



Bài giảng môn học:
Khai phá dữ liệu (7080508)

CHƯƠNG 2: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU (Phần 1)

08/2024



Nội dung chương 2 - Phần 1

1.1 Phương pháp thu thập dữ liệu

1.2 Lấy mẫu (Sampling)

1.3 Tích hợp dữ liệu

1.4 Bài tập thực hành



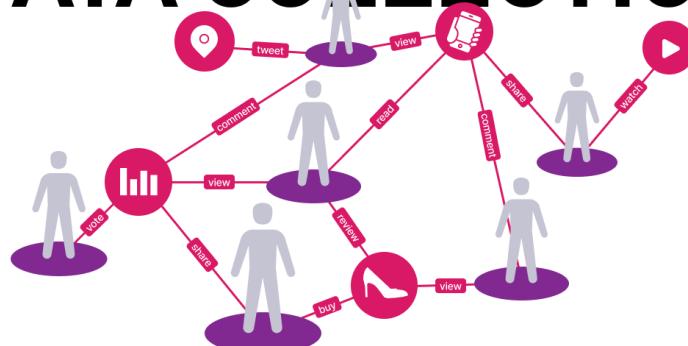
2.1 Phương pháp thu thập dữ liệu



Phương pháp thu thập dữ liệu

Khi xác định rõ mục tiêu của dự án, vấn đề quan trọng là cần **thu thập tất cả các dữ liệu liên quan**. Dữ liệu thu thập **đầy đủ và chính xác** sẽ quyết định đến **chất lượng của kết quả Phân tích dữ liệu**.

DATA COLLECTION



**Accuracy, Accuracy,
Accuracy**

Any raw, inaccurate data introduced will devalue all future work.

Phương pháp thu thập dữ liệu

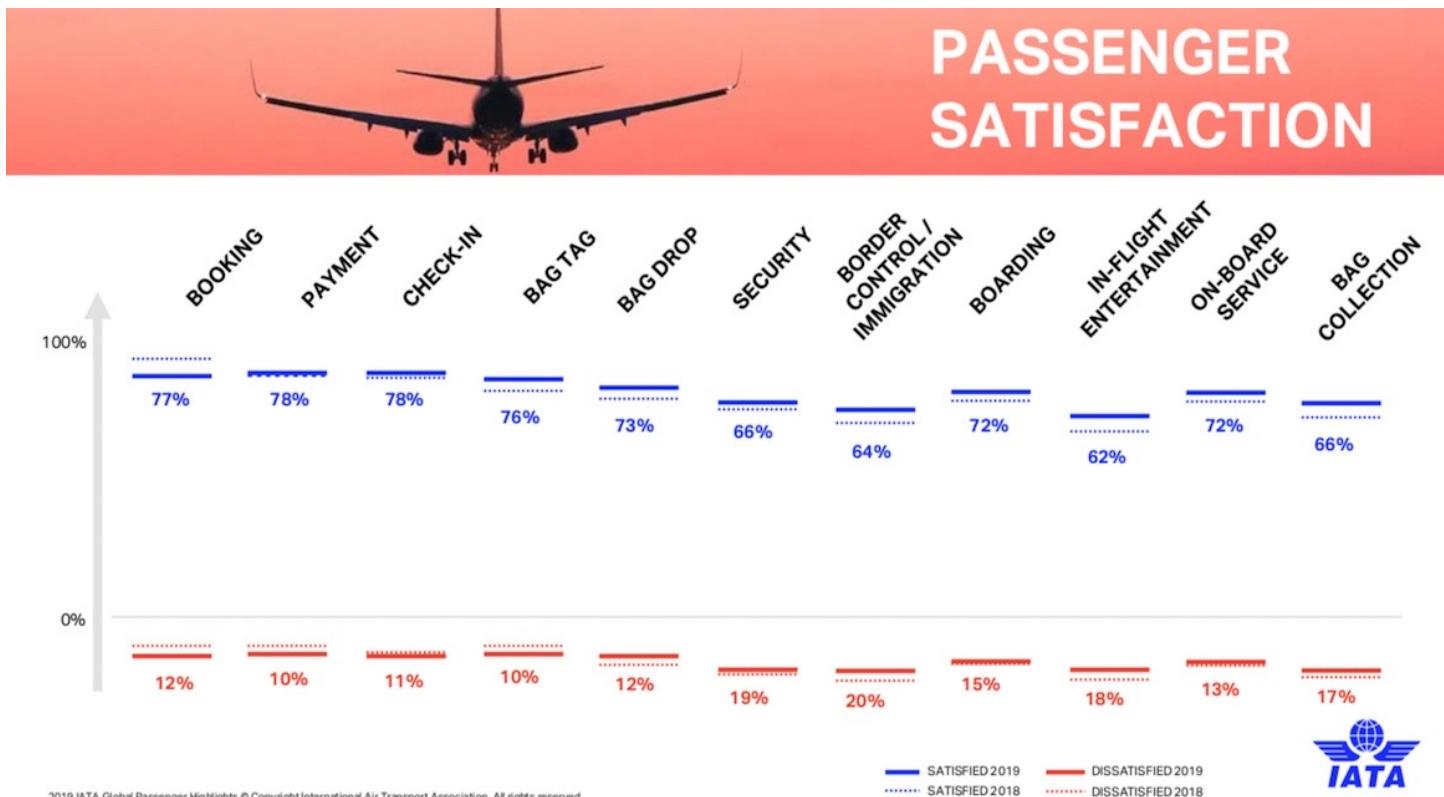
Bài toán 1:

Mục tiêu: Nhằm cải thiện, nâng cao chất lượng phục vụ hành khách. Giúp tìm ra những dịch vụ, những khâu còn hạn chế để đem đến sự hài lòng cho hành khách qua đó thu hút và giữ chân khách hàng sử dụng dịch vụ bay của hãng mình.



→ Để thực hiện được mục tiêu này cần tiến hành phân tích **dữ liệu đánh giá của hành khách** với các dịch vụ mà hãng bay cung cấp qua các chuyến bay mà hành khách đã bay.

Phương pháp thu thập dữ liệu



Thu thập các dữ liệu này như thế nào?





Phương pháp thu thập dữ liệu



Thu thập các dữ liệu này như thế nào?



Dữ liệu này chưa có, Cần thực hiện lấy khảo sát của hành khách thông qua các kênh khác nhau, thu thập và tổng hợp các dữ liệu này lại để phục vụ phân tích.

Airline Passenger Satisfaction quickpsurvey.com Your logo here

Please rate the following criteria based on your experience with our airline:

	Very Poor	Poor	Neutral	Good	Excellent
Luggage space	<input type="radio"/>				
Seating comfort	<input type="radio"/>				
Aircraft cleanliness	<input type="radio"/>				

Overall, how would you rate your interaction with our airline during your flight?

Do you participate in an airline rewards program?

START PLANETS

CUSTOMER FEEDBACK CARD

Name: _____ Date: _____

Email: _____ Phone: _____

Dear customer give us a small feedback About our Airlines Aviation services And suggest us new Services which is best for our Services

How is our Airlines Aviation Services
Good : Very Good : Poor :

How is Our Airlines Aviation Services performance
Good : Very Good : Poor :

How is our Staff performance
Good : Very Good : Poor :

How would you rate your overall experience with our service?
Good : Very Good : Poor :

how satisfied are you with our customer service representative?
Good : Very Good : Poor :

Comments and suggestions: _____

THANK YOU

Phương pháp thu thập dữ liệu

Bài toán 2:

Mục tiêu: Đánh giá hoạt động kinh doanh của chuỗi cửa hàng trong năm 2019

- Thực hiện phân tích tìm ra các thông tin có ích (insights) từ dữ liệu để cải thiện, tối ưu, nâng cao hoạt động kinh doanh tại các cửa hàng



→ Để thực hiện được mục tiêu này cần tiến hành phân tích **dữ liệu bán hàng** tại các cửa hàng trong chuỗi



Thu thập các dữ liệu này như thế nào?

Phương pháp thu thập dữ liệu



Thu thập các dữ liệu này như thế nào?



Lấy dữ liệu bán hàng trực tiếp từ CSDL của công ty các thông tin về Khách hàng và thông tin về doanh số, sản phẩm... đang được lưu trữ trong CSDL

Phương pháp thu thập dữ liệu

Bài toán 3:

Mục tiêu: Khách hàng đăng ký vay tiền, để tránh rủi ro tín dụng. Ngân hàng cần chấm điểm tín dụng của khách hàng này.



Dữ liệu cần thiết để chấm điểm tín dụng:

- Lịch sử vay tiền/ thanh toán/Khoản nợ
- Thông tin về khách hàng vay tiền (nghề nghiệp, độ tuổi, thu nhập, tình trạng hôn nhân,...)



Thu thập các dữ liệu
này như thế nào?

Phương pháp thu thập dữ liệu



Thu thập các dữ liệu này như thế nào?



Lịch sử vay tiền/thanh toán/Khoản nợ;
Lấy dữ liệu trực tiếp từ CSDL ngân hàng



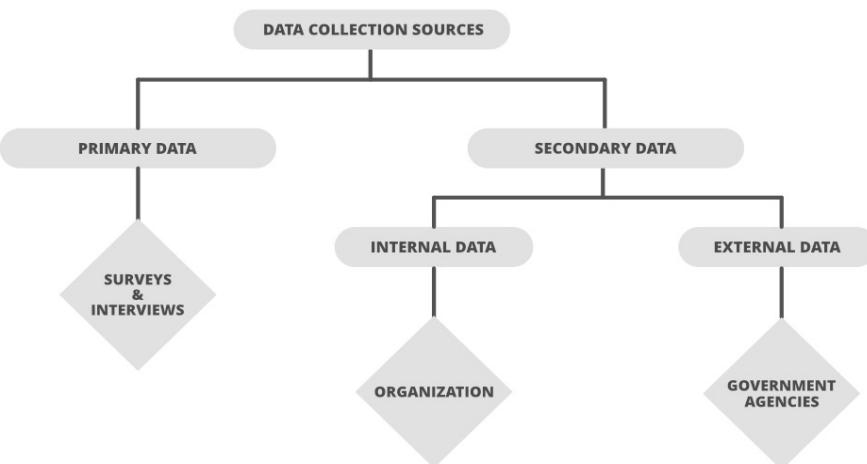
Thông tin về khách hàng vay tiền (nghề nghiệp, độ tuổi, thu nhập, tình trạng hôn nhân,...)
Mua dữ liệu từ bên thứ 3



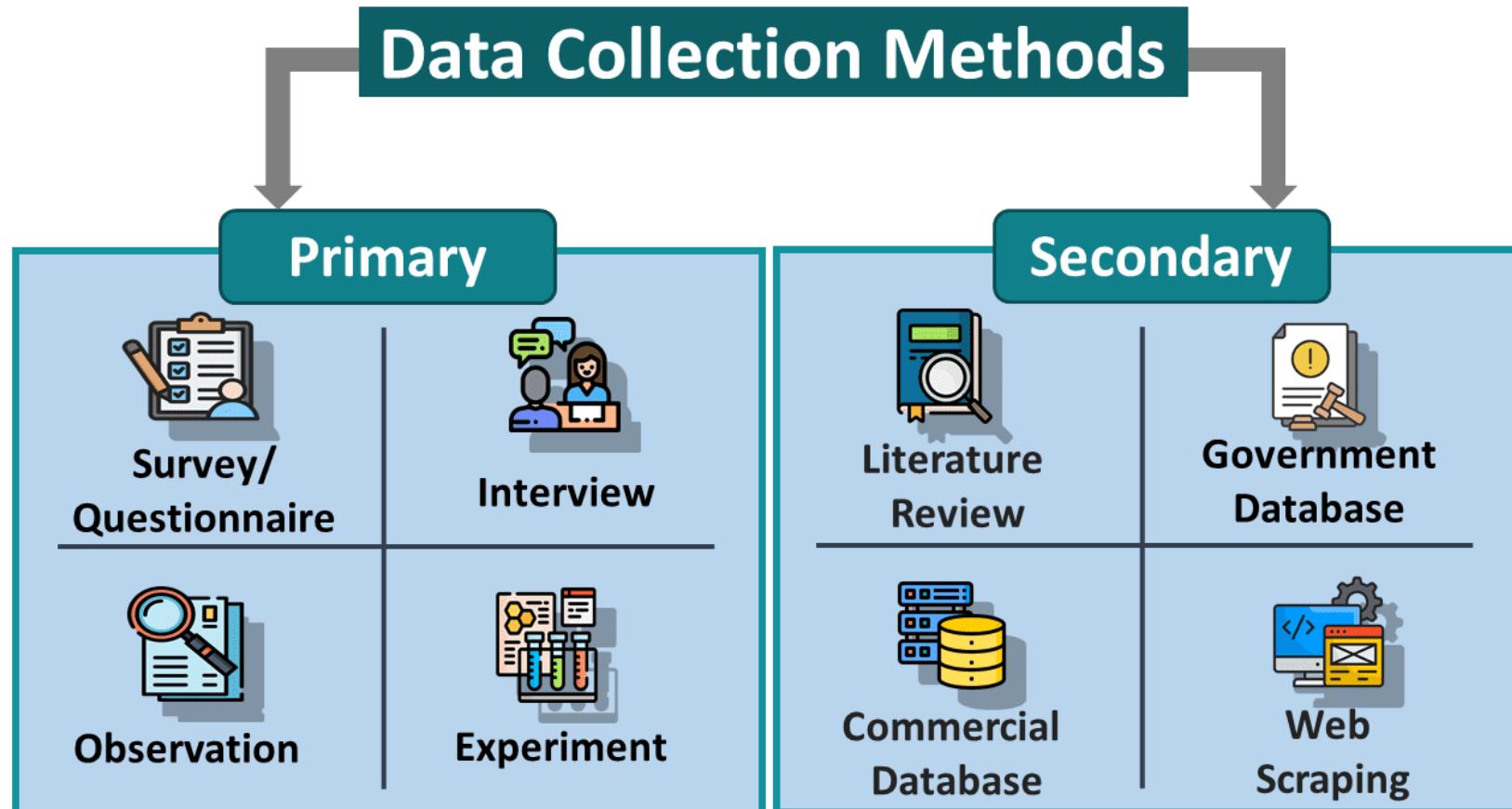
Phương pháp thu thập dữ liệu

Các nguồn thu thập dữ liệu bao gồm:

- **Nguồn dữ liệu chính (Primary Data):** Sử dụng các phương pháp thu thập trực tiếp thông qua phỏng vấn, khảo sát, bảng câu hỏi...
- **Nguồn dữ liệu thứ cấp (Secondary Data):** Là dữ liệu được thu thập, tổng hợp và sử dụng lại từ các nguồn dữ liệu đã có. Các dữ liệu này có thể đến từ:
 - **Nguồn dữ liệu bên trong (Internal Data):** Dữ liệu được thu thập trực tiếp từ bên trong công ty, tổ chức. Như hồ sơ bán hàng, khách hàng, giao dịch...từ trong CSDL của chính đơn vị.
 - **Nguồn dữ liệu bên ngoài (External Data):** Dữ liệu được thu thập hoặc mua từ bên thứ 3, bao gồm các dữ liệu mở (Open Data), Dữ liệu từ các cơ quan chính phủ, dữ liệu từ các công ty chuyên cung cấp có thể miễn phí hoặc có phí.



Phương pháp thu thập dữ liệu





Phương pháp thu thập dữ liệu

Dữ liệu được lưu trữ thế nào:

On-premises files	On-premises databases	Cloud Service	Cloud Platform	Other source
<ul style="list-style-type: none">• Excel files• CSV files• TXT• Json, XML	<ul style="list-style-type: none">• SQL server• MySQL• ...• Mongo DB	<ul style="list-style-type: none">• Google sheet• One drive	<ul style="list-style-type: none">• Google Bigquery• MS Azure	<ul style="list-style-type: none">• Web• Google analytics

2.2 Lấy mẫu (Sampling)

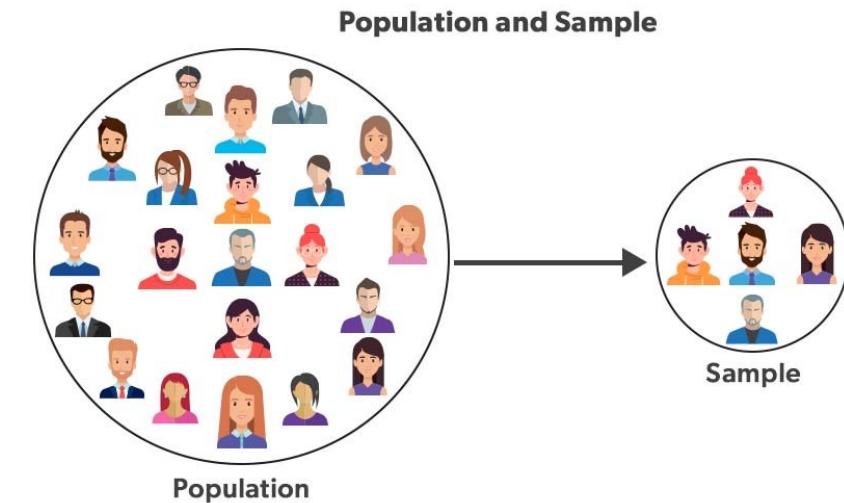


Lấy mẫu (Sampling)

Tổng thể (Population): Là toàn bộ tập hợp các đối tượng muốn nghiên cứu. Có thể là các cá nhân, sự kiện, hoặc kết quả trong một nhóm cụ thể.

Ví dụ: Muốn đánh giá mức độ hài lòng của khách hàng về một dịch vụ mà công ty cung cấp, thì Tổng thể là toàn bộ khách hàng đã sử dụng dịch vụ đó.

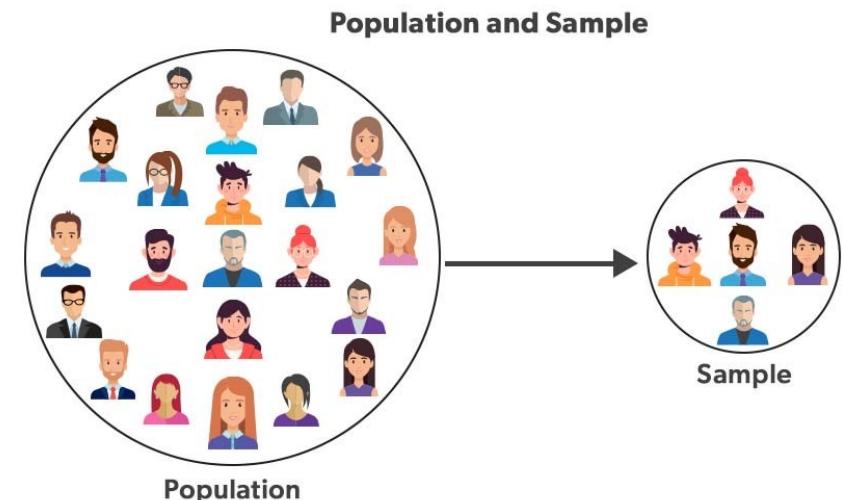
→ Population thường rất lớn và đôi khi không thể nghiên cứu toàn bộ do hạn chế về thời gian, chi phí và nguồn lực.



Lấy mẫu (Sampling)

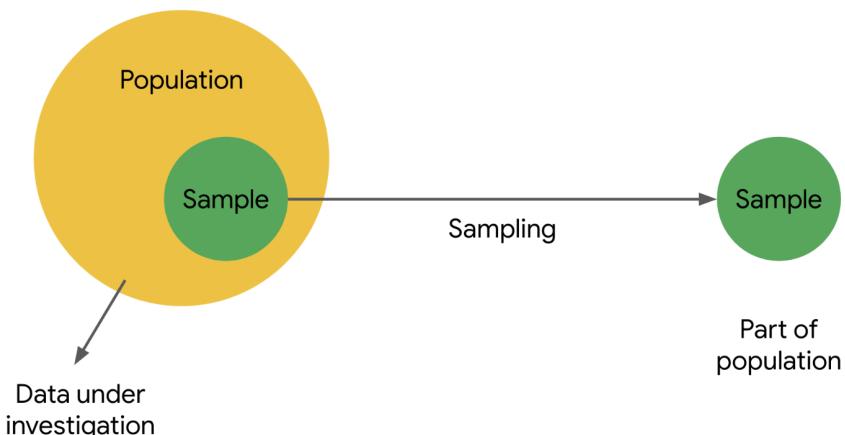
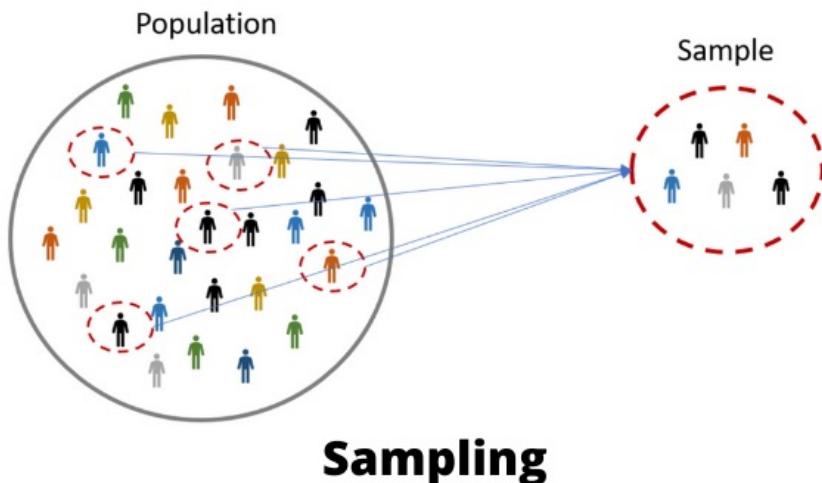
Mẫu (Sample): Là một phần nhỏ của tổng thể (Population) được chọn để nghiên cứu, với mục tiêu đại diện cho Population đó. Mẫu giúp phân tích, đánh giá các đặc điểm của Population mà không cần phải khảo sát toàn bộ.

Ví dụ: Muốn đánh giá mức độ hài lòng của khách hàng về một dịch vụ mà công ty cung cấp, chọn 1000 khách hàng đã sử dụng dịch vụ đó để nghiên cứu. Nhóm 1000 khách hàng này là Mẫu (Sample) trong Tổng thể (Population) toàn bộ KH sử dụng dịch vụ.



Lấy mẫu (Sampling)

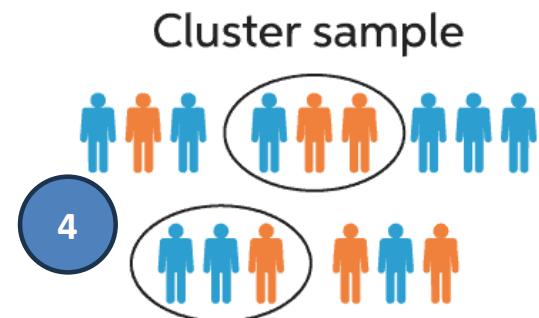
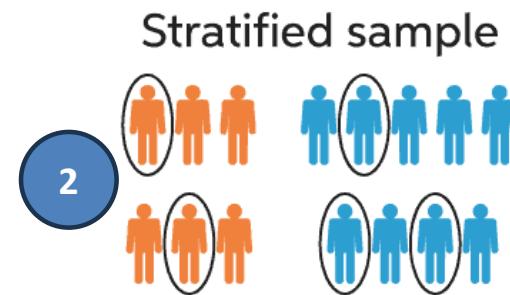
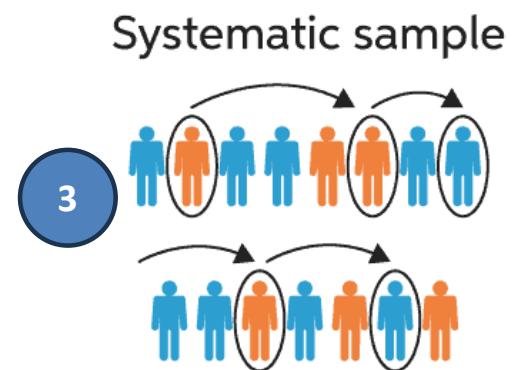
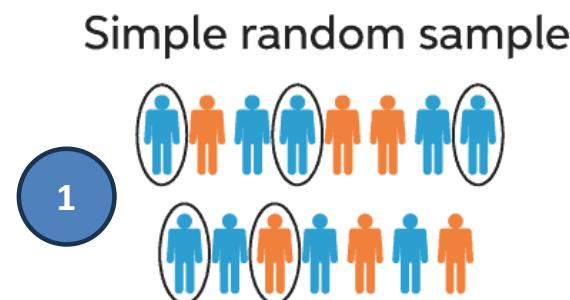
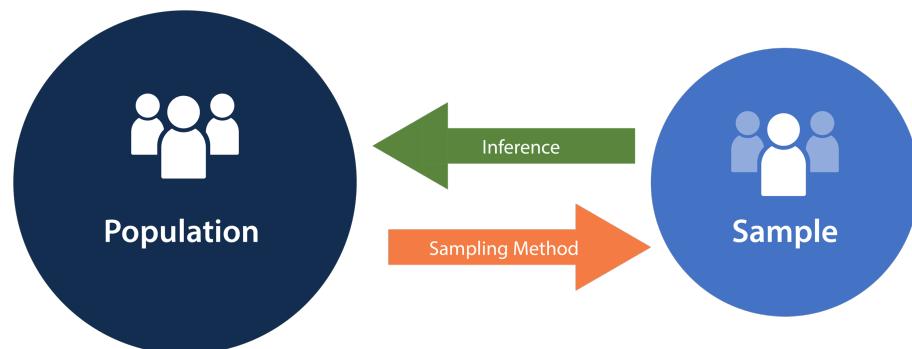
- Population là "toàn bộ" còn Sample là "một phần đại diện" của toàn bộ đó. Mẫu phải được chọn ngẫu nhiên và đủ lớn để đảm bảo tính đại diện cho Tổng thể (Population), giúp các kết quả nghiên cứu được suy rộng ra toàn bộ Tổng thể.
- Trong thống kê, việc chọn mẫu đúng cách và phân tích mẫu đúng phương pháp rất quan trọng để có thể đưa ra các kết luận chính xác về Population



Phương pháp lấy mẫu (Sampling)

- Một số phương pháp lấy mẫu:

 - 1. Lấy mẫu ngẫu nhiên đơn giản
 - 2. Lấy mẫu phân tầng
 - 3. Lấy mẫu hệ thống
 - 4. Lấy mẫu phân cụm





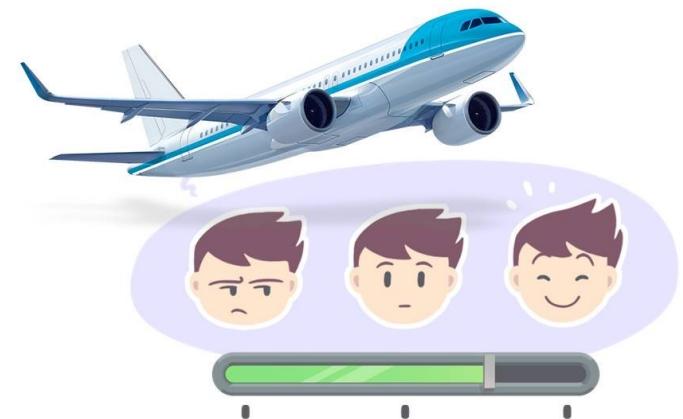
1.2 Tích hợp dữ liệu



Tích hợp dữ liệu (Data Integration)

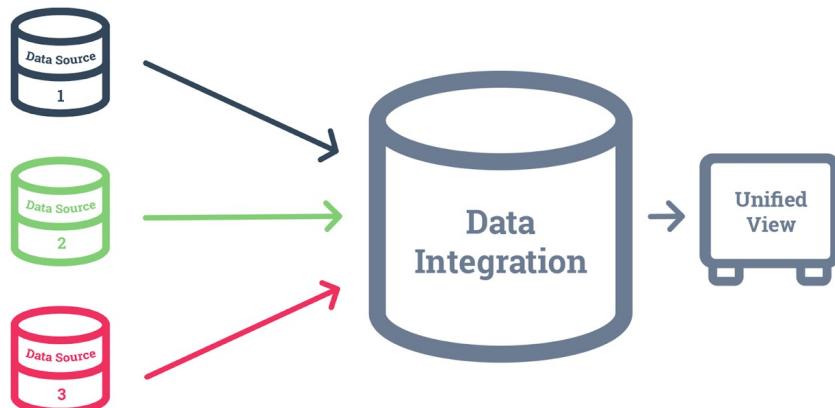
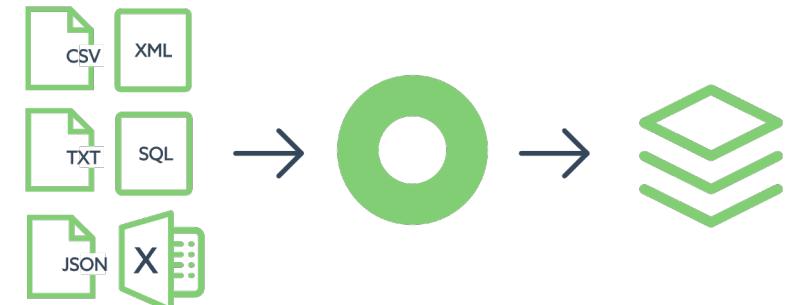
Vấn đề gặp phải: Để phân tích dữ liệu khảo sát, đánh giá các dịch vụ của hãng hàng không. Qua đó xác định được những dịch vụ nào còn hạn chế, những dịch vụ nào quan trọng ảnh hưởng đến sự hài lòng của khách hàng. Dự án cần 2 nguồn dữ liệu để thực hiện mục đích này:

- **Dataset 1:** Dữ liệu khảo sát hành khách cho các dịch vụ (tổng hợp trong file Excel) – Do Phòng Marketing quản lý.
- **Dataset 2:** Dữ liệu thông tin khách hàng đi trên các chuyến bay thực hiện khảo sát (lưu trữ trong các bảng của CSDL) – Trung tâm điều hành khai thác lưu trữ.



Tích hợp dữ liệu (Data Integration)

- Dữ liệu cần thiết bị phân mảnh ở nhiều định dạng, vị trí khác nhau dẫn đến các trở ngại và thách thức khi truy cập. Do đó cần phải thu thập, định dạng và tích hợp các tập dữ liệu cần thiết trước khi phân tích.
- Tích hợp dữ liệu liên quan đến các kỹ thuật, công cụ và phương thức về kiến trúc nhằm hợp nhất các nguồn, các định dạng dữ liệu khác nhau này cho hoạt động phân tích.

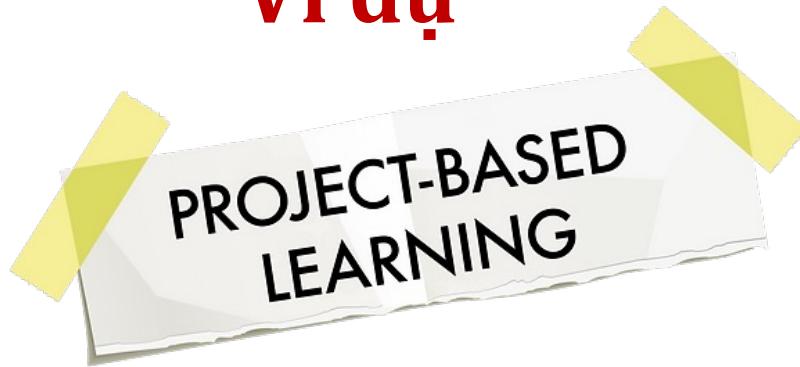


- Một số lợi ích của việc tích hợp dữ liệu:**
 - Cải thiện hiệu quả quản lý và sử dụng dữ liệu
 - Nâng cao chất lượng và tính toàn vẹn của dữ liệu
 - Khám phá các thông tin chuyên sâu, có ích từ dữ liệu nhanh và chính xác hơn.

Tích hợp dữ liệu (Data Integration)



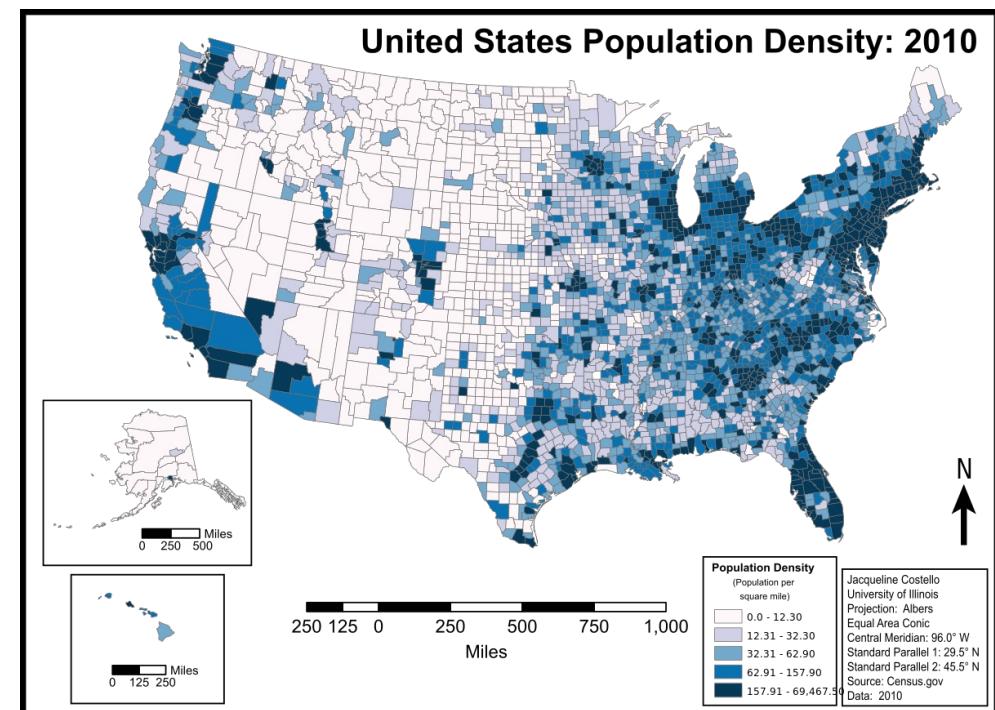
Ví dụ



Case study - Tích hợp dữ liệu

Yêu cầu: Xác định 5 bang của nước Mỹ có **mật độ dân số** cao nhất, thấp nhất năm 2010

Để xác định được mật độ dân số các bang của Mỹ năm 2010 cần thu thập được dữ liệu về dân số các bang năm 2010 và dữ liệu về diện tích của các bang.





Case study - Tích hợp dữ liệu

Dữ liệu dân số được thu thập từ Cục Điều tra Dân số Hoa Kỳ (U.S. Census Bureau): Lưu trữ trong file Excel **state_population.xlsx** từ năm 1990 đến năm 2013; Mỗi năm lưu trong một sheet bao gồm các thông tin:

- Số thứ tự
- **state/region:** Mã bang
- **ages:** nhóm tuổi: total (Tổng số của Bang)
- under 18 (Dân số dưới 18 tuổi)
- **year:** năm
- **population:** dân số

A	B	C	D	E
1		state/region	ages	year
2	0	AK	total	1998
3	1	AK	under18	1998
4	2	AL	total	1998
5	3	AL	under18	1998
6	4	AR	total	1998
7	5	AR	under18	1998
8	6	AZ	total	1998
9	7	AZ	under18	1998
10	8	CA	total	1998
11	9	CA	under18	1998
12	10	CO	total	1998
13	11	CO	under18	1998
14	12	CT	total	1998
15	13	CT	under18	1998
16	14	DC	total	1998



Case study - Tích hợp dữ liệu

Dữ liệu diện tích được thu thập từ Cục Quản lý Đất đai (Bureau of Land Management - BLM): Lưu trữ trong file CSV **state_areas.csv**, thông tin diện tích các bang của nước Mỹ:

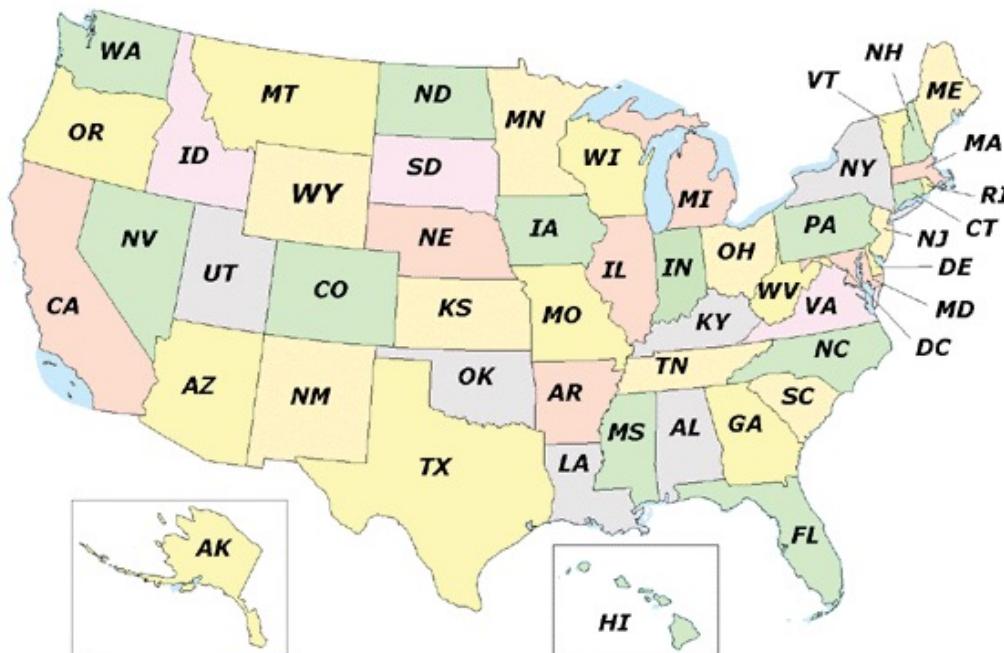
- **state**: Tên bang
- **area**: diện tích - dặm vuông sq.mi

The screenshot shows a spreadsheet application window titled 'state_areas'. The main area displays a table with two columns: 'state' and 'area (sq. mi)'. The table contains 17 rows of data, each representing a US state and its corresponding area in square miles. The data is as follows:

	state	area (sq. mi)
1	Alabama	52423
2	Alaska	656425
3	Arizona	114006
4	Arkansas	53182
5	California	163707
6	Colorado	104100
7	Connecticut	5544
8	Delaware	1954
9	Florida	65758
10	Georgia	59441
11	Hawaii	10932
12	Idaho	83574
13	Illinois	57918
14	Indiana	36420
15	Iowa	56276
16	Kansas	82282
17		

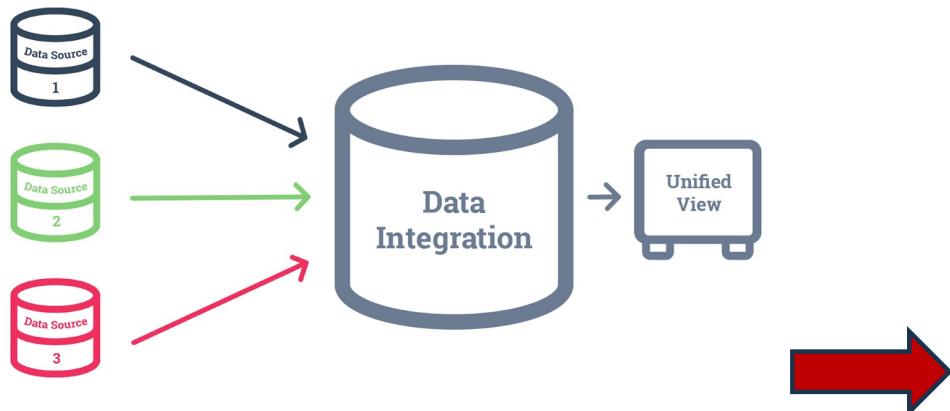
Case study - Tích hợp dữ liệu

Dữ liệu mã bang được thu thập trên Internet Lưu trữ trong file txt state_code.txt. Cho biết thông tin Tên và ký hiệu từng bang của Mỹ



state_code.txt	
Alabama,AL	
Alaska,AK	
Arizona,AZ	
Arkansas,AR	
California,CA	
Colorado,CO	
Connecticut,CT	
Delaware,DE	
District of Columbia,DC	
Florida,FL	
Georgia,GA	
Hawaii,HI	
Idaho,ID	
Illinois,IL	
Indiana,IN	
Iowa,IA	
Kansas,KS	
Kentucky,KY	
Louisiana,LA	
Maine,ME	
Montana,MT	
Nebraska,NE	
Nevada,NV	
New Hampshire,NH	
New Jersey,NJ	
New Mexico,NM	
New York,NY	
North Carolina,NC	

Case study - Tích hợp dữ liệu



data_population_2010.head(10) data_area.head(10) data_code.head(10)

	state/region	population
0	AK	713868
1	AL	4785570
2	AR	2922280
3	AZ	6408790
4	CA	37333601
5	CO	5048196
6	CT	3579210
7	DC	605125
8	DE	899711
9	FL	18846054

	state	area (sq. mi)		Name	Code
0	Alabama	52423	0	Alabama	AL
1	Alaska	656425	1	Alaska	AK
2	Arizona	114006	2	Arizona	AZ
3	Arkansas	53182	3	Arkansas	AR
4	California	163707	4	California	CA
5	Colorado	104100	5	Colorado	CO
6	Connecticut	5544	6	Connecticut	CT
7	Delaware	1954	7	Delaware	DE
8	Florida	65758	8	District of Columbia	DC
9	Georgia	59441	9	Florida	FL

	Name	Code	area (sq. mi)	population	population_density
0	District of Columbia	DC	68	605125	8898.897059
1	Puerto Rico	PR	3515	3721208	1058.665149
2	New Jersey	NJ	8722	8802707	1009.253268
3	Rhode Island	RI	1545	1052669	681.339159
4	Connecticut	CT	5544	3579210	645.600649
5	Massachusetts	MA	10555	6563263	621.815538
6	Maryland	MD	12407	5787193	466.445797
7	Delaware	DE	1954	899711	460.445752
8	New York	NY	54475	19398228	356.094135
9	Florida	FL	65758	18846054	286.597129



1.3 Bài tập



Thực hành

Tập dữ liệu Data_Movies.xlsx lưu trữ danh sách các bộ phim đã công chiếu. Mỗi một bộ film bao gồm 25 thuộc tính khác nhau. Dữ liệu được lưu thành 3 sheet:

- **Sheet 1900s:** Lưu trữ danh sách các bộ phim công chiếu trước năm 2000
- **Sheet 2000s:** Lưu trữ danh sách những bộ phim công chiếu từ năm 2000 đến trước năm 2010.
- **Sheet 2010s:** Lưu trữ danh sách những bộ phim công chiếu từ năm 2010 đến nay

	A	B	C	D	E	F	G	H	I	J	
1	Title	Year	Genre	Language	Country	Content Rating	Duration	Aspect Ratio	Budget	Gross Earnings	Director
2	Intolerance: Love's Struggle Throughout the Ages	1916	Drama History War	USA	Not Rated	123	1.33	385907			D.W.
3	Over the Hill to the Poorhouse	1920	Crime Drama	USA		110	1.33	100000			Harry
4	The Big Parade	1925	Drama Romance War	USA	Not Rated	151	1.33	245000			King
5	Metropolis	1927	Drama Sci-Fi	German	Germany	Not Rated	145	1.33	600000		Fritz
6	Pandora's Box	1929	Crime Drama Romance	German	Germany	Not Rated	110	1.33			Geor
7	The Broadway Melody	1929	Musical Romance	English	USA	Passed	100	1.37	379000		2808000 Harry
8	Hell's Angels	1930	Drama War	English	USA	Passed	96	1.2	3950000		How
9	A Farewell to Arms	1932	Drama Romance War	English	USA	Unrated	79	1.37	800000		Frank
10	42nd Street	1933	Comedy Musical Romance	English	USA	Unrated	89	1.37	439000		Lloy
11	She Done Him Wrong	1933	Comedy Drama History Musical Romance	English	USA	Approved	66	1.37	200000		Low
12	It Happened One Night	1934	Comedy Romance	English	USA	Unrated	65	1.37	325000		Frank
13	Top Hat	1935	Comedy Musical Romance	English	USA	Approved	81	1.37	609000		Mark
14	Modern Times	1936	Comedy Drama Family	English	USA	G	87	1.37	1500000		Char
15	The Charge of the Light Brigade	1936	Action Adventure Romance War	English	USA	Approved	100	1.37	1200000		Mich
16	Snow White and the Seven Dwarfs	1937	Animation Family Fantasy Musical	English	USA	Approved	83	1.37	2000000		Willia
17	The Prisoner of Zenda	1937	Adventure Drama Romance	English	USA	Approved	101	1.37			John
18	Alexander's Ragtime Band	1938	Drama Musical Romance	English	USA	Approved	106	1.37	2000000		Henr
19	You Can't Take It with You	1938	Comedy Drama Romance	English	USA	Approved	126	1.37	1644736		Frank
20	Gone with the Wind	1939	Drama History Romance War	English	USA	G	226	1.37	3977000		Victo
21	Mr. Smith Goes to Washington	1939	Comedy Drama	English	USA	Not Rated	120	1.37	1500000		Frank
22	The Wizard of Oz	1939	Adventure Family Fantasy Musical	English	USA	Passed	102	1.37	2800000		Victo
23	Boom Town	1940	Adventure Drama Romance Western	English	USA	Passed	119	1.37	1614000		Jack
24	Fantasia	1940	Animation Family Fantasy Music	English	USA	G	120	1.37	2280000		Jame
25	Pinocchio	1940	Animation Family Fantasy Musical	English	USA	Approved	88	1.37	2600000		84300000 Norm



Thực hành



Yêu cầu:

1. Đọc dữ liệu từng sheet trong file excel và tích hợp thành một bảng dữ liệu duy nhất chứa tất cả các bộ phim.
2. Tạo bảng dữ liệu mới chỉ sử dụng 7 thuộc tính quan trọng sau:
 - Title: Tên phim
 - Year: Năm phát hành
 - Genres: Thể loại phim
 - Country: Quốc gia
 - Director: Đạo diễn phim
 - User Votes: Số lượng người xem đánh giá
 - IMDB Score: Điểm đánh giá trung bình

Sắp xếp các bộ phim theo thứ tự IMDB Score giảm dần, nếu bộ phim có IMDB Score bằng nhau thì sắp xếp theo thuộc tính User Votes giảm dần.

3. Lưu dữ liệu ra file định dạng CSV: Data_Movies_OK.csv

