**TRƯỜNG ĐẠI HỌC TÀI NGUYÊN VÀ MÔI TRƯỜNG TP.HCM**

**KHOA: HỆ THỐNG THÔNG TIN VÀ VIỄN THÁM**

****

**Dịch: BIG DATA FOR REMOTE SENSING:**

**CHALLENGES AND OPPOTUNITIES**

Giảng viên hướng dẫn: **Phạm Trọng Huynh**

Sinh viên thực hiện: **Nhóm 5**

Lớp**: 09\_CNTT2**

**TP. Hồ Chí Minh, tháng năm 2022**

**MỤC LỤC**

[I. LỜI MỞ ĐẦU 3](#_Toc121380265)

[II. TÌM HIỂU DỮ LIỆU LỚN TRONG VIỄN THÁM 6](#_Toc121380266)

[A. KHÍA CẠNH ĐẦU TIÊN: SỞ HỮU DỮ LIỆU 6](#_Toc121380267)

[B. KHÍA CẠNH THỨ HAI: PHƯƠNG PHÁP DỮ LIỆU LỚN 7](#_Toc121380268)

[C. KHÍA CẠNH THỨ BA: ỨNG DỤNG DỮ LIỆU LỚN 7](#_Toc121380269)

[III. DỮ LIỆU LỚN, THÁCH THỨC LỚN 8](#_Toc121380270)

[A. NHỮNG THÁCH THỨC THƯỜNG GẶP 8](#_Toc121380271)

[1. Điện toán dữ liệu lớn: 8](#_Toc121380272)

[2. Kết hợp dữ liệu lớn: 9](#_Toc121380273)

[3. Phương pháp dữ liệu lớn: 9](#_Toc121380274)

[B. NHỮNG THÁCH THỨC RIÊNG BIỆT 10](#_Toc121380275)

[1. Xác minh dữ liệu chính xác: 10](#_Toc121380276)

[2. Các thách thức trong quyền sở hữu dữ liệu: 11](#_Toc121380277)

[3. Phát triển dữ liệu: 12](#_Toc121380278)

[4. Biểu diễn dữ liệu: 12](#_Toc121380279)

[5. Kết hợp dữ liệu: 13](#_Toc121380280)

[6. Diễn dịch và trực quan hóa dữ liệu: 13](#_Toc121380281)

[IV. NHỮNG CƠ HỘI LỚN 14](#_Toc121380282)

[V. NGHIÊN CỨU TÌNH HUỐNG 15](#_Toc121380283)

[A. Dữ liệu lớn trong việc phát hiện sự cố tràn dầu 15](#_Toc121380284)

[B. Truy xuất hình ảnh dựa trên nội dung từ các kho lưu dữ liệu siêu phổ 18](#_Toc121380285)

[VI. KẾT LUẬN 19](#_Toc121380286)

**Bài báo này phân tích những thách thức và cơ hội mà dữ liệu lớn có thể mang lại trong bối cảnh các ứng dụng viễn thám.**

**TÓM TẮT** | Mỗi ngày, một số lượng lớn các cảm biến trên không gian và trên trái đất từ nhiều quốc gia khác nhau cung cấp một lượng lớn dữ liệu cảm biến từ xa.

Những dữ liệu này được sử dụng cho các ứng dụng khác nhau, chẳng hạn như giám sát các nguy cơ tự nhiên, biến đổi khí hậu toàn cầu, quy hoạch đô thị, v.v. Dựa trên điều này, có thể nói rằng chúng ta hiện đang sống trong thời đại của dữ liệu viễn thám lớn. Hơn nữa, những dữ liệu này đang dần trở thành một lĩnh vực kinh tế và một nguồn lực quan trọng mới trong nhiều ứng dụng. Trong bài báo này, chúng mình phân tích cụ thể những thách thức và cơ hội mà dữ liệu lớn mang lại trong bối cảnh các ứng dụng viễn thám. Trọng tâm là phân tích chính xác ý nghĩa của dữ liệu lớn trong các ứng dụng viễn thám và làm thế nào dữ liệu lớn có thể cung cấp nhiều giá trị hơn trong bối cảnh này. Hơn nữa, bài báo này mô tả các vấn đề quan trọng nhất trong việc quản lý, xử lý và khai thác hiệu quả dữ liệu lớn cho các vấn đề viễn thám. Để diễn giải các khía cạnh nói trên, hai trường hợp điển hình giải thích việc sử dụng dữ liệu lớn trong viễn thám được minh họa.

Trong trường hợp đầu tiên, dữ liệu lớn được sử dụng để tự động phát hiện sự cố tràn dầu trên biển bằng cách sử dụng một kho lưu trữ lớn dữ liệu viễn thám.Trong trường hợp thứ hai, việc truy xuất thông tin dựa trên nội dung được thực hiện bằng cách sử dụng máy tính hiệu năng cao (HPC) để trích xuất thông tin từ một cơ sở dữ liệu lớn về hình ảnh viễn thám, thu được sau vụ tấn công khủng bố ở Trung tâm Thương mại Thế giới ở Thành phố New York (WTC). Cả hai trường hợp đều được sử dụng để minh họa những thách thức và cơ hội quan trọng do việc sử dụng dữ liệu lớn trong các ứng dụng viễn thám mang lại.

# LỜI MỞ ĐẦU

Vì có các bộ tạo dữ liệu động, nên con người có thể tạo ra dữ liệu mỗi ngày. Tất cả chúng ta đều được kết nối thông qua việc chia sẻ dữ liệu từ mạng xã hội, thiết bị thông minh, v.v. Viễn thám đã được sử dụng rộng rãi để quan sát hành tinh của chúng ta từ nhiều góc độ khác nhau và để làm cho cuộc sống của chúng ta dễ dàng hơn. Ta có thể nói rằng toàn bộ trái đất bây giờ đã được làm bằng kỹ thuật số. Do đó, việc Trái đất số hóa cộng với các bộ tạo dữ liệu động là tác nhân chính cho dữ liệu lớn trong viễn thám, được sử dụng để giúp cuộc sống dễ dàng hơn (ví dụ: cải thiện các dịch vụ như cảnh sát, chăm sóc sức khỏe và giao thông) và cũng để kinh doanh, tức là để cải thiện việc ra quyết định, việc sản xuất, đổi mới sản phẩm, trải nghiệm và dịch vụ người tiêu dùng v.v.

Theo báo cáo của IBM, 2,5 tỷ byte dữ liệu được tạo ra mỗi ngày. Nói cách khác, 90% dữ liệu trên thế giới ngày nay đã được tạo ra chỉ trong hai năm qua. Chúng ta thực sự đang sống trong thời đại dữ liệu lớn, và bây giờ các nhà lãnh đạo chính phủ, các doanh nghiệp và các tổ chức phi lợi nhuận đang nhanh chóng nhận ra rằng nó rất quan trọng trong việc thu thập dữ liệu lớn trong các bối cảnh khác nhau. Tuy nhiên, vẫn tồn tại một vấn đề phổ biến liên quan đến cách chúng ta có thể có được cái nhìn trong dữ liệu lớn. Vấn đề này là một câu hỏi hóc búa: Một mặt, rất nhiều dữ liệu lớn có thể mang lại cho chúng ta những cơ hội lớn. Mặt khác, chúng ta vẫn chưa biết làm thế nào để khai thác lượng dữ liệu lớn như vậy với sự linh hoạt, đa dạng, không đồng nhất, với các giá trị tiềm năng cao. Điều này làm cho dữ liệu rất khó để được xử lý và phân tích trong một thời gian hợp lý.

Dữ liệu lớn chủ yếu có thể được đặc trưng bởi ba yếu tố: kích cỡ, sự đa dạng và tốc độ, được định nghĩa là 3 "V" (volume, variety, velocity) bởi Meta Group (nay gọi là Gartner) vào năm 2001. Ta cần lưu ý rằng "giá trị" là một thành phần quan trọng của dữ liệu lớn, nhưng nó không phải là một đặc điểm xác định được. Do đó dữ liệu viễn thám lớn có thể được mô tả theo cách của chính nó (gọi là 3V).

1. Về kích cỡ: Dữ liệu lưu trữ được thiết kế bởi kích cỡ tăng dần của chúng, từ terabyte (TB = 1024 GB) đến petabyte (PB = 1024 TB), và thậm chí đến exabyte (EB = 1024 PB). Ví dụ, một lượng lớn dữ liệu viễn thám hiện đang có sẵn miễn phí từ Chương trình Chính phủ Mở rộng của NASA. Chỉ có một trong những kho lưu trữ của NASA, Hệ thống Thông tin và Dữ liệu Khoa học Trái đất (ESDIS), nắm giữ 7,5 PB dữ liệu với gần 7000 bộ dữ liệu độc nhất và 1,5 triệu người dùng trong năm 2013. Kích cỡ này chỉ chứa được trong dữ liệu viễn thám miền.
2. Về độ đa dạng: Chúng ta có thể thấy dữ liệu viễn thám lớn bao gồm đa nguồn (laser, radar, quang học, v.v.), đa thời gian (được thu thập vào các ngày khác nhau) và đa độ phân giải (độ phân giải không gian khác nhau) dữ liệu viễn thám, cũng như dữ liệu từ các ngành khác nhau tùy thuộc vào nhiều lĩnh vực ứng dụng.
3. Về tốc độ: Tốc độ của dữ liệu lớn trong viễn thám không chỉ tạo ra dữ liệu với tốc độ phát triển nhanh, mà còn hiệu quả trong quá trình xử lý và phân tích dữ liệu. Nói cách khác, dữ liệu này nên được phân tích theo thời gian thực hoặc thời gian hợp lý để đạt được một nhiệm vụ nhất định, ví dụ: chỉ vài giây có thể cứu sống hàng trăm nghìn người trong một trận động đất.

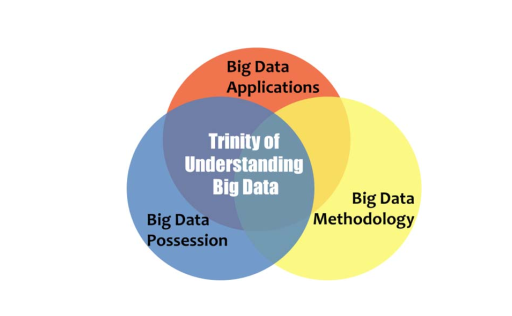
Mặc dù 3V có thể mô tả dữ liệu lớn, nhưng có thể nói rằng dữ liệu lớn trong viễn thám không cần thiết phải thỏa mãn cả 3 V. Ngoài những thách thức thường gặp của dữ liệu lớn đặc trưng bởi 3V, vẫn còn những thách thức khác cho các ứng dụng viễn thám, chẳng hạn như mở rộng khả năng tích hợp nhiều hệ thống quản lý khác nhau cho các trường hợp khác nhau cho một trung tâm dữ liệu viễn thám. Phần đặc biệt quan trọng là giá trị của dữ liệu, một chất lượng quan trọng ẩn trong dữ liệu lớn. Các phương thức xử lý dữ liệu có thể được sử dụng để khai thác các giá trị này, và sau đó giá trị của dữ liệu lớn có thể được phát huy trong một ứng dụng viễn thám thực tế.

Do đó, để hiểu rõ hơn về dữ liệu lớn, có ba điều cần được thống nhất: sở hữu dữ liệu, các ứng dụng dữ liệu, và các phương thức dữ liệu. Các khía cạnh của 3 điều như vậy (gọi là tam giác - trinity) đều có những thách thức chung và khác nhau, đều cho thấy những thách thức chung và khác nhau này đều có những thách thức riêng.

Trong phần này, những thách thức thường gặp và riêng biệt này được thảo luận trong bối cảnh các ứng dụng của viễn thám. Bất chấp những thách thức lớn như vậy, những tiềm năng của dữ liệu viễn thám lớn vẫn được trình bày 1 cách chi tiết. Những tiềm năng này đã được áp dụng để đối phó với các vấn đề thực tế khác nhau, chẳng hạn như khảo cổ học, đánh giá cây trồng và dự báo nông sản, an ninh lương thực, sức khỏe con người. Để minh họa sự hiệu quả của dữ liệu viễn thám lớn, hai bài test điển hình về việc sử dụng dữ liệu lớn trong viễn thám đã được trình bày trong bài báo này. Trong trường hợp đầu tiên, dữ liệu truyền thông xã hội cùng với hình ảnh viễn thám được xác định bao gồm dữ liệu viễn thám lớn để phát hiện tự động sự cố tràn dầu trên biển. Trong trường hợp thứ hai, việc truy xuất thông tin dựa trên nội dung được thực hiện bằng cách sử dụng máy tính hiệu suất cao (HPC) để trích xuất thông tin từ một cơ sở dữ liệu lớn về hình ảnh viễn thám, thu được sau vụ tấn công khủng bố vào Trung tâm Thương mại Thế giới ở Thành phố New York vào ngày 11 tháng 9 năm 2001.

Phần tiếp theo thảo luận về sự hiểu biết của chúng ta về dữ liệu lớn trong viễn thám từ ba góc nhìn khác nhau. Theo góc nhìn của tụi mình về dữ liệu lớn, Phần III chia các thách thức dữ liệu lớn thành các thách thức chung cho tất cả các ứng dụng viễn thám và các thách thức riêng biệt của cái gọi là ba phần của dữ liệu lớn (trinity – sở hữu dữ liệu, ứng dụng dữ liệu, phương thức dữ liệu). Các tiềm năng của dữ liệu viễn thám lớn được trình bày trong Phần IV. Phần V trình bày các nghiên cứu điển hình nói trên về dữ liệu lớn trong viễn thám. Cuối cùng, Phần VI rút ra một số kết luận việc và thảo luận về những phát triển trong tương lai.

# II. TÌM HIỂU DỮ LIỆU LỚN TRONG VIỄN THÁM



**Hình 1**

Từ góc nhìn chung, chúng ta có thể hiểu dữ liệu lớn có những ý nghĩa khác nhau tùy vào những người sở hữu nó, những người có thể xử lý, phân tích và sử dụng dữ liệu lớn. Theo đó, các phương pháp dữ liệu khác nhau có thể được khai thác để giải quyết các thách thức về dữ liệu lớn để thu được giá trị của những dữ liệu đó một cách hiệu quả. Một trinity (ba khía cạnh) được thảo luận về dữ liệu lớn (hướng cụ thể vào các ứng dụng viễn thám). Ở đây, chúng mình xác định ba khía cạnh để hiểu dữ liệu lớn, tức là sở hữu dữ liệu, phương thức dữ liệu và ứng dụng dữ liệu, chúng cùng đóng góp vào một chu kì dữ liệu lớn duy nhất. Khái niệm tam giác của dữ liệu lớn được minh họa trong Hình 1. Những thách thức phổ biến và khác nhau trong các khía cạnh riêng biệt trong việc hiểu được dữ liệu lớn được nêu chi tiết sau đây.

## KHÍA CẠNH ĐẦU TIÊN: SỞ HỮU DỮ LIỆU

Đây là một khía cạnh quan trọng của dữ liệu lớn dựa trên việc ta có thể xác định các ứng dụng và sử dụng hoặc thiết kế các phương pháp dữ liệu phù hợp để giải quyết một vấn đề thực tế (ví dụ vấn đề viễn thám). Các cơ hội tương ứng dựa trên thực tế dữ liệu đa dạng hơn có thể được thu thập bởi các thiết bị thông minh, chúng có thể trở thành các máy tạo dữ liệu động và riêng biệt, nơi hầu hết con người có thể truy cập internet hiện nay. Theo đó, các giá trị dữ liệu có thể được lấy từ các dữ liệu viễn thám phức tạp, đa dạng, không đồng nhất, nhiều chiều và dữ liệu khác từ không gian mạng. Tuy nhiên, những thách thức lớn này nảy sinh ở mỗi bước khi thu thập và tổ chức dữ liệu viễn thám lớn. Ví dụ: dữ liệu viễn thám được thu thập từ vệ tinh, máy bay hoặc các thiết bị viễn thám khác trong khi các dạng dữ liệu khác được truy xuất từ không gian mạng. Dữ liệu viễn thám được xử lý trước bằng cách hiệu chỉnh hình học và điện từ, loại bỏ tiếng ồn, v.v. và dữ liệu từ không gian mạng nên được làm sạch để giảm lỗi và tiếng ồn, khi đó chất lượng dữ liệu có thể được cải thiện. Dữ liệu viễn thám nên được chuyển từ vệ tinh đến các trạm mặt đất, và từ các trạm mặt đất đến người dùng. Các vấn đề khác là nén dữ liệu, lưu trữ dữ liệu, truy xuất dữ liệu, quyền và bảo vệ dữ liệu, v.v. Ở đây, chúng mình nhấn mạnh rằng dữ liệu không có giá trị cho đến khi chúng được sử dụng được trong các ứng dụng. Sự khác biệt chính giữa dữ liệu truyền thống và dữ liệu lớn là làm thế nào để xác định các bộ dữ liệu phù hợp và làm thế nào để kết hợp chúng để giải quyết một vấn đề khó khăn hoặc mới lạ.

## B. KHÍA CẠNH THỨ HAI: PHƯƠNG THỨC DỮ LIỆU LỚN

Một phương thức dữ liệu lớn nên được thiết kế để giải quyết một cách có hệ thống các vấn đề dữ liệu lớn từ các miền viễn thám khác nhau. Phương thức này được sử dụng để tạo ra các phương thức dữ liệu mới cho dữ liệu viễn thám lớn, triển khai dữ liệu, trích xuất thông tin, mô hình hóa dữ liệu, hợp nhất dữ liệu, trực quan hóa dữ liệu và giải thích dữ liệu. Những khía cạnh này đặc biệt quan trọng đối với các ứng dụng cảm biến tại chỗ, trong đó các bước tiền xử lý cũng quan trọng như các bước trích xuất thông tin. Tuy nhiên, xử lý và phân tích dữ liệu đại diện cho các phương pháp dựa trên dữ liệu có thể khác đáng kể so với góc nhìn của các ứng dụng và tên miền cụ thể.

Do tính không đồng nhất và độ chính xác cao nói trên của dữ liệu lớn trong viễn thám, ta cũng phải đối mặt với các thách thức về tính toán và thống kê quan trọng liên quan đến khả năng mở rộng xử lý, tích lũy tiếng ồn và lỗi đo lường. Những thách thức này đòi hỏi các kỹ thuật thống kê và tính toán mới để có thể phân tích và xử lý dữ liệu lớn. Các kỹ thuật phân tích và xử lý dữ liệu được thúc đẩy và có thể có được từ các lý thuyết và phương pháp từ các lĩnh vực thống kê, học máy, nhận dạng mẫu, trí tuệ nhân tạo, khai thác dữ liệu, v.v. Do vậy nên kiến thức về tên miền là một lĩnh vực quan trọng khác nên được liên kết chặt chẽ với việc phân tích dữ liệu.

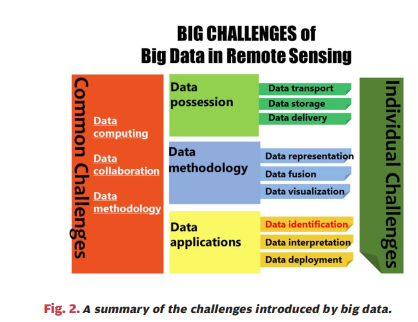
## C. KHÍA CẠNH THỨ BA: ỨNG DỤNG DỮ LIỆU LỚN

Mục tiêu chính trong các ứng dụng dữ liệu lớn là xác định dữ liệu chính xác để giải quyết các vấn đề hiện tại khó giải quyết hoặc hầu hết không thể thao tác với dữ liệu viễn thám truyền thống. Sau đó, vấn đề tiếp theo là làm thế nào để thu thập, tổ chức và tối ưu hóa những dữ liệu lớn này để giải quyết các vấn đề viễn thám thực sự.

Để xác định dữ liệu phù hợp, chúng ta nên liên kết chặt chẽ với khía cạnh đầu tiên. Nói cách khác, để khai thác dữ liệu lớn trước tiên, người ta nên lấy dữ liệu từ các đại lý dữ liệu liên quan (hoặc tổ chức ngành công nghiệp dữ liệu). Để truy cập dữ liệu, sự hợp tác giữa các lĩnh vực hoặc tổ chức nên được tính đến một cách hiệu quả. Đây là một trong những thách thức quan trọng trong các ứng dụng viễn thám. Sau khi có được dữ liệu phù hợp, chẳng hạn như dữ liệu viễn thám, dữ liệu văn bản và hình ảnh từ các mạng xã hội, các phương pháp dữ liệu sáng tạo nên được phát triển để khai thác, nhận thức được và chứng minh giá trị của các ứng dụng viễn thám trong dữ liệu lớn.

# III. DỮ LIỆU LỚN, THÁCH THỨC LỚN

Thách thức lớn của dữ liệu lớn trong viễn thám không chỉ liên quan đến khối lượng lớn dữ liệu. Đặc biệt, các thách thức về thu thập, lưu trữ, quản lý và phân tích dữ liệu cũng liên quan đến các vấn đề viễn thám liên quan đến dữ liệu lớn. Trong phần này, chúng mình đặc biệt phân tích những thách thức của dữ liệu lớn trong viễn thám liên quan đến các khía cạnh khác nhau trong các hiểu biết dữ liệu lớn ở phần trước.



Từ các quan điểm khác nhau trong việc hiểu dữ liệu lớn, chúng ta đang phải đối mặt với những thách thức lớn trong việc tận dụng giá trị mà dữ liệu mang lại. Trong ba khía cạnh này, có những thách thức tương tự được chia sẻ, chẳng hạn như tính toán dữ liệu, kết hợp dữ liệu và phương thức dữ liệu cho các ứng dụng khác nhau; trong khi đó, chúng ta đang phải đối mặt với những thách thức khác trong các khía cạnh hiểu biết riêng về dữ liệu lớn. Hình 2 tóm tắt những thách thức chung và riêng, được mô tả chi tiết trong các phần tiếp theo.

## NHỮNG THÁCH THỨC THƯỜNG GẶP

Trong những thách thức sau, có ba thách thức thường gặp: điện toán dữ liệu lớn, kết hợp dữ liệu lớn và phương thức dữ liệu lớn, được liệt kê theo mức độ hiểu biết về dữ liệu lớn trong viễn thám.

### Điện toán dữ liệu lớn:

Một thách thức trong việc thiết kế các hệ thống hiệu suất cao cho điện toán dữ liệu lớn là phát triển các hệ thống không đồng nhất hơn có thể tích hợp các tài nguyên ở các địa điểm khác nhau. Mặc dù các hệ thống điện toán đám mây đã được chứng minh là có hiệu suất tổng hợp cao trong các ứng dụng viễn thám, nhưng vẫn tồn tại những thách thức liên quan đến việc kết hợp dần khái niệm điện toán đám mây vào các nghiên cứu viễn thám. Mục tiêu cuối cùng là giúp cho các bộ dữ liệu phân tán dễ dàng truy cập từ những người dùng khác nhau. Tuy nhiên, thách thức còn lại là mức tiêu thụ năng lượng, vẫn khó có thể tận dụng trong các nền tảng song song lớn hoặc thậm chí trong các kịch bản xử lý ngay tại chỗ. Giải quyết những thách thức này sẽ rất quan trọng trong việc kết hợp đầy đủ các kỹ thuật tính toán dữ liệu lớn vào các ứng dụng viễn thám. Thông thường, một mô hình HPC được khai thác để xử lý dữ liệu lớn thời gian thực.

### Kết hợp dữ liệu lớn:

Quyền sở hữu dữ liệu trong các vấn đề viễn thám thường được phân mảnh qua các tác nhân hoặc các ngành công nghiệp dữ liệu. Theo đó, việc truy cập và kết nối dữ liệu có thể là một trở ngại. Các sự quan tâm hợp pháp có thể được đưa ra để có được sự hợp tác giữa các ngành thúc đẩy chia sẻ dữ liệu, chẳng hạn như văn bản xã hội hoặc phương tiện truyền thông xã hội. Tuy nhiên, các cá nhân thường chống lại việc chia sẻ dữ liệu cá nhân do vấn đề bảo mật và quyền riêng tư. Điều này mâu thuẫn với ý tưởng cá nhân hóa dữ liệu. Ngoài ra, nhiều công ty dữ liệu coi dữ liệu lớn là độc quyền và do đó không được khuyến khích chia sẻ dữ liệu. Đồng thời, đó cũng là một thách thức quan trọng đối với các tổ chức chính phủ trong việc chia sẻ dữ liệu trừ khi tất cả những người tham gia có thể đạt được lợi ích vật chất và ưu đãi trong việc chia sẻ dữ liệu nhiều hơn các rủi ro. Ví dụ, ngay cả khi NASA hiện đang chia sẻ một lượng đáng kể dữ liệu viễn thám theo sáng kiến của chính phủ mở, hầu hết các hình ảnh chất lượng cao, độ phân giải cao vẫn không có sẵn với công chúng. Do đó, cần phải tìm ra những cách thức hợp tác mới để cải thiện khả năng truy cập dữ liệu lớn trong các vấn đề viễn thám.

### Phương thức dữ liệu lớn:

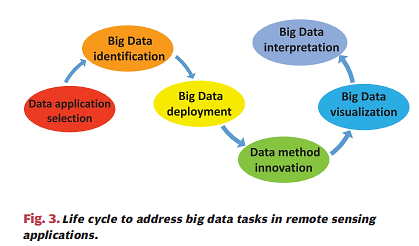
Vấn đề phân tích dữ liệu lớn trong viễn thám có thể được chính thức hóa một cách đơn giản như sau. Cho X là một tập dữ liệu đầu vào và cho f(x) là một hàm ánh xạ giữa một đầu vào x X và đầu ra y. Sau đó, một nhiệm vụ phân tích dữ liệu chung có thể được xây dựng như phương trình:

y = f(x)

Trong đó việc xử lý tương ứng có thể được thực hiện trong bộ nhớ của máy tính có chứa đầu vào dữ liệu.

Tuy nhiên, phân tích dữ liệu lớn nói chung nên áp dụng một cơ chế để phân chia đầu vào dữ liệu thành một kiến trúc phân tán và/hoặc song song, tức là X = {X1, X2, …XN} có nghĩa là chia nhỏ bộ dữ liệu lớn hơn X thành N bộ dữ liệu nhỏ hơn. Các phương pháp hoặc thuật toán dữ liệu được áp dụng nên được sửa đổi để thích ứng với các môi trường tính toán mới. Mặc dù, nói chung, đây là một sự đơn giản hóa (vì các bộ dữ liệu nhỏ hơn có thể không dễ dàng để xử lý độc lập và liên quan đến một số đồng bộ hóa và/hoặc truyền thông trong tác vụ xử lý liên quan), một thách thức quan trọng đối với sơ đồ xử lý này là không phải tất cả các thuật toán đều có thể được phân phối hoặc thực hiện hiệu quả ở dạng song song. Ngay cả khi các phương pháp xử lý dữ liệu có thể làm như vậy, thì việc thu thập dữ liệu phân tán và cung cấp dữ liệu đó đến các node điện toán phù hợp cũng là một thách thức. Do đó, xử lý dữ liệu lớn nói chung (và viễn thám nói riêng) cần có các mô hình tính toán và thống kê mới liên quan đến các chiến lược xử lý dữ liệu tiêu chuẩn.

## B. NHỮNG THÁCH THỨC RIÊNG BIỆT



Trong phần này, một vài thách thức quan trọng được thảo luận trong bối cảnh thiết kế một chu trình của dữ liệu viễn thám lớn (hình 3). Sau khi hiểu được những yêu cầu công việc (vd một ứng dụng viễn thám liên quan đến dữ liệu lớn), một vài bước quan trọng để xác định loại đúng của dữ liệu thông qua các quy tắc khác nhau để phát triển dữ liệu lớn, để tối ưu hóa hoặc thiết kế các phương thức dữ liệu mới và cuối cùng để trực quan hóa và diễn dịch các kết quả có được. Ở đây, các phương thức dữ liệu bao gồm phân tích dữ liệu, mô hình dữ liệu, xử lí dữ liệu, vv.

### Xác minh dữ liệu chính xác:

Dữ liệu viễn thám lớn thường bao gồm dữ liệu miền và ngoài miền. Trong quá khứ, các định dạng khác nhau của dữ liệu hiếm khi được kết hợp với nhau để hoàn thành các nhiệm vụ/ứng dụng viễn thám. Vì vậy, dữ liệu không có giá trị trừ khi được khai thác để hoàn thành 1 nhiệm vụ cụ thể. Do đó, sự khác biệt chính giữa dữ liệu viễn thám truyền thống và dữ liệu viễn thám lớn nằm ở cách chọn và kết hợp các định dạng khác nhau của dữ liệu để xác định được các vấn đề của thế giới thực trước đây được coi là nan giải. Đây là một thách thức chính của dữ liệu lớn, cách ta xác định và khai thác dữ liệu chính xác để giải quyết vấn đề trong tầm tay.

Trong viễn thám, ta có nhiều loại dữ liệu khác nhau, bao gồm dữ liệu quang học (vd đa phổ và siêu quang phổ), dữ liệu rada (vd rada tổng hợp (SAR)), hoặc dữ liệu laze (vd cảm biến ánh sáng và tầm xa (LiDAR)) được cung cấp bởi máy bay hoặc vệ tinh hoặc bộ cảm biến dưới mặt đất. Các loại dữ liệu khác cũng có thể được tích hợp ở các vấn đề về viễn thám, dữ liệu văn bản trên internet (các bài log trên web, tin tức,…) có thể được dùng để làm các họa tiết chú thích cho dữ liệu được cung cấp bởi việc viễn thám, như là thông qua việc học hoặc các kĩ thuật huy động cộng đồng, điều mà có thể làm được ở mức phí thấp hoặc không tốn phí. Bên cạnh đó, dữ liệu hình ảnh được chụp bởi các cá thể từ mạng xã hội có thể được đưa vào hỗ trợ cho các nhiệm vụ diễn dịch dữ liệu viễn thám. Các định dạng dữ liệu khác như là dữ liệu điều tra dân số, dữ liệu khí tượng, dữ liệu giao thông thông minh, dữ liệu địa lí chính xác cao, dữ liệu chăm sóc sức khỏe, vv có thể đóng vai trò quan trọng trong việc giải quyết một vấn đề cụ thể ở thế giới thực, ví dụ như giám sát an toàn thực phẩm.

### Các thách thức trong quyền sở hữu dữ liệu:

Sau khi dữ liệu đã được chuyển đến trạm ở dưới mặt đất, những dữ liệu đó nên được lưu trong 1 hệ thống. Một hệ thống chứa dữ liệu thường bao gồm phần cứng và phần mềm. Trước đây, cơ sở hạ tầng phần cứng nên thích ứng 1 cách linh động được với các môi trường ứng dụng khác nhau. Trong các điều kiện phần mềm, hệ thống lưu dữ liệu thường được trang bị với nhiều giao diện đa dạng, lưu trữ dữ liệu và các truy vấn từ web của người dùng. Với sự phát triển nhanh chóng của dữ liệu viễn thám, các hệ thống quản lý cơ sở dữ liệu liên quan đến các cấu trúc trước đây (theo cách truyền thống) (RDBMSs) không thể đáp ứng yêu cầu trong việc quản lí dữ liệu lớn trong viễn thám. Do vậy, việc tiếp nhận hoặc thiết kế 1 nơi lưu dữ liệu lớn trong phạm vi PB (petabyte) hoặc lớn hơn là điều hết sức cần thiết.

Chuyển giao dữ liệu cung cấp kết nối đến dữ liệu viễn thám và metadata đến người dùng, cả 2 đều ở các trạm dưới mặt đất hoặc các mạng lưới các trạm mặt đất nhận dữ liệu. Thông thường, việc này bao gồm các cổng web đồ họa cung cấp kết nối đến dữ liệu và tìm kiếm metadata đến người dùng. Theo cách truyền thống, người dùng tải dữ liệu của internet từ một trung tâm lưu trữ đến máy tính của họ để phân tích. Việc này không thể làm được trong các ứng dụng dữ liệu lớn vì sự gia tăng nhanh của kích thước dữ liệu không cho phép hệ thống hiện thời chuyển dữ liệu đến máy tính người dùng. Cụ thể, nếu một sự cố khẩn cấp như là động đất, 1 lượng lớn dữ liệu nên được phân tích trong 1 thời gian rất ngắn vì chỉ vài giây có thể cứu được rất nhiều người bằng cách cảnh báo đúng lúc. Đây là một thách thức lớn khác cho những người sở hữu dữ liệu lớn, vì các dữ liệu đa dạng và đa chiều nên được phân tích trong một thời gian ngắn dựa vào kích cỡ và tốc độ của dữ liệu lớn. Vì vậy, một nền tảng phân tích dữ liệu lớn thời gian thực nên được phát triển để giải quyết dữ liệu viễn thám online cùng với offline trong trung tâm dữ liệu địa phương hoặc ở các vùng hẻo lánh trong các ứng dụng thời gian thực, như là thông báo thời tiết, cảnh báo nguy hiểm, vv.

### Phát triển dữ liệu:

Như đã được thảo luận ở phần 3 B1, có 1 thách thức lớn là xác minh nguồn dữ liệu chính xác để đạt được mục tiêu cụ thể, điều mà khó thực hiện khi không có dữ liệu lớn. Một thách thức khác của dữ liệu lớn là cách triển khai dữ liệu cho các ứng dụng thực tế. Trong các ứng dụng dữ liệu lớn, việc triển khai dữ liệu lớn bao gồm sự chuẩn bị dữ liệu, các công nghệ quản lí dữ liệu, các phương thức dữ liệu và các kĩ thuật, vv. Đó là cách có được dữ liệu, chứa dữ liệu trong môi trường máy tính và cách xây dựng các mô hình để có cái nhìn xuyên suốt của dữ liệu lớn do đó nên đươc thiết kế 1 cách cẩn thận trong các bước phát triển dữ liệu lớn. Do các đặc tính kích cỡ và tốc độ của dữ liệu lớn, các phương pháp truyền thống không thể dùng cho các mục đích triển khai được. Do đó, các công nghệ mới nên được đem vào sử dụng, như là công nghệ quản lí dữ liệu phân tán, các mô hình dữ liệu giản đồ, các kĩ thuật trực quan hóa và vâng vâng để có được các góc nhìn quan trọng trong dữ liệu lớn.

### Biểu diễn dữ liệu:

Nhiều nguồn dữ liệu viễn thám đa dạng có nhiều độ phân giải quang phổ và không gian khác nhau và thường thu được vào nhiều ngày khác nhau. Ví dụ, trong dữ liệu quang phổ các kí hiệu quang phổ của mọi vật đều là độc nhất dựa vào phép đo ở phòng lab. Tuy nhiên, các kí hiệu quang phổ của trường dữ liệu có thể thay đổi dựa vào sự đa dạng của các vật, hiệu ứng môi trường, ô nhiễm ở bề mặt, các ảnh hưởng của các vật xung quanh, thay đổi theo mùa, vv. Điều này có thể dẫn đến hiện tượng các kí hiệu tương đồng có thể biểu thị các vật khác nhau trong khi các kí hiệu khác nhau có thể biểu thị các vật giống nhau. Hiện tượng này giống với “khoảng cách ngôn ngữ” ở góc nhìn máy tính, sự đa dạng giữa thông tin từ dữ liệu và kiến thức được diễn dịch bởi người dùng. Trong những năm gần đây, DNNs đã thành công giải quyết các vấn đề cụ thể ở góc nhìn máy tính để xóa bỏ khoảng cách đó thông qua các tính năng tự động giải nén. Mặc dù DNNs đã du nhập các lựa chọn tính năng và các tác vụ cụ thể khi phân tích các hình ảnh viễn thám, trong hầu hết các công việc được tiến hành hiện nay chỉ có các tính năng quang phổ hoặc chuyển đổi quang phổ (các thành phần chủ chốt) được sử dụng như là đầu vào cho DNNs tạo ra các hình ảnh viễn thám “tốt hơn”. Trong các công việc, thông thường các thành phần từ tính năng gốc trao đổi sự phản xạ quang phổ của các mục tiêu dưới mặt đất, các thành phần này đều có 1 vai trò vật lí rõ ràng. Ở đây, các hình ảnh viễn thám có thể được diễn dịch như là cấu trúc dữ liệu. Mặc dù độ mạnh và chính xác cụ thể có thể được cải thiện 1 chút bằng việc kết hợp những mẫu chưa được dán nhãn cho việc cải thiện tính năng, ta không rõ hiệu năng cụ thể là bao nhiêu dựa vào các đại diện dữ liệu “tốt hơn” do sự thiếu sót dữ liệu huấn luyện cho máy (training data) và số lượng các lớp quang phổ có hạn có thể được dùng trong việc thực thi DNNs. Ví dụ, độ chính xác cụ thể không thể được cải thiện đáng kể với hơn 5 lớp quang phổ.

Bên cạnh đó, mặc dù nhiều loại dữ liệu viễn thám (có được từ các bộ cảm khác nhau, từ nhiều địa điểm khác nhau ở nhiều ngày khác nhau) có được và được khai thác để giải quyết vấn đề thách thức của ứng dụng cùng với dữ liệu phi viễn thám, nhưng vẫn tồn taị các phương thức dữ liệu không thể vận dụng những dữ liệu đó để lấy lại giá trị dữ liệu. Trong khi đó, dữ liệu viễn thám bao gồm nhiều chiều và độ phân giải không gian khác nhau, như là máy quang phổ sóng điện từ hình ảnh có độ phân giải trung bình trong không gian (MODIS) trong 36 dải quang phổ trong phạm vi 250m dưới mặt đất (dải 1 – 2), 500m (dải 3 – 7) và 1000m (dải 8 – 36) và máy quang phổ hình ảnh hệ thống phản quang trong không trung (ROSIS) với độ phân giải dưới mặt đất ít hơn 1m trong 115 dải quang phổ. Hơn nữa, sự hiện diện của dữ liệu phi viễn thám có thể phi cấu trúc được (vd các hình ảnh cá nhân) điều mà khác đáng kể với các dữ liệu viễn thám quang học hoặc sóng điện từ. Vì vậy, các loại dữ liệu khác nhau trở thành một chướng ngại lớn trong việc khai thác dữ liệu viễn thám lớn.

### Kết hợp dữ liệu:

Do các thách thức về biểu diễn dữ liệu được thảo luận ở phần 3 B4, một khó khăn tiếp theo là cách tích hợp dữ liệu từ nhiều nguồn khác nhau nơi mà các loại dữ liệu khác nhau nhiều (vd các kí hiệu quang phổ trong dữ liệu viễn thám quang học, sóng điện từ trong dữ liệu rada, các tính năng cấu trúc của văn bản, các tính năng phi cấu trúc của các hình được chụp bởi máy kĩ thuật số, vv).

Thông thường, việc kết hợp dữ liệu có thể được thực hiện ở mức độ pixel, mức độ tính năng và ở mức độ ra quyết định. Tuy nhiên, dữ liệu lớn trong viễn thám thường bao gồm nhiều tỉ lệ và/hoặc định dạng. Kết quả là, các phương pháp tiếp cận truyền thống không thể tối ưu hóa trong việc tích hợp thông tin cho việc kết hợp dữ liệu. Vì vậy, những phương pháp mới phải được phát triển để giải quyết việc kết hợp dữ liệu lớn trong viễn thám. Ví dụ, trong các ứng dụng đô thị, mỗi pixel có thể được chú thích bởi các hình chụp bởi những cá thể từ mạng xã hội ở cùng một địa điểm bằng các kĩ thuật huy động cộng đồng. Đo lường sự tương quan giữa nhiều nguồn dữ liệu khác nhau cũng có thể trở thành 1 thách thức bổ sung bởi sự giúp đỡ của trí tuệ nhân tạo, khai thác dữ liệu, học máy, hoặc số liệu thống kê.

### Diễn dịch và trực quan hóa dữ liệu:

Trực quan hóa không chỉ cho phép người dùng, người ra quyết định có được cái nhìn tốt hơn về dữ liệu lớn, nó còn quan trọng trong việc hiểu và phân tích dữ liệu lớn trong viễn thám để đưa ra các chi tiết dữ liệu thích hợp về các mục tiêu và vật thể hiện tại. Do đó, việc trực quan hóa nên được cân nhắc sớm cùng với các tác vụ quan trọng ở hình 3, như là thu thập và xử lí dữ liệu. Điều này yêu cầu kĩ thuật trực quan hóa mới cùng với kiến thức lĩnh vực chuyên ngành trước đó về miền thông qua sự hợp tác chặt chẽ với các chuyên gia về miền đã đặt ra nhiệm vụ giải quyết các vấn đề thực tế.

Để ứng dụng trực quan hóa 1 cách hiệu quả, dữ liệu viễn thám lớn nên được tổng hợp lại từ nhiều nguồn khác nhau ở 1 lượng lớn và nhập vào các mô hình cho phép ra quyết định trong vài phút thay vì vài tuần hoặc vài tháng. Đây là 1 thách thức lớn với lượng đầu vào dữ liệu ở cấp độ PB hoặc lớn hơn, ví dụ trong các ứng dụng liên quan đến việc giám sát nguy hiểm. Vì vậy, việc trực quan hóa dữ liệu lớn phải giải quyết được các thách thức của dữ liệu cũng như sự khai thác dữ liệu để phát triển chúng. Ta cũng nên lưu ý rằng việc trực quan hóa dữ liệu phải được tiếp tục xuyên suốt chu trình của dữ liệu lớn, bên cạnh đó cũng có các thách thức riêng biệt trong nhiều bối cảnh khác nhau.

# IV. NHỮNG CƠ HỘI LỚN

Mặc cho các thách thức đã được nêu ở trên, giá trị tiềm năng của dữ liệu viễn thám lớn vẫn khá ấn tượng. Trong thực tế, các kĩ thuật viễn thám đã được sử dụng thành công trong nhiều ứng dụng khác nhau, như là ứng dụng nông nghiệp (vd giám sát an toàn thực phẩm, giám sát đồng ruộng), ứng dụng về biển (vd phát hiện tàu bè, phát hiện tràn dầu), kế hoạch hóa đô thị, giám sát đô thị, việc định cư (cả vùng đô thị và nông thôn), giảm sát an toàn thực phẩm, chất lượng nguồn nước, đánh giá năng lượng, các ca dịch bệnh, đánh giá hệ sinh thái, ấm lên toàn cầu, thay đổi toàn cầu, thẩm định tài nguyên rừng toàn cầu, khảo cổ, vv.

Được kết hợp với các hoạt động và dữ liệu của con người từ khoa học xã hội, nhiều kĩ thuật viễn thám nêu trên đã trở thành các công cụ mạnh mẽ trong việc cải thiện đáng kể sự hiệu quả trong sản xuất và nghiên cứu trong đời sống con người. Theo hướng này, dữ liệu viễn thám lớn cho ta tiềm năng đạt được các mục tiêu mà trước đây được cho là khó hoặc không thể theo cách truyền thống. Ví dụ, 1 khu di tích ẩn có thể được tìm thấy bởi dữ liệu viễn thám phân giải cao trong 1 khu rừng rậm không có cơ sở hạ tầng hiện đại, điều là một rào cản lớn với các nhà khảo cổ để tiếp cận. Một ứng dụng đã thành công ở Maya vùng Peten phía bắc Guatemala.

Trong các ứng dụng kế hoạch hóa đô thị, các hình ảnh viễn thám đo lường mặt đất cũng như là không trung và không gian được tích hợp để cho ra kết quả tốt hơn và đúng lúc hơn trong việc kế hoạch hóa, quản lí bền vững đô thị. Trong bối cảnh này, dữ liệu viễn thám có thể có được ở một khu vực rộng lớn theo 1 trình tự ở độ phân giải rất cao (ít hơn 1m/pixel) sử dụng các kĩ thuật viễn thám nâng cao. Các dự án liên quan bao gồm 100 dự án thành phố cho việc đặc tính môi trường đô thị, giám sát và ra quyết định của chính phủ và dấu chân trong các đô thị toàn cầu sử dụng độ phân giải không gian cao với tổng cộng 180000 ảnh TerraSAR-X và TanDEM-X cho việc xây dựng bản đồ định cư toàn cầu. Kết hợp với dữ liệu điều tra dân số, dữ liệu viễn thám được tích hợp để phát triển, sử dụng đất đai và đô thị ở Puerto Rico. Cùng với các thông số kinh tế xã hội, các hình ảnh vệ tinh phân giải cao đã được sử dụng để phân tích tốc độ tăng dân số điều mà liên quan chặt chẽ với tăng trưởng kinh tế.

Từ những năm 1990, dữ liệu viễn thám đã được sử dụng cho các ứng dụng nông lâm với sự thay đổi theo vùng và liên kết với các yếu tố khí tượng. Trong nông nghiệp, vai trò của dữ liệu viễn thám ngày càng trở nên quan trọng trong nông nghiệp bền vững, bao gồm an toàn thực phẩm hoặc đánh giá điều kiện mùa vụ và dự báo năng suất. Thêm vào đó, khoảng cách năng suất trung bình giữa các nước khá là lớn ở các cây ngũ cốc, bắp, lúa mì và gạo, vv. Thông thường, sự tăng cường nông nghiệp có thể rút ngắn đáng kể khoảng cách này. Ở đây, viễn thám đã chứng tỏ nó có công dụng lớn trong việc quan sát vụ mùa trong một khu vực lớn ví dụ như là lập bản đồ năng lượng tiềm năng của các vụ mùa bắp bằng cách kết hợp các ảnh hưởng của khí hậu và đất.

Cụ thể, an toàn thực phẩm là yếu tố chính của hệ thống nông nghiệp thông minh và chỉ có viễn thám từ các vệ tinh quan sát Trái Đất (vd Landsat, Resourcesat, MODIS) có thể cung cấp 1 cách thích hợp, lâu dài và dữ liệu chất lượng cao trong việc đặc tính hóa và lập bản đồ các thông số vụ mùa chính cho việc ước tính đất trồng trọt toàn cầu và phân tích an toàn thực phẩm trong việc kết hợp với các bản ghi (records) thông số của quốc gia, dữ liệu thực địa và dữ liệu thứ cấp (50 – 100 năm) về lượng mưa và nhiệt độ. Cùng với dữ liệu khảo sát sức khỏe và nhân khẩu học, nhiều ứng dụng có thể có lợi từ các phân tích của dữ liệu viễn thám và trong tương lai mối quan hệ giữa sức khỏe con người và thay đổi môi trường có thể được mô hình hóa một cách chính xác. Thêm vào đó, việc dữ liệu viễn thám có thể được dùng trong thị trường bảo hiểm toàn cầu, như là thiệt hại vụ mùa, lũ lụt và thẩm định nguy cơ cháy.

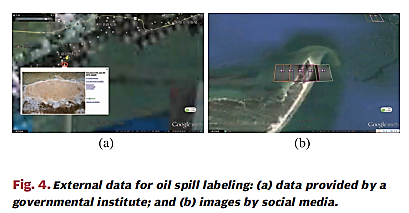
Nói chung, dữ liệu viễn thám cũng như là dữ liệu miền cho ta nhiều cơ hội lớn trong các ứng dụng trong khoa học tự nhiên, như là lập bản đồ mật độ cây phạm vi toàn cầu, cũng như khoa học xã hội như là nghiên cứu đô thị, nhân khẩu học, khảo cổ, chiến tranh và nghiên cứu xung đột, vv.

# V. NGHIÊN CỨU TÌNH HUỐNG

Trong phần này, hai nghiên cứu điển hình chứng minh tính hiệu quả của dữ liệu lớn trong các ứng dụng viễn thám được mô tả. Trong cả hai trường hợp, sử dụng các phương thức xử lý dữ liệu mới và kiến trúc điện toán mạnh mẽ là những điều cần thiết. Các vấn đề được nêu là phát hiện tràn dầu tự động và truy xuất thông tin dựa trên nội dung từ một kho dữ liệu viễn thám đa phổ, siêu phổ và dữ liệu liên quan từ các lĩnh vực khác, tương ứng.

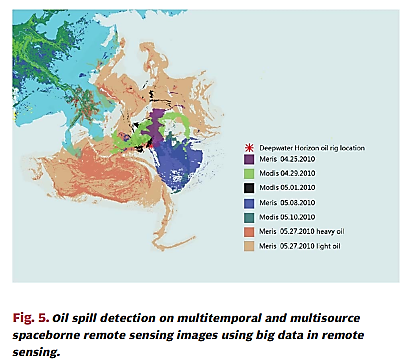
## Dữ liệu lớn trong việc phát hiện sự cố tràn dầu

Trong các ứng dụng phân loại viễn thám truyền thống, các mẫu được đánh dấu thu được dựa vào các cuộc khảo sát mặt đất, phân tích hình ảnh, hoặc kết hợp các kĩ thuật nói trên. Các cuộc khảo sát mặt đất có thể có độ chính xác cao khi dán nhãn nhưng những kỹ thuật tốn này kém và mất nhiều thời gian. Diễn dịch phân tích hình ảnh thì nhanh và rẻ, nhưng không thể đảm bảo chất lượng đánh dấu cao. Mặc dù các giải pháp kết hợp có thể tận dụng việc khảo sát mặt đất và phân tích hình ảnh trong hầu hết các vấn đề viễn thám, nhưng vẫn còn khó khăn trong việc đánh dấu phát hiện tràn dầu trên các hình ảnh viễn thám trên không gian đa thời gian và đa nguồn sử dụng dữ liệu lớn trong viễn thám. Việc đánh dấu tràn dầu trên biển bằng cách sử dụng các giải pháp kết hợp trong các điều kiện của dữ liệu viễn thám được cung cấp bởi các thiết bị trên tàu vũ trụ/ trên không. Do đó, việc đánh dấu tràn dầu trên biển mang lại một thách thức lớn cho nhiệm vụ phát hiện sự cố tràn dầu. Trong nghiên cứu điển hình này, trước tiên chúng mình xác định dữ liệu phù hợp bao gồm dữ liệu viễn thám lớn và sau đó giải quyết thách thức đánh dấu bằng phương pháp dữ liệu mới, tức là bằng cách tích hợp dữ liệu truyền thông xã hội với sự hỗ trợ của kĩ thuật cộng đồng và các kỹ thuật chủ động học.



Cụ thể, chúng mình đã chọn sự kiện tràn dầu lớn ở Vịnh Mexico (Hoa Kỳ) vào năm 2010 làm trường hợp nghiên cứu vì có thể thu được nhiều dữ liệu truyền thông xã hội và các dạng dữ liệu khác để phân tích dữ liệu lớn. Dữ liệu viễn thám quang học được sử dụng có chứa hình ảnh đa thời gian và đa nguồn, tức là dữ liệu quang phổ có độ phân giải trung bình (Meris), do Cơ quan Vũ trụ châu Âu (ESA) vận hành và quang phổ có độ phân giải vừa phải (MODIS), do NASA vận hành. Các hình thức dữ liệu khác trong ngữ cảnh này bao gồm dữ liệu truyền thông xã hội, tức là hình ảnh từ phương tiện truyền thông xã hội và từ mô tả văn bản. Ví dụ, hình ảnh từ Panoramio, một trang web chia sẻ ảnh định hướng theo vị trí địa lý, có thể dễ dàng thu được và thường được gắn thẻ địa lý, dưới dạng tọa độ chính xác của vị trí từ nơi những hình ảnh này đã được chụp, cũng như gắn thẻ văn bản (Hình 4a). Ngoài ra, dữ liệu trên không trong khu vực ô nhiễm có thể được sử dụng để đánh dấu các hình ảnh viễn thám, chẳng hạn như sự cố tràn dầu được phát hiện bởi các cảm biến trên không từ một trạm (Hình 4b). Các dạng dữ liệu lớn khác nhau có thể được sử dụng để cải thiện độ chính xác phát hiện sự cố tràn dầu trong bối cảnh này.

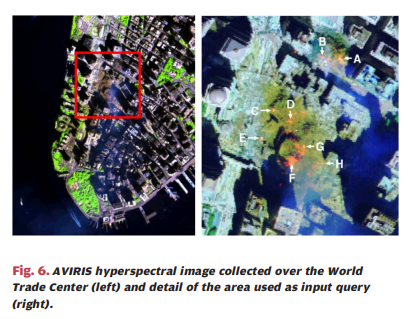
Ta cần lưu ý rằng quá trình gắn nhãn tốn nhiều thời gian để kết hợp ý tưởng về kĩ thuật cộng đồng và dữ liệu bên ngoài không thể bao gồm tất cả các pixel trong ảnh viễn thám. Theo đó, điều quan trọng là phải chọn một số lượng thông tin đơn giản để đánh dấu nhằm đảm bảo tính chính xác của nhiệm vụ phân loại. Ở đây, quá trình gán nhãn đã được thực hiện thông qua việc chủ động học theo cách lặp đi lặp lại.



Sau khi loại bỏ dữ liệu bị hỏng nặng do ảnh hưởng của mây, hình ảnh viễn thám đa phổ từ các ngày khác nhau (tức là hình ảnh đa thời gian) và hình ảnh từ các bộ cảm biến khác nhau (tức là hình ảnh đa nguồn) đã được khai thác để phát hiện sự cố tràn dầu bằng cách sử dụng thuật toán học máy. Ở đây, chúng mình đã sử dụng các bộ phân loại phổ biến như máy vector cổng sup (SVM), mạng nơ-ron truyền ngược và bộ phân loại k-nearest. Trong các thí nghiệm của chúng mình, SVM đưa ra sự phân loại tốt nhất và cũng mạnh mẽ nhất. Bản đồ phân loại thu được bởi SVM trong vấn đề tràn dầu ở Vịnh Mexico được đưa ra trong Hình 5, cho thấy sự cố tràn dầu lan rộng xung quanh vị trí giàn khoan dầu.

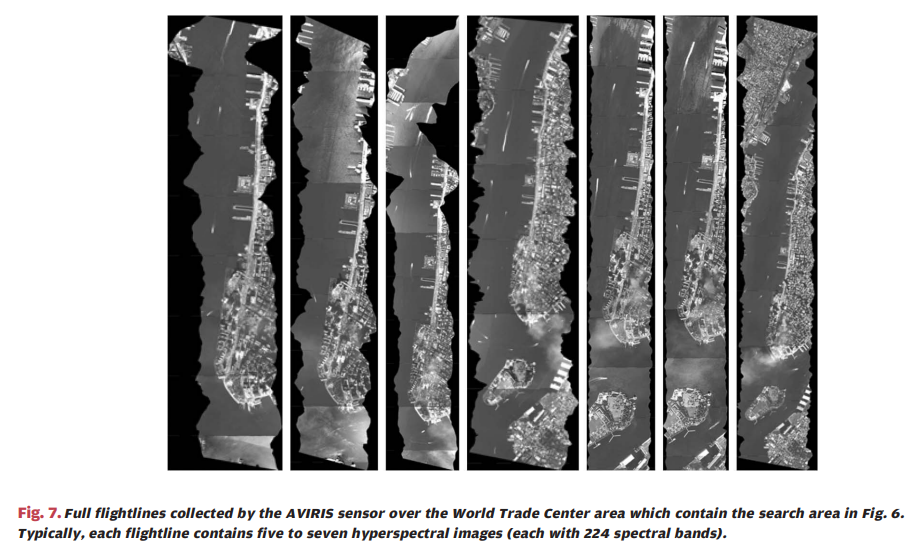
Vẫn còn nhiều vấn đề mở trong việc đánh dấu khi kết hợp hình ảnh viễn thám và dữ liệu truyền thông xã hội. Ví dụ, một chiến lược hiệu quả nên được phát triển để có được dữ liệu bên ngoài liên quan nhất cho một nhiệm vụ cụ thể. Trong lúc đó, những dữ liệu bên ngoài như hình ảnh và thông tin văn bản nên được tự động liên kết với các mẫu tương ứng.

## Truy xuất hình ảnh dựa trên nội dung từ các kho lưu dữ liệu siêu phổ



Trong trường hợp thứ hai này, chúng mình đề cập đến một trường hợp cụ thể về truy xuất hình ảnh dựa trên nội dung (CBIR) được áp dụng cho dữ liệu siêu phổ được cảm nhận từ xa, được đánh giá bởi sự nhiều chiều trong miền phổ [55]. Hệ thống được xác nhận bằng cách sử dụng một cơ sở dữ liệu hình ảnh siêu quang phổ phức tạp và được triển khai trên một trạm Beowulf tại Trung tâm vũ trụ Goddard của NASA. Trong bối cảnh này, thách thức chính của trường hợp này là để giải quyết thách thức lớn về dữ liệu viễn thám lớn, trong đó thí nghiệm của chúng mình bao gồm bộ 154 dữ liệu siêu phổ có độ phân giải cao (hơn 20 TB dữ liệu) được NASA thu thập trên khu vực Trung tâm Thương mại Thế giới (WTC) ở thành phố New York trong hai tuần cuối của tháng 9 năm 2001, chỉ vài ngày sau các cuộc tấn công khủng bố làm sụp đổ hai tòa tháp chính và các tòa nhà khác trong khu phức hợp WTC. Độ phân giải không gian của dữ liệu là 3,7 m/pixel, và độ phân giải quang phổ là 224 dải quang phổ hẹp giữa Hình 6 cho thấy một hỗn hợp màu giả của những hình này, với 614 × 512 pixel và 224 dải phổ. Thành phần màu giả đã được tổ hợp và sử dụng các kênh 1682-, 1107- và 655-nm, được hiển thị lần lượt dưới dạng màu đỏ, màu xanh lá cây và màu xanh dương. Các khu vực thực vật có màu xanh lá cây trong hình 6, trong khi các khu vực bị cháy xuất hiện màu xám sẫm. Khói đến từ khu vực WTC xuất hiện màu xanh nhạt do sự phản xạ quang phổ mạnh trong kênh 655-nm. Khu vực được sử dụng làm truy vấn đầu vào được hiển thị trong hình chữ nhật màu đỏ và nằm ở giữa khu vực nơi các tòa tháp sụp đổ.

Sử dụng khu vực tìm kiếm ở phần ngoài cùng bên phải của Hình 6 làm truy vấn đầu vào, hệ thống CBIR đã truy xuất thành công tất cả các trường hợp hình ảnh có chứa khu phức hợp WTC trên cơ sở dữ liệu. Để minh họa, Hình 7 hiển thị bảy đường bay đầy đủ trong cơ sở dữ liệu AVIRIS được coi là có chứa khu vực ta muốn tìm ở giữa khu WTC.



Để nghiên cứu các thuộc tính song song của hệ thống CBIR, chúng mình đã đánh giá hiệu suất của nó khi được triển khai trên trạm Thunderhead Beowulf của NASA, một hệ thống bao gồm 256 node Intel Xeon 2,4 GHz kép, mỗi node có 1 GB bộ nhớ và 80 GB bộ nhớ chính, được kết nối với nhau bằng sợi quang Myrinet 2GHz. Sử dụng 256 bộ xử lý trên Thunderhead, hệ thống có thể tìm kiếm hầu hết các cảnh tương tự trên toàn bộ 154 hình ảnh trong bộ cơ sở dữ liệu đầy đủ (với các metadata được tính toán trước) chỉ trong 4 giây, kết quả trong tổng cộng khoảng 10 giây để ghi lại và mô tả đầy đủ một mục nhập mới trong cơ sở dữ liệu. Điều này thể hiện sự cải thiện đáng kể so với việc thực hiện cùng một quy trình CBIR trên một bộ xử lý Thunderhead, cũng cùng 1 quy trình đó nhưng lại mất hơn 1 giờ tính toán.

# VI. KẾT LUẬN

Trong bài báo này, các ý nghĩa của dữ liệu viễn thám lớn đã được thảo luận. Dữ liệu lớn trong viễn thám có thể chứa nhiều loại dữ liệu viễn thám từ các dải phổ phản xạ khác nhau, độ phân giải không gian mặt đất khác nhau và các vị trí khác nhau (như quang học, radar, vi sóng, v.v.), cũng như dữ liệu từ các lĩnh vực khác, chẳng hạn như khảo cổ học, nhân khẩu học, kinh tế học (liên quan đến "sự đa dạng" của thuộc tính 3V của dữ liệu lớn). Kết quả là dữ liệu viễn thám lớn cũng có 3V như dữ liệu lớn nói chung với khối lượng dữ liệu viễn thám được gia tăng liên tục từ TB (terabyte) đến PB (petabyte) và thậm chí đến thang EB (exabyte). Với kích cỡ dữ liệu, một mặt, các nhiệm vụ khó lại thể hoàn thành trong một thời gian hợp lý (liên quan đến “tốc độ” của thuộc tính 3V và cho thấy được các tiềm năng lớn); mặt khác, dữ liệu cảm biến từ xa lớn với bất kỳ 2V hoặc 3V nào lại mang lại những thách thức lớn cho những người sở hữu chúng, trong việc phân tích và sử dụng dữ liệu lớn tương ứng.

Tiếp đó, một trinity (3 khía cạnh) đã được đề xuất cho những người sở hữu dữ liệu lớn, những người có thể cung cấp các phương thức dữ liệu và những người cần khai thác dữ liệu lớn để giải quyết các vấn đề trong thế giới thực. Các thách thức chung và riêng của dữ liệu lớn đã được thảo luận trong bối cảnh các ứng dụng viễn thám. Với vai trò là một thách thức chính thường gặp, dữ liệu viễn thám lớn trước tiên cần được xác định để sử dụng trong các ứng dụng viễn thám thực tế. Sau đó, việc cần thiết là có khả năng điện toán hiệu quả cao đặt để giải quyết dữ liệu kích cỡ lớn. Các phương thức dữ liệu mới hoặc thậm chí các phương thức dữ liệu hoàn toàn mới nên được phát triển để thích ứng được với tính linh hoạt của dữ liệu viễn thám lớn. Tất nhiên, vẫn tồn tại những thách thức thường gặp khác khi xử lý dữ liệu viễn thám lớn, chẳng hạn như cách phân tích chất lượng dữ liệu từ các góc nhìn khác nhau của các nguồn cung cấp dữ liệu như là cá nhân và người nhận dữ liệu.

Ngoài các thách thức thường gặp chính của dữ liệu lớn trong viễn thám, những thách thức riêng lẻ cũng nên được giải quyết theo 3 góc nhìn của trinity trong dữ liệu lớn. Với những người sở hữu dữ liệu viễn thám lớn, 3 yếu tố trọng tâm nên được làm 1 cách cẩn thận, ví dụ như truyền dữ liệu từ hệ thống bộ cảm trên không trung/không gian đến trạm dưới mặt đất, sau đó là chứa dữ liệu trong hệ thống, rồi chuyển tiếp dữ liệu đến các người dùng có hứng thú. Với những người khai thác dữ liệu lớn trong các ứng dụng viễn thám, thì những thách thức chính là việc xác định dữ liệu phù hợp để đạt được nhiệm vụ nhất định, việc triển khai dữ liệu lớn để xử lý và phân tích dữ liệu sau này cũng như diễn giải các kết quả được cung cấp bởi các phương pháp dữ liệu. Với những người có khả năng phát triển các phương thức dữ liệu mới cho các ứng dụng viễn thám, trước tiên cần quản lý việc biểu diễn dữ liệu do tính đa dạng của dữ liệu viễn thám đa nguồn và đa thời gian cũng như những dữ liệu từ các miền khác. Sau đó, dữ liệu được mô tả ở các định dạng thuộc tính khác nhau sẽ được tích hợp để phân tích và xử lý dữ liệu lớn tốt hơn. Tiếp theo, các kết quả được cung cấp bởi các phân tích dữ liệu kỹ thuật cần được cấu hình rõ ràng để cải thiện việc phân tích dữ liệu một cách thuận lợi.

Mặc dù tiềm năng của dữ liệu viễn thám lớn đã được dự đoán trước, điều quan trọng cần lưu ý là dữ liệu thường đến từ các nguồn không đồng nhất và đòi hỏi những nỗ lực tính toán đáng kể về mặt diễn dịch. Do đó, cơ hội lớn là tích hợp dữ liệu viễn thám cùng với các dữ liệu bên ngoài khác để biến những tiềm năng này thành hiện thực. Trong bối cảnh này, chúng ta có thể hưởng lợi từ dữ liệu viễn thám lớn, nhất quán, lặp đi lặp lại và chất lượng cao để giải quyết các ứng dụng liên quan đến giám sát an ninh lương thực, tiến trình đô thị hóa, mật độ dân số, v.v. Điều này có thể được sử dụng để giải quyết các ứng dụng liên quan khác liên quan đến sức khỏe con người, thay đổi môi trường hoặc các hoạt động của con người nói chung.

Để có được lợi ích từ viễn thám trong dữ liệu lớn, dữ liệu thích hợp từ các nguồn khác nhau trước tiên cần được xác định để sử dụng trong một ứng dụng cụ thể. Ngoài các dữ liệu viễn thám đa thời gian, đa độ phân giải, đa bức xạ, ta cũng cần biết làm thế nào để xác định dữ liệu ngoài miền liên quan và làm thế nào để có được những bộ dữ liệu đó là 1 thách thức lớn nhất đối với ứng dụng dữ liệu viễn thám lớn. Sau đó, một phương thức dữ liệu mới nên được thiết kế cẩn thận để xử lý dữ liệu, hợp nhất dữ liệu, v.v. Mặc dù có nhiều ứng dụng kết hợp dữ liệu viễn thám và dữ liệu đến từ các lĩnh vực khác nhau, hầu hết các công việc thực hiện được đều dựa trên một kỹ thuật lấy mẫu để ước tính, kể cả các công việc gần đây ví dụ như ước tính mật độ cây toàn cầu dựa trên 429775 phép đo mật độ cây từ các nguồn dưới mặt đất từ tất cả các châu lục trên Trái Đất. Làm cách nào chúng ta có thể sử dụng tất cả dữ liệu có sẵn để nghiên cứu sau này cho các tác vụ dữ liệu viễn thám lớn. Cuối cùng nhưng không kém phần quan trọng, làm thế nào để đánh giá hiệu suất và làm thế nào để đảm bảo chất lượng dữ liệu là những điều thú vị cần được khám phá thêm sau này.