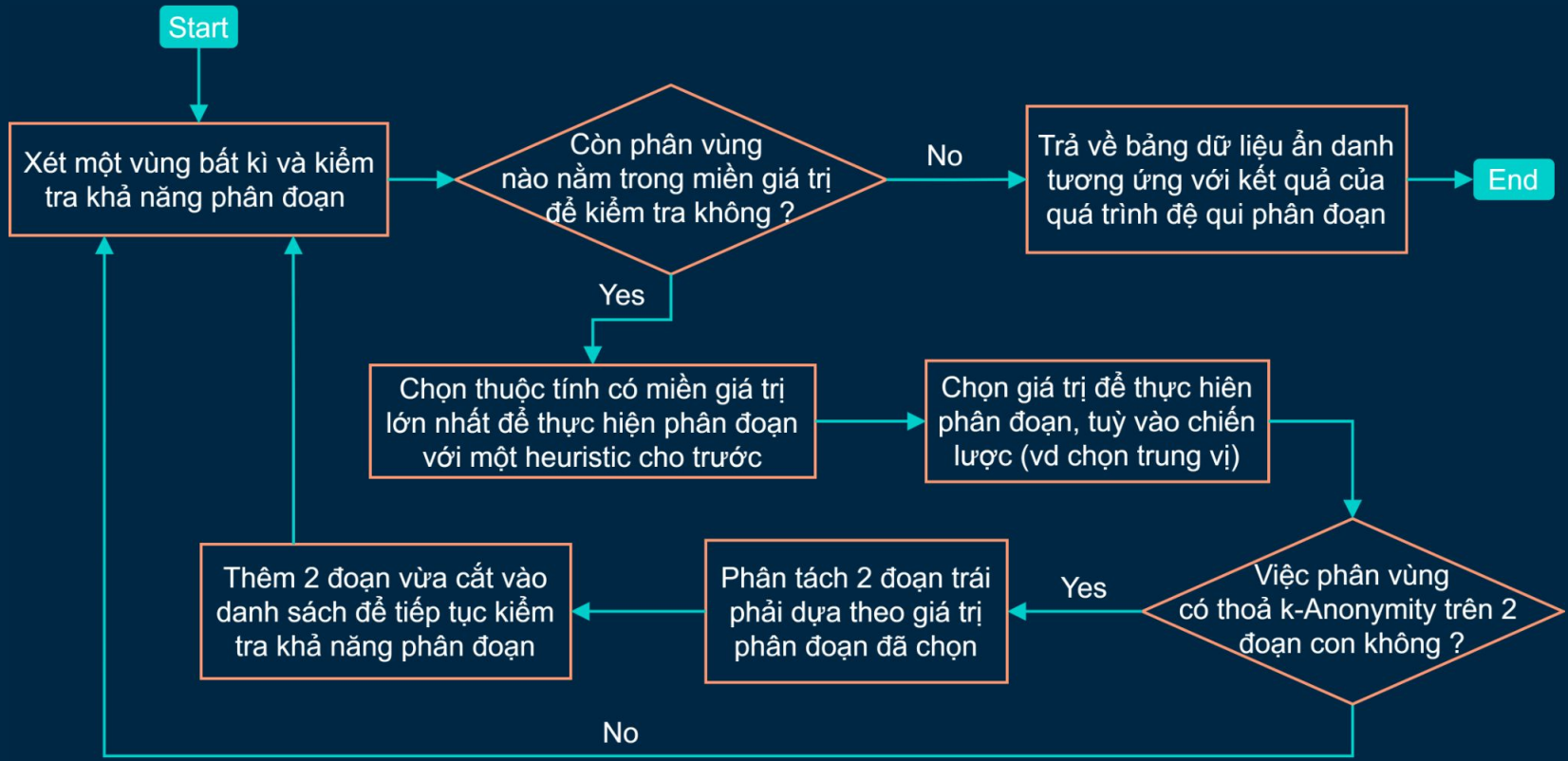
# MỘT SỐ THUẬT TOÁN K-ANONYMITY

- Thuật toán Datafly (*1997, Latanya Arvette Sweeney*)
- Thuật toán Incognito (*2005, Kristen LeFevre*)
- Thuật toán Top-Down Greedy (*2006, J. Xu*)
- Thuật toán dựa trên KNN (*2008, Jun-Lin Lin và Meng-Cheng We*)
- Thuật toán Mondrian (*2006, Kristen LeFevre*)

# THUẬT TOÁN MONDRIAN

- Tổng quát hóa QIDs từ mức tổng quát cao nhất và đệ qui dần vào các vùng nhỏ hơn bằng các phép cắt đa chiều cho đến khi không thể cắt được nữa.
- Mỗi vòng lặp của thuật toán chọn ra một chiều (thuộc tính) để thực hiện việc cắt. Việc chọn sẽ ưu tiên chiều có miền giá trị lớn nhất. Vị trí cắt sẽ là vị trí trung vị của miền, phép cắt sẽ chia đôi miền này.
- Lặp lại đến khi k-anonymity không thỏa nữa.

# THUẬT TOÁN MONDRIAN



# HƯỚNG TIẾP CẬN & THỰC NGHIỆM

03

# BỘ DỮ LIỆU

Nhóm thực nghiệm trên 6 bộ dữ liệu chính

# BỘ DỮ LIỆU

- The Adult Dataset (ADULT)
- The California Housing Price Dataset (CAHOUSING)
- The Contraceptive Method Choice Dataset (CMC)
- The Mammographic Mass Dataset (MGM)
- INFORMS Data Mining Dataset (INFORMS)
- Italia Healthcare Dataset (ITALIA)

# HÀM ĐÁNH GIÁ MÔ HÌNH

Nhóm cài đặt 3 hàm đánh giá mô hình ẩn danh

# HÀM ĐÁNH GIÁ MÔ HÌNH

- Discernibility Metric (DM)
- Average Equivalence class size Metric (CAV G)
- Normalized Certainty Penalty (NCP)

# HÀM ĐÁNH GIÁ MÔ HÌNH

- Discernibility Metric (DM)
- Average Equivalence class size Metric (CAV G)
- Normalized Certainty Penalty (NCP)

# DISCERNIBILITY METRIC (DM)

- Tính toán khả năng phân biệt giữa các thực thể của dữ liệu trong bảng T
- Ý tưởng đằng sau metric này là với EQ càng lớn nghĩa là độ mất mát thông tin càng lớn

$$DM(T) = \sum_{EQ \subset T, |EQ| \geq k} |EQ|^2 + \sum_{EQ \subset T, |EQ| < k} |EQ| * |T|$$

# HÀM ĐÁNH GIÁ MÔ HÌNH

- Discernibility Metric (DM)
- Average Equivalence class size Metric (CAV G)
- Normalized Certainty Penalty (NCP)

# AVERAGE EQUIVALENCE CLASS SIZE METRIC ($C_{AVG}$)

- Metric này mô tả việc tạo ra các EQ tốt như thế nào.
- Giá trị bằng 1 ám chỉ độ tổng quát hóa tối ưu khi kích thước của tất cả các EQ bằng giá trị k.

$$C_{AVG}(T) = \frac{|T|}{|EQs| \times k}$$

# HÀM ĐÁNH GIÁ MÔ HÌNH

- Discernibility Metric (DM)
- Average Equivalence class size Metric (CAV G)
- Normalized Certainty Penalty (NCP)

# NORMALIZED CERTAINTY PENALTY (NCP)

- NCP được tính dựa trên cây phân tầng tổng quát hóa, được hiểu đơn giản là dùng khoảng giá trị của thuộc tính của các bộ chia cho khoảng giá trị của thuộc tính của toàn bộ dữ liệu
- Do các thuộc tính có hai loại là rời rạc (ví dụ như zipcode, giới tính) và liên tục (số tuổi, thu nhập), nên khoảng giá trị của chúng sẽ được tính toán khác nhau

# NORMALIZED CERTAINTY PENALTY (NCP)

- Đối với thuộc tính liên tục

$$NCP_{A_i}(t) = \frac{r_i - l_i}{|A_i|}$$

- Đối với thuộc tính rời rạc

$$NCP_{A_i}(t) = \frac{leaf(u)}{|A_i|}$$

- Kết hợp 2 biểu thức trên, đối với một bảng dữ liệu tổng quát T, tổng độ lỗi được tính như sau

$$NCP(T) = \frac{1}{|T| \times n} \sum_{t \in T} \sum_{i=1}^{n} NCP_{A_i}(t)$$

# CÁC MÔ HÌNH MÁY HỌC

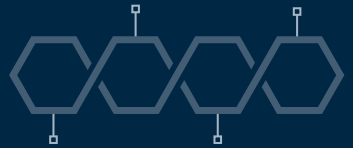
Nhóm cài đặt 3 mô hình máy học để đánh giá các mô hình ẩn danh

# CÁC MÔ HÌNH MÁY HỌC

- K-nearest neighbors (KNN)
- Support Vector Machines (SVMs)
- Random Forests (RFs)

# THỰC NGHIỆM

Chi tiết thực nghiệm của nhóm



# THỰC NGHIỆM

- Tiến hành thử nghiệm 5 phương pháp ẩn danh k-anonymity trên 6 bộ dữ liệu và quan sát các điểm metric của từng cặp.

- Giá trị k được chọn lần lượt trong khoảng khác nhau để quan sát khả năng của thuật toán ẩn danh với ràng buộc càng lúc càng khó hơn

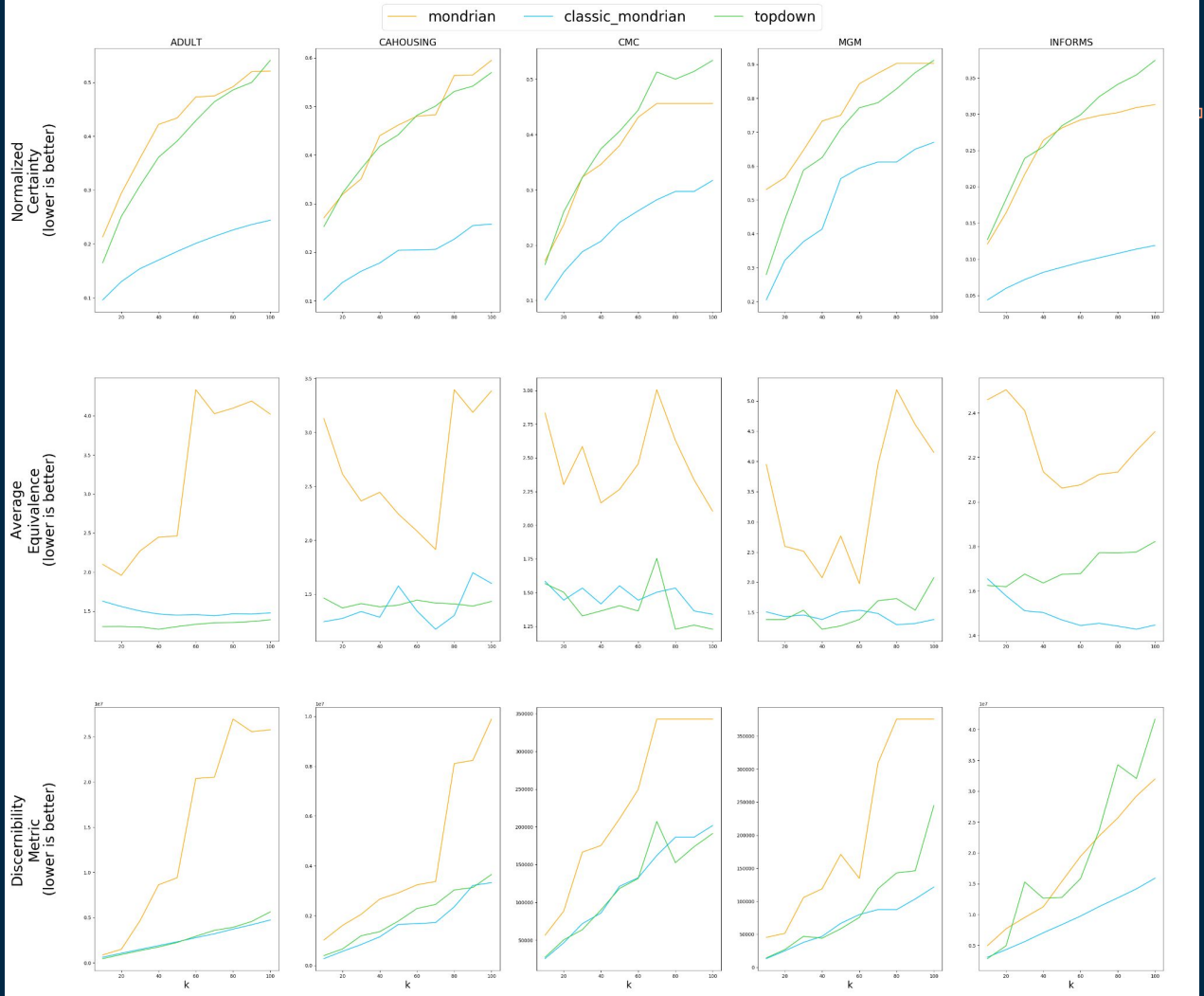
# THỰC NGHIỆM

- Đánh giá mô hình:
  - Dựa trên số lượng thông tin mất mát
  - Dựa trên ý nghĩa thông tin mất mát
- Kết luận

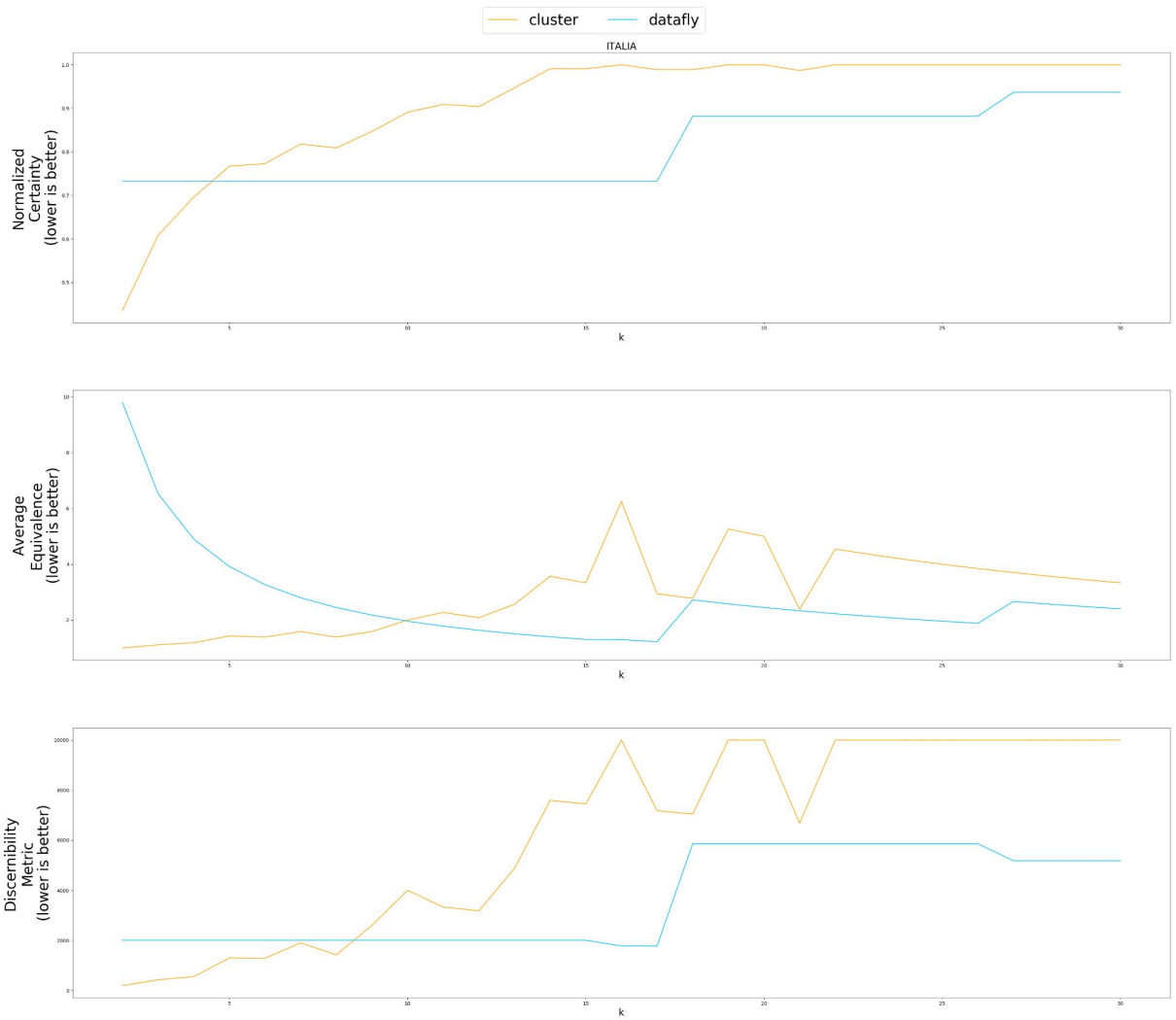
# THỰC NGHIỆM

- Đánh giá mô hình:
  - Dựa trên số lượng thông tin mất mát
  - Dựa trên ý nghĩa thông tin mất mát
- Kết luận

# KẾT QUẢ


mondrian
classic_mondrian
topdown
ADULT
CAHOUSING
CMC
MGM
INFORMS
Normalized Certainty (lower is better)
Average Equivalence (lower is better)
Discernibility Metric (lower is better)
k

# KẾT QUẢ


cluster
datafly
ITALIA
Normalized Certainty (lower is better)
k
Average Equivalence (lower is better)
k
Discernibility Metric (lower is better)
k

# THỰC NGHIỆM

- Đánh giá mô hình:
  - Dựa trên số lượng thông tin mất mát
  - Dựa trên ý nghĩa thông tin mất mát
- Kết luận

# ONE-HOT ENCODING

- Để đưa dữ liệu vào các mô hình máy học yêu cầu việc tiền xử lý, số hóa các giá trị.
- Đối với các thuộc tính liên tục, ta có thể giữ nguyên để đưa vào mô hình, tuy nhiên để xử lý giá trị liên tục đã bị tổng quát hóa, nhóm lấy giá trị trung vị của khoảng này.
- Đối với các giá trị là thuộc tính rời rạc, nhóm tiến hành one-hot encoding thuộc tính này.

# ONE-HOT ENCODING

| id | age | city_birth | zip_code | disease |
|---|---|---|---|---|
| 1 | 55 | San Giovanni | 17049 | Cancer |
| 2 | 23 | Comina | 26151 | AIDS |
| 3 | 17 | Marina Di Camerota | 73015 | AIDS |
| 4 | 62 | San Giovanni | 17028 | Autism |
| 5 | 40 | Marina Di Camerota | 58014 | Autism |

A

| id | age | city_birth_San Giovanni | city_birth_Comina (La) | city_birth_Marina Di Camerota | zip_code | disease |
|---|---|---|---|---|---|---|
| 1 | 55 | 1 | 0 | 0 | 17049 | Cancer |
| 2 | 23 | 0 | 1 | 0 | 26151 | AIDS |
| 3 | 17 | 0 | 0 | 1 | 73015 | AIDS |
| 4 | 62 | 1 | 0 | 0 | 17028 | Autism |
| 5 | 40 | 0 | 0 | 1 | 58014 | Autism |

B

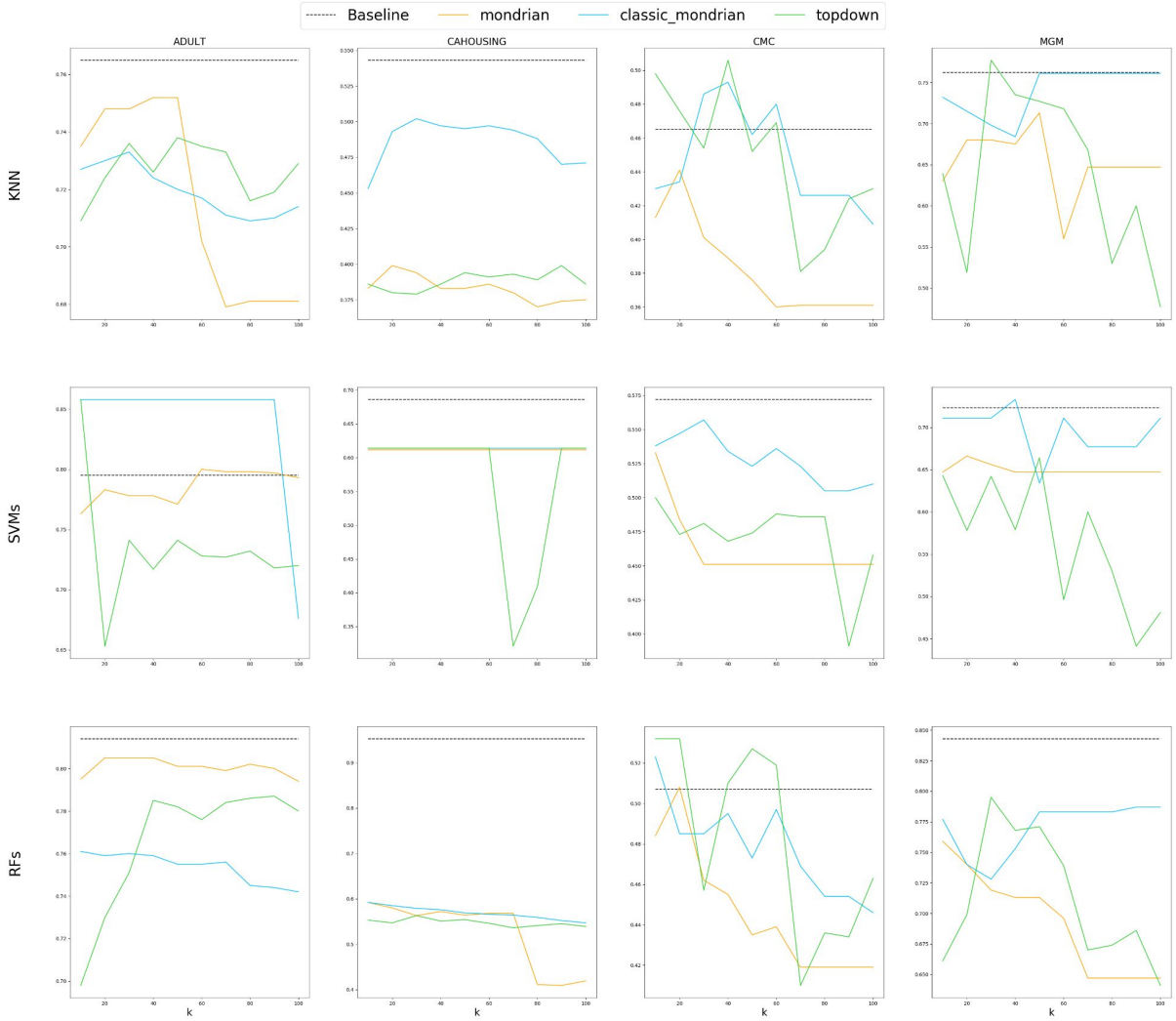
| id | age | city_birth | zip_code | disease |
|---|---|---|---|---|
| 1 | 50~100 | San Giovanni | 17*** | Cancer |
| 2 | 0~50 | Italy | ***** | AIDS |
| 3 | 0~50 | Italy | ***** | AIDS |
| 4 | 50~100 | San Giovanni | 17*** | Autism |
| 5 | 0~50 | Italy | ***** | Autism |

C

| id | age | city_birth_San Giovanni | city_birth_Comina (La) | city_birth_Marina Di Camerota | zip_code | disease |
|---|---|---|---|---|---|---|
| 1 | 75 | 1 | 0 | 0 | 17499.5 | Cancer |
| 2 | 25 | 1 | 1 | 1 | 49999.5 | AIDS |
| 3 | 25 | 1 | 1 | 1 | 49999.5 | AIDS |
| 4 | 75 | 1 | 0 | 0 | 17499.5 | Autism |
| 5 | 25 | 1 | 1 | 1 | 49999.5 | Autism |

**A: One-hot encoding**
**B: 2-anonymize**
**C: Deanonymize one-hot encoding**

# KẾT QUẢ


Baseline
mondrian
classic_mondrian
topdown
ADULT
CAHOUSING
CMC
MGM
KNN
SVMs
RFs
k
k
k
k

# KẾT QUẢ

# THỰC NGHIỆM

- Đánh giá mô hình:
  - Dựa trên số lượng thông tin mất mát
  - Dựa trên ý nghĩa thông tin mất mát
- Kết luận

# KẾT LUẬN

- Tác động qua lại giữa phương pháp ẩn danh và mô hình máy học
- Cái nhìn tổng thể về sự nhạy cảm của việc thông tin bị thay đổi (bằng cách tổng quát hóa hoặc loại bỏ) có thể ảnh hưởng đến chất lượng dữ liệu
- Nếu những giá trị nhãn (target variable) có độ tương quan cao với các QID bị tổng quát hóa mạnh thì sẽ gây ảnh hưởng đến khả năng học của mô hình huấn luyện

# KẾT LUẬN

- Với mỗi bộ dữ liệu khác nhau tồn tại nhiều cách để đạt được k-anonymity khác nhau. Thuật toán được chọn nên mang lại sự tổng quát hóa ít nhất đối với các thuộc tính tương quan với nhãn

- Lựa chọn giá trị k cũng quan trọng. Cần phải thử nghiệm nhiều hơn.

Cảm ơn các thầy và trợ giảng đã lắng nghe!

# THE END

Nếu có câu hỏi, liên hệ email: 18120043@student.hcmus.edu.vn
Mã nguồn: https://github.com/kaylode/k-anonymity