

Khoa học: Hai, ba hay bốn chân?

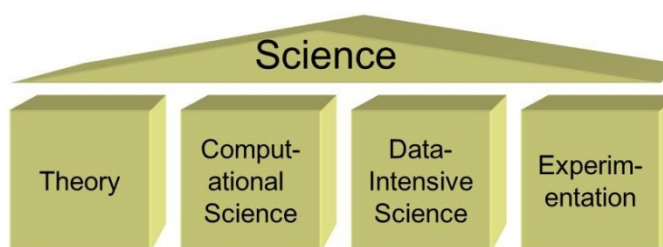
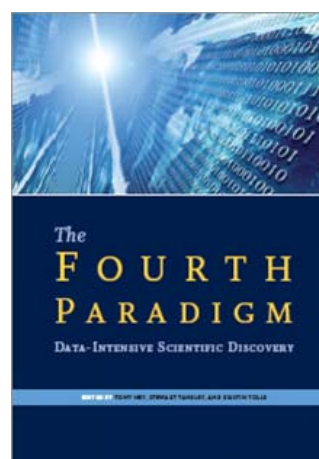
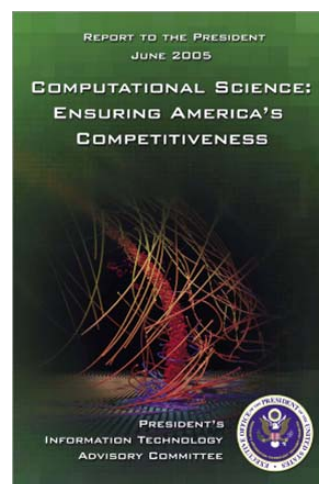
Câu chuyện về khoa học tính toán và khoa học với dữ liệu lớn

1. Hai, ba hay bốn chân?

Tháng sáu năm 2005, Hội đồng Cố vấn về Công nghệ Thông tin (PITAC) của Tổng thống Mỹ đã trình lên Tổng thống George W. Bush một báo cáo dài 117 trang với tựa đề “Khoa học Tính toán: Đảm bảo tính cạnh tranh của nước Mỹ”¹. Báo cáo bắt đầu bằng: “Cùng với lý thuyết và thực nghiệm, *khoa học tính toán* hiện đang tạo ra ‘trụ cột thứ ba’ (pillar) của khoa học, cho phép các nhà nghiên cứu xây dựng và thử nghiệm mô hình của các hiện tượng phức tạp”. Sinh động hơn, giới chuyên môn và báo chí thường gọi khoa học tính toán là cái ‘chân thứ ba’ của khoa học (the third leg), hoặc đôi khi là ‘dạng thức thứ ba’ (paradigm) của khoa học.

Báo cáo của PITAC đưa ra một lời cảnh báo cho vị trí dẫn đầu của Mỹ và cho rằng sức cạnh tranh của nước Mỹ sẽ phụ thuộc vào sự phát triển khoa học tính toán, đồng thời đề nghị một cấu trúc của nghiên cứu và giáo dục cho thế kỷ 21, một lộ trình nhiều chục năm của khoa học tính toán, một hạ tầng cơ sở bền vững cho sáng tạo và cạnh tranh, cũng như xác định các thách thức của nghiên cứu và phát triển. Rất nhiều người chia sẻ vai trò của khoa học tính toán với 24 chuyên gia công nghệ thông tin làm việc trong công nghiệp (như Chủ tịch của AT&T, Phó Chủ tịch của Dell, ...) hoặc ở các đại học MIT, Berkeley, Columbia, ... trong Hội đồng Cố vấn PITAC này². Tuy nhiên cũng có những ý kiến không đồng ý khoa học tính toán là dạng thức thứ ba của khoa học, như bài viết của Moshe Y. Vardi, Tổng biên tập tạp chí Communication of the ACM của Hiệp hội Máy tính Mỹ³. Các tranh luận như nhiều hơn khi xuất hiện đề nghị về một dạng thức thứ tư: *khoa học với dữ liệu lớn* (data-intensive science)⁴. Cuốn sách về cái ‘chân thứ tư’ này xuất bản năm 2009, để tưởng nhớ Jim Gray⁵, người đi tiên phong và đề xuất dạng thức khoa học này.

Thay vì đưa ra ý kiến riêng về khoa học có hai, ba hay bốn chân, bài này giới thiệu về khoa học tính toán và khoa học với dữ liệu lớn, và dành để mỗi người tự lựa chọn ý kiến của mình.



Science Has Only Two Legs

Science has been growing new legs of late. The traditional “legs” (or “pillars”) of the scientific method were *theory* and *experimentation*. That was then. In 2005, for example, the U.S.

CACM, Sep. 2010

DOI:10.1145/1859204.1859206

Science Has Four Legs

CACM, Dec. 2010

2. Câu chuyện khoa học tính toán

Khoa học tính toán được hiểu là khoa học về việc xây dựng các mô hình tính toán và thực hiện chúng trên máy tính để giải quyết các vấn đề phức tạp trong các khoa học khác⁶.

Báo cáo của PITAC đưa ra một định nghĩa khá chi tiết cho lĩnh vực đang phát triển nhanh này, cấu thành từ ba thành phần: (1) Mô hình hóa và mô phỏng, (2) Khoa học máy tính và thông tin, (3) Hạ tầng cơ sở cho tính toán.

Trước hết ta nói về mô hình hóa và mô phỏng. Mỗi *mô hình* là một mô tả khái quát của một thực thể, và *mô hình hóa* là quá trình tìm ra mô hình.

Thách thức của mô hình hóa các thực thể phức tạp là việc tìm ra và mô tả được những gì bản chất nhất và bỏ đi những gì không bản chất. Mô hình trong khoa học tính toán là những mô hình có thể tính toán được trên máy tính, tức chúng cần được mô tả bằng ngôn ngữ toán học (như các hệ phương trình) hay các ngôn ngữ tính toán hình thức ... Nhiều câu hỏi lý thú và quan trọng có thể đặt ra, như mô hình tính toán thế nào cho giao thông tại Hà Nội và thành phố Hồ Chí Minh? Mô hình thế nào thị trường và giá cả sau Tết? Mô hình thế nào diễn biến của một dịch bệnh trên cả nước do một virus gây ra?

Mô phỏng là việc tạo ra như thật trên máy tính các thực thể sao cho có thể thấy chúng xảy ra thế nào⁷. Mô phỏng và mô hình luôn đi với nhau. Một cách nôm na, muốn mô phỏng một thực thể trên máy tính thì phải có mô hình tính toán của thực thể đó và thực hiện được mô phỏng dựa trên việc tính toán của mô hình này. Có rất nhiều thách thức trong mô phỏng cho những thực thể phức tạp, như làm sao xây dựng được mô hình tính toán đúng cho những thứ rất lớn hoặc rất bé con người không quan sát được hay làm sao thực hiện được những lượng tính toán vô cùng lớn của mô phỏng. Thí dụ như việc xây dựng các hệ tác tử mô phỏng toàn bộ thị trường thế giới với những kịch bản khác nhau, mô phỏng sự hình thành của vụ trụ, mô phỏng toàn bộ hành vi phức tạp của các tế bào trong cơ thể người, hoặc mô phỏng để tính được với một bệnh cụ thể cần đưa thuốc gì vào chỗ nào của protein nào trong cơ thể người để chữa chạy ...

Khoa học tính toán là một lĩnh vực liên ngành, nhằm dùng các khả năng tính toán tiên tiến để hiểu và giải quyết các bài toán phức tạp. Khoa học tính toán hòa trộn ba thành phần sau:

1. Các thuật toán (số và phi-số) và phần mềm được phát triển cho *mô hình hóa* và *mô phỏng* để giải các bài toán về khoa học, kỹ nghệ và con người.
2. *Khoa học máy tính và thông tin* nhằm phát triển và tối ưu các hệ thống tiên tiến về phần cứng, phần mềm, mạng máy tính và các thành phần quản trị dữ liệu cần thiết để giải các vấn đề cần đến tính toán.
3. *Hạ tầng cơ sở tính toán* cần cho việc giải các bài toán khoa học và kỹ thuật cũng như để phát triển khoa học máy tính và thông tin.

Định nghĩa của khoa học tính toán theo PITAC



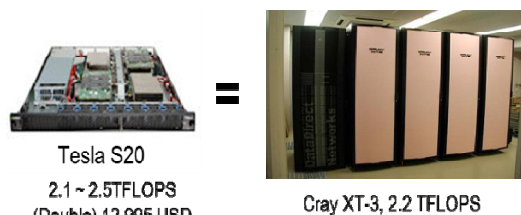
Vai trò của *khoa học máy tính và thông tin*, thành phần thứ hai của khoa học tính toán, là đương nhiên vì khoa học không chỉ cần mô hình và mô phỏng, mà còn cần nhiều thứ khác từ máy tính, nhất là các phương pháp và phần mềm quản lý và khai phá dữ liệu (sẽ nói ở phần khoa học với dữ liệu lớn).

Thành phần thứ ba *hạ tầng cơ sở cho tính toán* liên quan đến việc phát triển các máy tính hiệu năng cao gồm cả các siêu máy tính cho phép tính rất nhanh những khối lượng tính toán cực lớn. Có thể nói hầu hết các vấn đề khoa học lớn hoặc các bài toán ở tầm cỡ quốc gia đều đòi hỏi lượng tính toán cực lớn này.

Ta bắt đầu bằng một sự kiện rất đáng chú ý trong lĩnh vực tính toán hiệu năng cao: *Tính toán vạn năng trên các bộ xử lý đồ họa*, viết tắt là GPGPU.

Đây là sự kết hợp giữa ‘xử lý vạn năng’ (GP, general purpose) và ‘đơn vị xử lý đồ họa’ (GPU, graphical processing unit). GPU là bộ xử lý làm các nhiệm vụ tính toán đồ họa thường được dùng

trong các hệ nhúng, điện thoại di động, máy tính cá nhân, trạm làm việc, và các màn hình game. Chuyên dụng cho nhiệm vụ xử lý đồ họa, các GPU với cấu trúc tính toán song song thích hợp đã thực hiện với hiệu năng rất cao một lớp bài toán phức tạp. Trong những năm gần đây, GPU đã được phát triển và tiến hóa thành những bộ xử lý hết sức linh hoạt và mạnh về khả năng lập trình, độ chính xác và hiệu quả. GPGPU chính là việc mở rộng GPU cho các nhiệm vụ tính toán khác. Khó khăn lớn của việc lập trình trên GPGPU được vượt qua gần đây bởi kiến trúc CUDA của hãng NVIDIA, cho phép xây dựng các ứng dụng trên các ngôn ngữ thông dụng nhất như C, và cả với Python, Fortran, Java, hay Matlab. Điều đáng để ta quan tâm nhất là giá thành rất thấp của các bộ GPGPU. Chẳng hạn một bộ GPGPU Tesla S2050 của hãng NVIDIA với giá 12,995 USD khi lắp vào một máy tính thông thường có thể đạt tốc độ tính toán 2,1-2,5 teraflops (triệu triệu phép tính mỗi giây)⁸. Chú ý rằng đây là tốc độ tính toán của siêu máy tính CRAY XT-3 (2,2 teraflops) phiên bản năm 2004 với giá khoảng 2 triệu USD.



Máy tính với bộ xử lý Tesla S2050 có hiệu năng như CRAY XT-3

Từ khi chiếc máy tính lớn đầu tiên trên thế giới ENIAC ra đời tại Mỹ vào năm 1946 đã luôn có một cuộc thi đua của các cường quốc về sản xuất máy tính, đặc biệt là các siêu máy tính.

Thập kỷ 80 của thế kỷ trước đã sôi động với đề án máy tính thế hệ 5 của Nhật Bản, dù sau được xem là không thành công⁹. Câu chuyện siêu máy tính như lắng xuống trong hai thập kỷ tiếp theo khi kỹ thuật vi tính và các máy tính cá nhân, internet và các tiến bộ của truyền thông ra đời và bùng nổ, đến với từng người trong toàn xã hội. Tuy nhiên, những bài toán tầm quốc gia hay những bài toán lớn của khoa học và công nghệ cần phải giải quyết trên các siêu máy tính, và cuộc thi đua sản xuất siêu máy tính vẫn ‘âm thầm’ diễn ra. Gần đây rộ lên tin tức về các dự án siêu máy tính. Tháng 10 năm 2010, Trung Quốc công bố hoàn



Máy tính Tianhe-1A của Trung Quốc

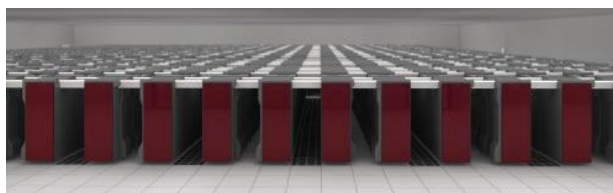
thành máy tính Thiên Hà 1A (Tianhe-1A) đạt tốc độ 2,507 petaflops. Vào thời điểm này đây là máy tính được công nhận có tốc độ nhanh nhất thế giới. Chú ý là Thiên Hà-1A hoàn toàn dựa trên các bộ xử lý của Mỹ, kết nối 7.168 bộ Tesla M2050 GPUs của NVIDIA và 14.336 bộ CPUs của hãng Intel.

Chính phủ Mỹ đặt hàng hãng IBM sản xuất hai siêu máy tính, đặt tên BlueGene và BlueWater nhằm đạt tốc độ 20 petaflops dự kiến hoàn thành vào đầu năm 2012. Để có ý niệm về 20 petaflops nhanh thế nào, hãy hình dung nếu toàn bộ 6,7 tỷ người trên trái đất với máy tính cầm tay liên tục cùng làm các phép tính 24 giờ mỗi ngày, 365 ngày mỗi năm thì 6,7 tỷ người này phải làm việc 320 năm để đạt được số phép tính BlueGene làm trong một tiếng đồng hồ. Để biết một petabyte dữ liệu lớn thế nào, hãy hình dung 10 petabytes có thể chứa hết toàn bộ gần 22 triệu cuốn sách tại thư viện lớn nhất thế giới ở New York (the U.S. Library of Congress) khi chúng được số hóa (và 22 triệu cuốn sách nếu xếp đứng kề nhau theo gáy sách trên quốc lộ 1 sẽ dài từ Hà Nội đến Huế, tức dài quãng 630 km).



Máy tính IBM BlueGene của Mỹ

Nước Nhật có đề án trọng điểm quốc gia trong giai đoạn 2007-2012 với đầu tư hơn một tỷ USD nhằm xây dựng siêu máy tính với tốc độ 10 petaflops. Đề án được lãnh đạo bởi Bộ Giáo dục, Văn hóa, Thể thao, Khoa học và Công nghệ (MEXT), thực hiện bởi công ty Fujitsu và Viện RIKEN. Máy tính gồm 800 tủ máy, với các CPU siêu nhanh của công ty Fujitsu, có tên 'K computer'. Tên gọi này có hai nghĩa. Chữ 'kei' trong tiếng Nhật có nghĩa 10^{16} , nhằm chỉ mục tiêu 10 petaflops. Chữ 'kei' trong tiếng Nhật còn có nghĩa cửa ngõ lớn, nhằm chỉ hy vọng máy tính này sẽ là cửa ngõ lớn cho khoa học tính toán. 'K computer' được dự kiến đặt tại Viện Tiên tiến về Khoa học Tính toán của RIKEN ở thành phố Kobe.



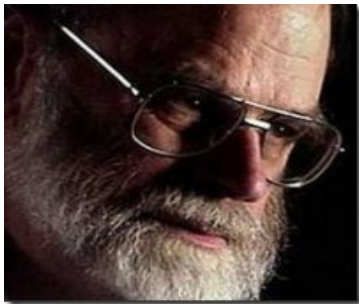
'K computer' của Nhật Bản

Giới công nghiệp dự tính sẽ sản xuất được máy tính đạt tốc độ 1 exaflops (1 exa = 1000 peta) vào năm 2020. Đáng nói là cuộc thi đua về siêu máy tính không chỉ có việc làm ra máy tính, mà khó không kém và thậm chí nhiều thách thức hơn là việc làm ra những ứng dụng trên các siêu máy tính này. Các nhà khoa học Mỹ đang thực hiện nhiều dự án tính toán về khí hậu, về vũ trụ, về y học và sinh học, ... trên BlueGene. Ở Nhật Bản, 21 đề tài ứng dụng 'K computer' cũng đã được xây dựng từ đầu đề án quốc gia, gồm 6 về khoa học nano, 6 về khoa học sự sống, 3 về khí hậu và địa cầu, 4 về kỹ nghệ, và 2 về vật lý và thiên văn.

Làm được siêu máy tính Thiên Hà-1A, Trung Quốc chứng tỏ một tầm nhìn, một kế hoạch được đầu tư ít nhất từ hơn mười năm trước và một tiến bộ vượt bậc, nhất là trong việc kết nối và truyền tin giữa các máy nhỏ thành phần. Tuy nhiên có ít nhất hai hạn chế đáng lưu ý. Thứ

nhất là Thiên Hà-1A dùng toàn bộ các bộ xử lý của Mỹ, và điều này tạo ra sự phụ thuộc lớn vào Mỹ. Thứ hai là các ứng dụng trên siêu máy tính ở Trung Quốc còn hạn chế. Nie Hua, phó chủ tịch công ty Dawning Information Industry tại Bắc Kinh (nơi làm ra siêu máy tính Nebulae hiện được xếp nhanh thứ ba thế giới), nêu ý kiến: “Do chưa phát triển được các ứng dụng, siêu máy tính sẽ không tạo được nhiều tác động vào nền kinh tế và cuộc sống thường nhật của chúng tôi. Thậm chí nếu phát triển được các ứng dụng ở Trung Quốc, chúng tôi vẫn đi sau Mỹ 5 năm vì không biết cách sử dụng những ứng dụng đó”¹⁰.

3. Câu chuyện về khoa học với dữ liệu lớn



Jim Gray (1944-2007)

Jim Gray nhận giải Turing năm 1988 về “đóng góp có tính dẫn đường trong nghiên cứu về cơ sở dữ liệu và xử lý giao dịch, cũng như vai trò lãnh đạo kỹ thuật của việc cài đặt hệ thống”¹¹.

Hơn mười năm cuối tại Microsoft, Gray thường làm việc với với các nhà khoa học nghiên cứu về đại dương, về địa cầu, về thiên văn, và thấy trong nhiều lĩnh vực khoa học ngày nay con người không còn quan sát các hiện tượng bằng mắt được nữa, mà chính máy móc ‘quan sát’, ‘đo đạc’ và chuyển kết quả là dữ liệu lên máy tính. Gray thấy rõ các nhà khoa học đang thu thập và tích tụ những lượng dữ liệu lớn ở mọi thang bậc, mọi khuôn dạng, đang ‘ngập chìm’ trong đồng dữ liệu khổng lồ vì không biết dùng và khai thác chúng thế nào. Từ hiện thực này và với một tầm nhìn sâu sắc, Gray đã nhiều năm đề cập đến một dạng thức mới của khoa học, tức quá trình khám phá khoa học với dữ liệu lớn (data-intensive scientific discovery). Trong bài nói cuối tại Hội đồng Công nghệ Thông tin và Truyền thông ngày 11/1/2007 (trước khi ra khơi và mất tích trên biển ngày 28/1/2007), Gray đã mô tả lần nữa quan điểm của mình về dạng thức thứ tư này của nghiên cứu khoa học¹².

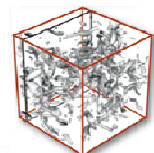
Ba dạng thức khoa học đã tồn tại trong lịch sử thường được Gray nhắc lại trong các bài nói. Trước hết, khoa học từ hàng nghìn năm trước đã dựa vào *thực nghiệm*, chủ yếu nhằm mô tả các hiện tượng thiên nhiên. Tiếp theo, trong vài năm trăm vừa qua, các *lý thuyết* đã trở thành một dạng thức thứ hai của khoa học với bản chất là việc khái quát hóa các hiện tượng thành mô hình trong các lĩnh vực của khoa học, như các định luật Kepler, các định luật về chuyển động

Các dạng thức khoa học

- Hàng nghìn năm trước:
khoa học là **thực nghiệm**
Mô tả các hiện tượng thiên nhiên
- Vài trăm năm vừa qua:
thêm nhánh **lý thuyết**
Dùng các mô hình, các khái quát hóa
- Vài thập kỷ vừa qua:
thêm nhánh **tính toán**
Mô phỏng các hiện tượng phức tạp
- Ngày nay: **Khai thác dữ liệu (eScience)**
Hợp nhất lý thuyết, thực nghiệm và mô phỏng
 - Dữ liệu từ đo đạc bằng máy hoặc mô phỏng
 - Được xử lý bởi các phần mềm
 - Thông tin và tri thức được chứa trong máy tính
 - Nhà khoa học phân tích cơ sở/tập dữ liệu với công cụ quản trị dữ liệu và thống kê.



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



Bốn dạng thức của nghiên cứu khoa học (Jim Gray, 11.1.2007)

của Newton, các phương trình Maxwell, ... Trong quá trình phát triển, nhiều mô hình lý thuyết trở nên quá phức tạp không thể phân tích được dẫn con người đến việc phải *mô phỏng*, cốt lõi của dạng thức thứ ba của khoa học. Các kỹ thuật mô phỏng đã phát triển rất sôi động trong vòng mấy chục năm qua cùng sự phát triển của khoa học máy tính. Mô phỏng cũng đã tạo ra những lượng dữ liệu khổng lồ, bên cạnh các dữ liệu thu được từ thực nghiệm.

Quan sát sự thay đổi của thế giới khoa học, của lượng dữ liệu đang tăng lên gấp đôi hằng năm, Gray cho rằng khoa học ngày nay cần những mô hình mới để thu được và quản lý được dữ liệu trước khi phân tích chúng nhằm tìm ra các thông tin và tri thức cần thiết. Gray khẳng định cần xem *nguyên cứu với nguồn dữ liệu lớn* là một con đường mới của khoa học.

Lấy vài thí dụ để thấy lượng dữ liệu khoa học quanh ta lớn thế nào. Chẳng hạn đề tài Khảo sát Bầu trời Số (Sloan Digital Sky Survey) tài trợ bởi Quỹ Sloan – với sự tham gia của nhiều đại học và viện nghiên cứu ở Mỹ, Nhật và Đức, khảo sát khoảng một triệu đối tượng trên bầu trời – thu được 40 terabytes dữ liệu thô và 2,5 petapixels ảnh. Máy va chạm hạt Hadron lớn của CERN (Tổ chức châu Âu về Nghiên cứu Hạt nhân) có thể tạo ra vài petabytes dữ liệu mỗi ngày. Khi sắp các dãy nucleotides trong hệ gene người, khoảng 25 kilobytes có thể chứa vài trăm nghìn cặp cơ sở, do vậy mỗi hệ gene người với ba tỷ cặp cơ sở sẽ chứa trong khoảng 1 gigabyte, và toàn bộ hệ gene của 6,7 tỷ người trên trái đất sẽ chứa trong khoảng 7000 petabytes (tức 7 exabytes). Đây là một lượng rất lớn nếu biết là tất cả dữ liệu trên toàn thế giới có trong vòng một năm được chứa trong 2 exabytes.

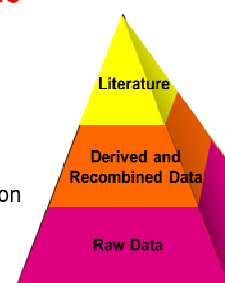
Nguồn dữ liệu khoa học còn gồm số lượng khổng lồ các ấn phẩm khoa học và sáng chế phát minh đã được công bố, và rất nhiều đã được số hóa và tổ chức trực tuyến (online) trên internet. Cơ sở dữ liệu điện tử các ấn phẩm y học MEDLINE chứa hơn 18 triệu tài liệu của khoảng 5000 tạp chí y học và sinh học từ năm 1950 đến nay.

Việc khai thác dữ liệu khoa học đòi hỏi hợp nhất cả lý thuyết, thực nghiệm và mô phỏng. Gray tóm tắt dạng thức khám phá khoa học này như sau: Dữ liệu thô thu nhận từ đo đạc bằng máy móc hoặc mô phỏng được xử lý bởi các phần mềm và các dữ liệu, thông tin, tri thức chứa trong các cơ sở dữ liệu sẽ được nhà khoa học phân tích với các công cụ quản trị, khai phá dữ liệu và thống kê.

Điều mấu chốt trong đề xuất của Gray cho tương lai là cộng đồng khoa học thế giới cần hướng đến việc chia sẻ cùng nhau mọi dữ liệu khoa học (dữ liệu thô, dữ liệu đã xử lý và tích hợp, tài liệu khoa học) và các chương trình xử lý chúng. Những điểm chính trong bức tranh khoa học khi các tài nguyên khoa học được chia sẻ online là: Nhà khoa học được dùng dữ liệu từ nhiều lĩnh vực liên quan; Internet là môi trường cho phép hợp nhất mọi nguồn tài liệu và dữ liệu khoa học;

All Scientific Data Online

- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere
- Increase scientific information velocity
- Huge increase in science productivity



Đề xuất của Gray về dữ liệu khoa học online

Để khám phá tri thức mới nhà khoa học sẽ đi từ tài liệu đến dữ liệu và tính toán rồi trở về với tài liệu; Thông tin khoa học của mỗi người sẽ đến được bất kỳ ai khác, bất kỳ nơi nào khác; Tốc độ của thông tin khoa học sẽ tăng lên rất nhiều, và do vậy năng suất khoa học sẽ tăng lên bội phần.

Phải chăng đây là một giấc mơ về một ‘thế giới đại đồng’ của khoa học? Nhưng lẽ nào ta không có quyền tưởng tượng và mơ ước cùng Gray. Chia sẻ đề xuất này Bill Gate viết: “Ảnh hưởng của suy nghĩ của Jim Gray sẽ tiếp tục làm con người nghĩ theo một cách mới về dữ liệu và phần mềm đang được nhìn nhận lại ra sao cho sự phát triển của khoa học”.

Jim Gray không phải người duy nhất nói về khoa học với dữ liệu lớn. Liên quan điều này, hơn mười năm trước một khoa học mới đã ra đời: khoa học về *phát hiện tri thức và khai phá dữ liệu* (knowledge discovery and data mining), viết tắt là KDD. Với tiến bộ của xã hội, của khoa học và công nghệ, của máy tính và mạng máy tính, quanh ta chưa bao giờ nhiều dữ liệu như ngày nay. Lượng dữ liệu rất lớn với nhiều triệu bản ghi của nhiều kiểu dữ liệu phức tạp không ở dạng vectơ thông thường (như dữ liệu về các hợp chất hóa học hoặc các mạng trao đổi chất hoặc các liên kết trên web dưới dạng đồ thị; hoặc các dãy axit amino trong dữ liệu protein hay các dãy nucleotide biểu diễn DNA, ...) đã vượt quá khả năng phân tích của các phương pháp và phần mềm con người đã có, và thúc đẩy sự ra đời của KDD.

Hội nghị quốc tế đầu tiên về KDD được tổ chức vào năm 1995 tại Montreal và từ đó đến nay ngành khoa học này đã phát triển mạnh mẽ, đi được những bước tiến dài. Rất gần với KDD là ngành *máy học* (machine learning), với mục tiêu làm cho máy tính có một số khả năng học tập như con người và cốt lõi ngành máy học là việc phân tích các bảng dữ liệu, đặc biệt là các dữ liệu phức tạp. Ngành máy học đã phát triển rất nhanh trong vòng hơn ba mươi năm qua, cùng với KDD tạo thành một trong những lĩnh vực sôi động của công nghệ thông tin.

Đáng kể nữa là hai lĩnh vực về các cơ sở dữ liệu rất lớn và phân tích dữ liệu nhiều biến trong thống kê toán học cũng có những phát triển sâu sắc trong vòng mấy chục năm qua, vốn được thúc đẩy bởi sự xuất hiện của rất nhiều dữ liệu. Các phương pháp và kỹ thuật của hai lĩnh vực này cũng như của KDD và máy học chính là những công cụ cơ bản để thực hiện khoa học tính toán và khoa học với dữ liệu lớn.

Đến đây mỗi người có thể tùy thích tự trả lời rằng khoa học có hai, ba hay bốn chân. Điều

Người viết và nhóm nghiên cứu làm việc trong lĩnh vực máy học và khai phá dữ liệu, và đang cùng các đồng nghiệp trong ngành vật lý, y học và công nghệ thông tin thực hiện một số đề tài nghiên cứu về khoa học tính toán:

1. Xác định cấu trúc và quá trình động học của các phân tử nước trong các hệ sinh học bằng kỹ thuật khai phá dữ liệu, với dữ liệu mô phỏng trên 6 petabytes.
2. Xác định cơ chế phân tử của bệnh viêm gan C và phương pháp điều trị dựa trên tích hợp của protein NS5A, các yếu tố ngoài hệ gene và RNAi.
3. Xây dựng tài nguyên, công cụ và phương pháp của tính toán ngôn ngữ cho tiếng Việt.

Một vài đề tài về tính toán khoa học

quan trọng hơn là ta thấy được tầm quan trọng của những chuyện này và biết chúng liên quan đến ta ra sao. Dù chưa có siêu máy tính như ở các nước có nền khoa học phát triển, hiện chúng ta cũng đang có rất nhiều cơ hội và nội dung nghiên cứu về khoa học tính toán. Các bộ xử lý GPGPU với giá rẻ cho phép ta có được dễ dàng nhiều máy tính hiệu năng cao với tốc độ nhiều teraflops. Ta có thể sớm đào tạo và tập trung được một lực lượng nghiên cứu đáng kể về mô phỏng, cơ sở dữ liệu lớn, phân tích dữ liệu nhiều biến, máy học và khai phá dữ liệu. Khi này, rất nhiều bài toán ý nghĩa có thể đặt ra và giải quyết, như nghiên cứu các bệnh nhiệt đới với nhiều dữ liệu chỉ ta có, như mô phỏng và tính toán các dòng chảy và thời tiết để dự đoán lũ lụt ở miền Trung, như tính toán và sử dụng tự động tiếng nói và văn bản tiếng Việt trên máy tính và các phương tiện truyền thông di động, ...

Và còn rất nhiều việc khác nữa trong cái mệnh mông của thế giới khoa học, mà đi vào đây ta mới tìm ra những con đường phát triển bền vững của đất nước.

Hồ Tú Bảo

Tài liệu tham khảo và ghi chú

1. President's Information Technology Advisory Committee, *Computational Science: Ensuring America's Competitiveness*, June 2005.
2. Moshe Y. Vardi, "Science has only two legs", *Communication of the ACM*, September 2010.
3. Tony Hey, "Science has four legs", *Communication of the ACM*, December 2010.
4. Tony Hey, Stewart Tansley, Kristin Tolle (Eds.), *The Four Paradigm: Data Intensive Scientific Discovery*, Microsoft 2009.
5. Jim Gray (1944-2007) là nhà khoa học máy tính có nhiều đóng góp vào nhận thức nghiên cứu khoa học với nguồn dữ liệu lớn (tạm dịch 'khoa học vớidữ liệu lớn').
6. Thuật ngữ 'khoa học tính toán' đã xuất hiện từ nhiều năm trước ở ta, như 'Viện khoa học tính toán và điều khiển' của Viện khoa học Việt Nam ra đời vào năm 1976, sau đổi tên thành 'Viện tin học' rồi 'Viện công nghệ thông tin'. Các từ có cùng nghĩa và dùng phổ biến khi đó là 'khoa học tính toán' (ở Nga) và 'tin học' (ở Pháp và một số nước châu Âu), và về cơ bản cùng nghĩa với 'khoa học máy tính' hay 'công nghệ thông tin' (theo nghĩa rộng của từ này).
7. Ở đây ta hạn chế chỉ nói về mô phỏng trên máy tính, và 'mô phỏng' được hiểu là 'mô phỏng trên máy tính'.
8. Về đơn vị đo: kilo = 10^3 , mega = 10^6 , giga = 10^9 , tera = 10^{12} , peta = 10^{15} , exa = 10^{18} .
9. Hồ Tú Bảo, "Đề án máy tính thế hệ 5 của Nhật Bản", *Tạp chí Khoa học Tính toán và Điều khiển*, 1990.
10. CNN, "Is China now the world's top supercomputer power?", Nov. 18, 2010.

11. Giải Turing là giải thưởng cao khoa học quý nhất trong ngành công nghệ thông tin và truyền thông, được xem là giải Nobel của ngành này. Giải thưởng trao mỗi năm cho một người.
12. Jim Gray đã nhiều lần trình bày quan điểm của mình và được ủng hộ rộng rãi. Người viết cũng đã nghe Gray trình bày về khoa học với dữ liệu lớn tại hội nghị về Khai phá Dữ liệu KDD ở Washington D.C., tháng 8 năm 2003.