# KHAI THÁC DỮ LIỆU LỚN ĐỐI VỚI GIÁO DỤC ĐẠI HỌC TRONG BỐI CẢNH CUỘC CÁCH MẠNG CÔNG NGHIỆP 4.0

*Lê Thị Cẩm Bình**

***Tóm tắt:*** *Tốc độ phát triển đột phá của công nghệ số trong những năm gần đây như điện thoại thông minh, điện toán đám mây, Internet vạn vật, mạng xã hội, các dịch vụ online,... đã phát sinh một lượng dữ liệu khổng lồ và đến từ nhiều nguồn khác nhau, chủ yếu từ các phương tiện truyền thông xã hội như Twitters, Youtube, Facebook; các giao dịch kinh doanh như Amazon, eBay, giao dịch qua mạng hoặc giao dịch từ các thiết bị di động; các máy móc thu nhận dữ liệu như máy gia tốc hạt lớn của CERN, các thiết bị cảm biến,... Các dữ liệu này thường không có cấu trúc như các tài liệu, ảnh, video, audio, email, dữ liệu trên các trang web,... Theo thống kê, chúng chiếm khoảng trên 80% các loại dữ liệu hiện nay và không ngừng lớn lên. Lượng dữ liệu khổng lồ đó là nguồn gốc ra đời của khái niệm dữ liệu lớn(Big data). Bài viết sau đây đề cập đến tác động, cơ hội và thách thức của việc khai thác dữ liệu lớn đối với đổi mới giáo dục Đại học Việt Nam trong bối cảnh diễn ra cuộc cách mạng công nghiệp 4.0 trong thời đại hiện nay.*

## 1. Một số khái niệm

### *1.1. Cuộc cách mạng công nghiệp*

Trong Hội thảo Khoa học dữ liệu và Cách mạng công nghiệp lần thứ tư diễn ra ngày 15 tháng 5 năm 2017, giáo sư Hồ Tú Bảo - Viện Khoa học và Công nghệ tiên tiến Nhật Bản nhận định: để xác định một cuộc cách mạng công nghiệp cần có hai đặc trưng: Thứ nhất, cần có sự đột phá của khoa học và công nghệ. Thứ hai, tạo ra sự thay đổi về bản chất của sản xuất.

Nhìn lại lịch sử, khoa học công nghệ thế giới đã trải qua 3 cuộc cách mạng công nghiệp lớn, đó là cách mạng công nghiệp 1.0 về sản xuất cơ khí dựa trên động cơ hơi nước; cách mạng công nghiệp 2.0 sản xuất máy móc dựa trên năng lượng điện; cách mạng công nghiệp 3.0 sản xuất tự động với công nghệ số hóa trong lĩnh vực điện tử và công nghệ thông tin. Cuộc cách mạng công nghiệp 4.0 gần đây sản xuất thông minh dựa trên những tiến bộ đột phá kết hợp các công nghệ của cuộc cách mạng 3.0. Theo Klaus Schwab, tốc độ đột phá của cách mạng công nghiệp 4.0 hiện "không có tiền lệ lịch sử". So với các cuộc cách mạng công nghiệp trước đây, cuộc cách mạng công nghiệp hiện nay đang tiến triển nhanh hơn rất nhiều.

### *1.2. Dữ liệu lớn*

Chúng ta xem xét một số định nghĩa về dữ liệu lớn:

* Thạc sĩ, Khoa Lý luận Chính trị và Khoa học cơ bản - Trường Đại học Văn hóa Hà Nội

- Theo Wikipedia: Dữ liệu lớn (Big data) là một thuật ngữ chỉ bộ dữ liệu lớn hoặc phức tạp mà các phương pháp truyền thống không đủ các ứng dụng để xử lý dữ liệu này.

- Theo Gartner: Dữ liệu lớn là những nguồn thông tin có đặc điểm chung khối lượng lớn, tốc độ nhanh và dữ liệu định dạng dưới nhiều hình thức khác nhau, do đó muốn khai thác được đòi hỏi phải có hình thức xử lý mới để đưa ra quyết định, khám phá và tối ưu hóa quy trình.

- Theo IBM: Dữ liệu lớn là sự thu thập, quản lý và phân tích dữ liệu, những việc đó đã vượt xa dữ liệu cấu trúc tiêu biểu, nó có thể được truy vấn với hệ thống quản trị dữ liệu quan hệ - thường với những tệp phi cấu trúc, video kỹ thuật số, hình ảnh, dữ liệu cảm biến, tệp lưu nhật ký, bất cứ dữ liệu nào không có trong hồ sơ với các phạm vi tìm kiếm khác.

*Đặc trưng và thuộc tính của dữ liệu lớn*

Năm 2001, dữ liệu lớn được Doug Laney mô tả bằng ba tính chất (3Vs) đó là sự gia tăng về dung lượng (volumn), vận tốc (velocity) và chủng loại (variety). Các tính chất này sau đó đã được các công ty và tổ chức sử dụng để định nghĩa về dữ liệu lớn. Sau đó, Gartner bổ sung thêm rằng dữ liệu lớn ngoài ba tính chất cơ bản nêu trên thì còn phải "cần đến các dạng xử lí mới để giúp đỡ việc đưa ra quyết định, khám phá sâu vào sự vật/sự việc và tối ưu hóa các quy trình làm việc". Hiện nay, dữ liệu lớn được mô tả bởi năm đặc trưng (5Vs):

• Volume (dung lượng): số lượng dữ liệu được tạo ra và lưu trữ. Kích thước của dữ liệu xác định liệu nó có thể thực sự được coi là dữ liệu lớn hay không. Đây là đặc điểm tiêu biểu nhất của dữ liệu lớn. Dữ liệu truyền thống có thể lưu trữ trên các thiết bị đĩa mềm, đĩa cứng. Nhưng với dữ liệu lớn, công nghệ "đám mây" mới có thể đáp ứng khả năng lưu trữ được dữ liệu lớn.

| | | | Dữ liệu lớn | | |
|---|---|---|---|---|---|
| Bytes | $10^6$ | $10^8$ | $10^{10}$ | $10^{12}$ | $10^{>12}$ |
| Kích thước | Medium | Large | Huge | Monster | Very large |

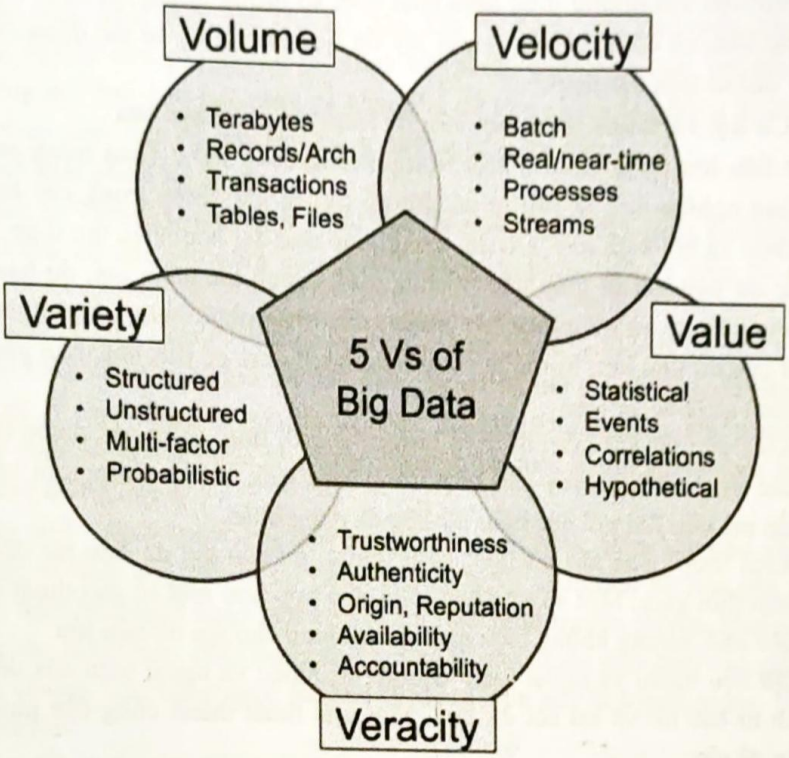
*Hình 1. Bảng phân loại dữ liệu dựa trên dung lượng* [1]

• Variety (tính đa dạng): các dạng và kiểu của dữ liệu. Dữ liệu được thu thập từ nhiều nguồn khác nhau và các kiểu dữ liệu cũng có rất nhiều cấu trúc khác nhau. Người ta ước tính có tới hơn 80% dữ liệu được sinh ra là phi cấu trúc như tài liệu văn bản, email, dữ liệu từ website, ảnh, âm thanh, video,...

• Velocity (vận tốc): tốc độ sản sinh, lưu thông, lưu trữ, phân tích, xử lý dữ liệu. Bao gồm tốc độ gia tăng dữ liệu rất nhanh và xử lý dữ liệu phát sinh trong thời gian thực.

• Veracity (tính xác thực): chất lượng và độ tin cậy của dữ liệu thu được có thể khác nhau, làm ảnh hưởng đến sự phân tích chính xác. Việc loại bỏ dữ liệu thiếu chính xác và nhiễu là nhiệm vụ quan trọng của phân tích và xử lý dữ liệu lớn.

• Value (giá trị): xác định giá trị của thông tin mang lại. Đây là một đặc tính quyết định trong vấn đề có khai thác và xử lý dữ liệu lớn hay không.




*Hình 2. Năm đặc trưng của dữ liệu lớn* [2]

*Phân biệt dữ liệu lớn với các dữ liệu khác*

Dựa vào đặc trưng tính chất nêu trên, có thể thấy rằng về cơ bản dữ liệu lớn khác với dữ liệu truyền thống ở bốn điểm sau đây [3]:

• Dữ liệu đa dạng hơn: Dữ liệu lớn có thể có cấu trúc, bán cấu trúc hoặc phi cấu trúc, nhưng thông thường là phi cấu trúc. Khi khai thác dữ liệu truyền thống (dữ liệu có cấu trúc), chúng ta thường quan tâm đến kiểu và định dạng của chúng. Tuy nhiên, đối với dữ liệu lớn, ta không cần quan tâm đến các đặc điểm đó mà tập trung ở giá trị mà dữ liệu mang lại có đáp ứng được cho công việc hay không.

• Lưu trữ dữ liệu lớn hơn: Lưu trữ dữ liệu truyền thống quan tâm đến khía cạnh cách thức, dung lượng kho lưu trữ và chi phí. Giải pháp lưu trữ thường đầu tư cho CPU, RAM, đĩa cứng,... Giải pháp này ngày càng kém hiệu quả, đòi hỏi đầu tư mới liên tục nếu áp dụng để lưu trữ dữ liệu lớn do có kích thước lớn và đặc tính phức tạp. Việc lưu trữ dữ liệu lớn ngày nay đã được đơn giản hóa nhờ những công nghệ như lưu trữ đám mây, phân phối lưu trữ dữ liệu phân tán, có thể kết hợp chúng lại với nhau một cách chính xác và xử lý nhanh trong thời gian thực.

• Truy vấn dữ liệu nhanh hơn: Dữ liệu lớn được cập nhật liên tục, trong khi cơ sở dữ liệu truyền thống không được cập nhật thường xuyên. Do đó có thể gây ra tình trạng lỗi cấu trúc truy vấn dẫn đến không tìm kiếm được thông tin đáp ứng theo yêu cầu.

• Độ chính xác cao hơn: Dữ liệu lớn khi đưa vào sử dụng thường được kiểm định lại dữ liệu với những điều kiện chặt chẽ, số lượng thông tin được kiểm tra thông thường rất lớn, và đảm bảo về nguồn lấy dữ liệu không có sự tác động của con người vào thay đổi số liệu thu thập.

**2. Cơ hội và thách thức của vấn đề khai thác dữ liệu lớn**

Dữ liệu lớn được xem là một trong những công nghệ quan trọng của cuộc cách mạng công nghiệp 4.0, là yếu tố cốt lõi để tìm ra giải pháp trong các hoạt động của mọi tổ chức xã hội hiện nay. Do đó, trong giáo dục đại học, việc thu thập, phân tích và khai thác dữ liệu lớn sẽ giúp các tổ chức giáo dụccó thể đánh giá, dự báo, đưa ra các giải pháp phù hợp và hiệu quả.Tuy nhiên, cần khắc phục những khó khăn do đặc thù tính chất của dữ liệu lớn, trong đó có những thách thức cơ bản nhất bao gồm:

*Thách thức về dữ liệu*

- Kiểu của tập dữ liệu: dữ liệu trong thế giới thực chứa các thuộc tính đa dạng. Do đó các kỹ thuật hiện nay cần hoạt động hiệu quả, không những đối với dữ liệu số đơn thuần mà còn đối với các kiểu dữ liệu đa dạng khác.

- Kích thước của tập dữ liệu: kích thước lớn của tập dữ liệu tác động đến hiệu quả về mặt thời gian. Một số kỹ thuật tỏ ra phù hợp hơn một số giải thuật khác đối với tập dữ liệu nhỏ, nhưng không hiệu quả khi áp dụng cho tập dữ liệu lớn.

- Dữ liệu nhiễu và ngoại biên: các giá trị nhiễu và ngoại biên của dữ liệu có thể phát sinh từ các lỗi và sai sót dữ liệu. Một giải thuật thành công cần phải giải quyết được vấn đề này.

- Số chiều dữ liệu lớn: Đặc trưng của dữ liệu lớn là phức tạp, có thể có từ vài trăm cho đến vài trăm ngàn chiều thuộc tính khác nhau hoặc nhiều hơn, dẫn đến việc khó khăn trong phân tích dữ liệu. Nếu thực hiện lưu trữ và tính toán trực tiếp trên các tập dữ liệu này thì sẽ tốn kém trong việc lưu trữ và tốc độ tính toán. Vấn đề này được Richard E. Bellman gọi là "The curse of dimensionality"- sự thiệt hại của số chiều dữ liệu lớn, đề cập đến sự phức tạp của các bài toán xử lý dữ liệu có số chiều thuộc tính lớn

*Thách thức về công nghệ*

Phát triển công cụ quản trị dữ liệu lớn, nghiên cứu về các kỹ thuật hiển thị dữ liệu lớn, về mối quan hệ phức tạp trong chúng, là những thách thức không nhỏ, nhưng thách thức chính của dữ liệu lớn là các phương pháp phân tích dữ liệu, trong đó chủ yếu là các phương pháp của hai lĩnh vực học máy (machine learning) và khai phá dữ liệu (data mining).

Học máy là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc phát triển các phương pháp kỹ thuật, cho phép máy móc "học" tự động từ dữ liệu để hỗ trợ con người giải quyết vấn đề nhanh chóng với một lượng thông tin khổng lồ phát sinh hàng

ngày. Học máy làm cho máy tính có một số khả năng học tập của con người, chủ yếu là học để khám phá, giúp phân tích các tập dữ liệu để phát hiện ra các quy luật, các mẫu dạng, hay các mô hình... Học máy thường kết hợp với các lĩnh vực liên quan trong toán học như thống kê và tối ưu,đã được chứng minh là hiệu quả trong vấn đề phân tích các dữ liệu phức tạp.

Khai phá dữ liệu là một lĩnh vực tập trung vào việc đưa các phương pháp học máy vào phân tích, khai thác các tập dữ liệu lớn có trong các lĩnh vực khác nhau. Những hướng nghiên cứu gần đây về mô hình làm thưa, giảm số chiều, mô hình đồ thị xác suất... trong hai lĩnh vực học máy và khai phá dữ liệu chính là những hướng đi tới các phương pháp phân tích dữ liệu lớn trong những năm tới đây.

**3. Một số giải pháp kỹ thuật**

Theo giáo sư Phùng Quốc Định - Đại học Deakin, Australia, chìa khóa khoa học và công nghệ của dữ liệu lớn gồm ba vấn đề: Quản trị, phân tích và hiển thị dữ liệu. Trong đó:

1- *Quản trị dữ liệu - data management* (bao gồm lưu trữ, bảo trì và truy nhập các nguồn dữ liệu lớn): công nghệ lưu trữ yêu cầu giải quyết bài toán lượng dữ liệu khổng lồ và tốc độ xử lý cao bằng cách phân mảnh dữ liệu và phân tán trên nhiều server lưu trữ. Khi truy xuất dữ liệu thì cho phép truy xuất đồng thời nhiều server lưu trữ cùng một lúc để tăng thông lượng.

2- *Phân tích - data modeling and analytics* (cách hiểu được dữ liệu và tìm ra thông tin hoặc tri thức quý báu từ dữ liệu). Để xử lý và phân tích dữ liệu lớn cần mô hình phương thức tính toán khác biệt so với các mô hình truyền thống. Phương pháp xử lý dữ liệu lớn là kết hợp, phối hợp năng lực xử lý của nhiều máy tính vào giải quyết một bài toán chung. Các công nghệ thường sử dụng hiện nay là MapReduce, Hadoop, Spark,TensorFlow.

3- *Trao đổi, hiển thị, kết quả phân tích dữ liệu* - visualization decisions and values (nhằm tạo ra sản phẩm hay giá trị). Hiển thị trực quan lượng dữ liệu khổng lồ và các tri thức khai thác được từ dữ liệu là đòi hỏi cần thiết khi làm việc với dữ liệu lớn. Việc hiển thị dữ liệu dưới dạng trực quan giúp người khai thác có cái nhìn toàn cảnh về dữ liệu và tri thức mang lại từ dữ liệu. Các công cụ cho phép hiển thị và tương tác trực quan với dữ liệu lớn hiện nay phổ biến là các công cụ như Tableau, Pentahoo, SAS, vv...

**4. Tác động của việc khai thác dữ liệu lớn đối với giáo dục đại học Việt Nam hiện nay**

Giáo dục đại học nhằm mục đích đào tạo một lực lượng lao động có trình độ chuyên môn cao, sáng tạo, có khả năng tự học tập nhằm đáp ứng nhu cầu của xã hội. Theo Bộ trưởng Bộ giáo dục và Đào tạo Phùng Xuân Nhạ: "Giáo dục đại học đóng vai trò then chốt trong việc cung cấp nguồn nhân lực bậc cao và đóng góp trực tiếp vào sự phát triển kinh tế - xã hội của đất nước". Cũng theo Bộ trưởng, "trong những năm qua, hệ thống giáo dục đại học của Việt Nam đã góp phần quan trọng vào sự nghiệp đổi mới và hội nhập quốc tế của đất nước thông qua việc mở rộng về quy mô, đa dạng hóa

các loại hình và ngành nghề đào tạo, nâng cao chất lượng và bước đầu hội nhập quốc tế".

Tuy nhiên, trong giai đoạn cuộc cách mạng công nghiệp 4.0 với những thay đổi nhanh chóng đã đặt ra những thách thức mới đối với giáo dục đại học. Với sự phát triển của khoa học công nghệ như hiện nay thì việc thu thập dữ liệu lớn không còn khó khăn nữa. Vấn đề là các nhà quản lý giáo dục đại học cần làm gì với dữ liệu đã được phân tích dựa trên các công nghệ nêu trên để nó trở nên có ích.

Dữ liệu lớn được ứng dụng trong quản lý giáo dục sẽ làm thay đổi sâu sắc bản chất của giáo dục đại học, giúp tăng mức độ minh bạch và dân chủ trong quản lý giáo dục, thúc đẩy tăng trưởng, cung cấp dịch vụ phù hợp với giáo dục đại học hiện nay,... Bên cạnh đó, dữ liệu lớn là nguồn khai thác để có thể dùng để sản xuất số liệu thống kê chính xác hơn và kịp thời hơn so với các nguồn số liệu thống kê truyền thống, có thể giúp các trường đại học dự đoán được tỉ lệ thất nghiệp, xu hướng nghề nghiệp của tương lai để đầu tư hoặc cắt giảm cho những hạng mục đó,... Ngoài ra, cuộc cách mạng công nghiệp 4.0 sẽ tạo ra những thay đổi sâu sắc về các ngành lao động và đặt ra những yêu cầu mới về năng lực và kỹ năng của người học. Việc phân tích hiệu quả dữ liệu lớn là căn cứ rất quan trọng để lựa chọn những mô hình đào tạo tiên tiến, phương thức dạy và học, nhằm làm tăng năng lực cạnh tranh trong khu vực và trên thế giới, đào tạo một lực lượng lao động chất lượng cao, có thể cạnh tranh về cung cấp nguồn nhân lực bậc cao trong khu vực và trên thế giới./.

## TÀI LIỆU THAM KHẢO


1. R. Hathaway and J. Bezdek (2006), "*Extending fuzzy and probabilistic clustering to very large data sets*," Comput. Stat. Data Anal, vol. 51, no. 1, pp. 215-234.

2. https://www.researchgate.net/figure/5-Vs-of-Big-Data-and-security-related-properties-of-Veracity-Variety-and-Volume_fig1_273945634

3. "Tổng quan về dữ liệu lớn", http://vienthongke.vn/attachments/article/2249/2.%20Tong%20quan%20ve%20DL%20lon.pdf

4. "Dữ liệu lớn trong Tài chính", https://vi.routestofinance.com/big-data-in-finance

5. Hồ Tú Bảo, *Khoa học phân tích dữ liệu lớn và Học máy thống kê*

6. Hong He, Yonghong (2017), *Automatic pattern recognition of ECG signals using entropy-based adaptive dimensionality reduction and clustering*, Elsevier, pp. 238-252(DOI: 10.1016/j.asoc).

7. Indrajit Saha, Jnanendra Prasad Sarkar, Ujjwal Maulik (2015), *Ensemble based rough fuzzy clustering for categorical data*, Elsevier, pp. 114-127 (DOI:10.1016/j.knosys).