

## TỔNG QUAN VỀ KHOA HỌC DỮ LIỆU

ThS Nguyễn Danh Minh Trí

Trường ĐHKHXH&NV - ĐHQG Tp. Hồ Chí Minh

**Tóm tắt:** Khoa học dữ liệu là một lĩnh vực nghiên cứu mới xuất hiện thời gian gần đây. Cùng với sự phát triển không ngừng của các kho dữ liệu lớn trên khắp thế giới, ngành rút trích và phân tích dữ liệu đang trở nên ngày càng quan trọng hơn. Thông tin và tri thức có được từ dữ liệu lớn đang giúp ích cho rất nhiều công ty và tổ chức trên thế giới. Bài viết giới thiệu những khái niệm cơ bản về Khoa học dữ liệu, các phương pháp được sử dụng và tương lai phát triển của lĩnh vực này.

**Từ khóa:** Khoa học dữ liệu; hồi quy; dữ liệu; khai thác dữ liệu; thống kê; mô hình thống kê.

### Overview on data science

**Abstract:** Data science is an emerging research field. With the development of Big Data storages all over the world, data retrieval and analysis is proved to play a more and more important role. Information and knowledge from Big Data is very helpful for many companies and organizations in the world. The article introduces definitions, methodologies and future of data science.

**Keywords:** Data science; regression; data utilization; statistics; statistical model.

### Mở đầu

Dữ liệu đóng một vai trò rất quan trọng trong việc vận hành, ứng dụng cũng như lưu trữ thông tin của người dùng. Ngày nay, cùng với sự phát triển không ngừng của nhân loại, lượng dữ liệu được phát sinh là vô cùng lớn. Ý nghĩa của những tri thức có được từ dữ liệu là nền tảng cực kỳ hữu ích cho các hệ thống ra quyết định và hỗ trợ cuộc sống. Ngày nay, hầu hết các công ty và tập đoàn lớn đều đã có những đội ngũ, chuyên gia phân tích dữ liệu của riêng họ. Có thể kể đến trong danh sách này là Google, Facebook, Yahoo, Youtube,... Sự thành công của các công ty và tổ chức trên thế giới ngày nay đều ít nhiều có liên quan với ngành Khoa học dữ liệu (KHDL). KHDL đang lan rộng ảnh hưởng của nó và mang lại ý nghĩa ngày càng quan trọng hơn đối với

đời sống con người.

### 1. Khái niệm về Khoa học dữ liệu

#### 1.1. Lịch sử của Khoa học dữ liệu

Trong khoảng hơn 30 năm (1960-1996), thuật ngữ "Khoa học dữ liệu" (data science) đã được sử dụng trong nhiều tài liệu nói về các phương pháp tính toán. Đến tháng 11/1997, thuật ngữ KHDL mới được dùng chính thức bởi một nhà nghiên cứu tên là Chien-Fu Jeff Wu. Trong bài thuyết trình mang tên "Statistics = Data Science?" tại Đại học Michigan, Chien-Fu Jeff Wu đã phổ biến thuật ngữ "Khoa học dữ liệu" và nói rằng thống kê nên được đổi tên thành KHDL và nhà thống kê thành nhà KHDL vì họ đã dành phần lớn thời gian của mình để thao tác và thử nghiệm với dữ liệu [4].

Năm 2001, William S. Cleveland đã

giới thiệu KHDH như là một ngành độc lập. Đến tháng 4/2002, International Council for Science cho ra đời Tạp chí KHDH, một ấn phẩm tập trung vào các vấn đề như mô tả hệ thống dữ liệu, ấn phẩm của họ trên internet, các ứng dụng và các vấn đề pháp lý. Vào tháng 01/2003, Đại học Columbia bắt đầu xuất bản Tạp chí KHDH, nhằm cung cấp một công cụ cho tất cả nhân viên dữ liệu trình bày quan điểm của mình và trao đổi ý kiến. Đến năm 2008, DJ Patil và Jeff Hammerbacher mới sử dụng thuật ngữ “nhà KHDH” để xác định công việc của họ tại LinkedIn và Facebook. [4].

Năm 2013, Nhóm công tác của IEEE về KHDH và Phân tích nâng cao đã được đưa ra, và hội nghị quốc tế đầu tiên về KHDH và Phân tích nâng cao của IEEE đã được tổ chức vào năm 2014. Năm 2015, Tạp chí Quốc tế về KHDH và Phân tích đã được lập bởi Springer để xuất bản tác phẩm ban đầu về KHDH và phân tích dữ liệu lớn. KHDH vẫn còn được xem là một khái niệm mới, và nó chỉ mới thật sự xuất hiện vào những thập niên đầu thế kỷ 21.

### **1.2. Khoa học dữ liệu và các khái niệm liên quan**

KHDH là một lĩnh vực nghiên cứu mới xuất hiện thời gian gần đây. Như tên gọi của nó, KHDH là một ngành khoa học nghiên cứu về dữ liệu. Điều này nhấn mạnh đối tượng nghiên cứu chính của ngành là dữ liệu. Dữ liệu rất đa dạng và có khối lượng, tốc độ phát sinh rất lớn do nhu cầu phát triển của các ứng dụng và cuộc sống con người. Dữ liệu có thể đến từ mọi nơi, mọi lĩnh vực trong cuộc sống chúng ta [8]. Do vậy, ở góc độ chuyên ngành, KHDH là một lĩnh vực nghiên cứu liên ngành vì nó khảo sát rất nhiều loại dữ liệu đến từ các lĩnh vực chuyên ngành khác nhau, về các quá trình và các hệ thống rút trích tri thức hoặc hiểu biết từ dữ liệu ở các dạng khác nhau (có cấu trúc hay phi cấu trúc) và nó là sự tiếp nối của một số lĩnh vực phân tích dữ liệu

như khoa học thống kê, khai thác dữ liệu, tương tự như khám phá tri thức ở các cơ sở dữ liệu (KDD) [5].

Mục tiêu chính của ngành KHDH là để có được cái nhìn sâu hơn vào dữ liệu và tạo ra những điều hữu ích cho cuộc sống con người. Thông qua quá trình tiếp nhận, phân tích các đặc tính và rút được các kết quả từ dữ liệu sẽ hỗ trợ chúng ta trong việc đưa ra các quyết định, các dự đoán tốt hơn cho các hệ thống. Quá trình nghiên cứu KHDH cần sự hỗ trợ của các phương pháp tính toán, thống kê và trực quan hóa. Theo đó, chúng ta sẽ thường xuyên sử dụng các phương pháp tính toán để xây dựng các mô hình, áp dụng cách thức phân tích để thâm nhập tốt hơn vào dữ liệu. Bên cạnh đó, chúng ta cũng áp dụng các phương pháp thống kê để hỗ trợ đưa ra các dự đoán và cuối cùng là sử dụng phương pháp trực quan hóa để thể hiện tốt hơn các kết quả nghiên cứu lên các công cụ trực quan. Trực quan hóa có vai trò cực kỳ quan trọng trong việc giao tiếp với các kết quả có được từ quá trình nghiên cứu, phân tích dữ liệu. Nhưng nếu việc trực quan hóa thể hiện không rõ ràng hoặc không chuyển tải được thông tin thì kết quả đó coi như cũng không có giá trị gì. Điều này cho thấy vai trò quan trọng trong việc thể hiện các kết quả lên các công cụ trực quan hóa.

Khoa học thống kê (hỗ trợ cho KHDH) sử dụng các kỹ thuật và lý thuyết rút từ nhiều lĩnh vực với biên độ rộng bao gồm: toán học, khoa học thống kê, khoa học thông tin và khoa học máy tính. Các lĩnh vực chuyên ngành liên quan đến KHDH bao gồm: xử lý tín hiệu, lý thuyết xác suất, học máy, lý thuyết học thống kê, khai thác dữ liệu, cơ sở dữ liệu, kỹ thuật thông tin, nhận dạng mẫu, trực quan dữ liệu, các phân tích dự đoán, lý thuyết quyết định, kho dữ liệu, nén dữ liệu, lập trình máy tính, trí tuệ nhân tạo, và siêu máy tính [5]. Điều này giải thích cho lý do vì sao KHDH có tính liên ngành và có biên độ nghiên cứu rất rộng lớn.

Những người làm việc trong lĩnh vực KHDH thường được gọi với cái tên là: “nhà KHDH”. Theo Josh Blumenstock: “Nhà KHDH là người biết thống kê nhiều hơn nhà khoa học máy tính và biết khoa học máy tính nhiều hơn nhà thống kê”. Đây là một định nghĩa vui, nó ngụ ý KHDH là nơi hội tụ chủ yếu của 2 lĩnh vực là khoa học máy tính và thống kê. Ngoài ra, cũng có một định nghĩa khác của Shlomo Aragon như sau: “Nhà KHDH = Nhà Khoa học + Lập trình viên + Huấn luyện viên + Người kể chuyện + Nghệ sĩ”. Định nghĩa này thể hiện công việc đa dạng của một nhà KHDH. Điều này cho thấy vai trò của họ là người hỗ trợ cho các khoa học chuyên ngành phát triển thông qua việc đi sâu hơn trong tương tác với dữ liệu.

Để trở thành một nhà KHDH thì chúng ta cần học thêm toán học, thống kê và học máy, học lập trình và hiểu biết về cơ sở dữ liệu. Bên cạnh đó, chúng ta cũng cần làm quen với các công nghệ dữ liệu lớn, tìm thêm các trải nghiệm từ thực tế thông qua các cuộc thi và các chương trình thực tập liên quan [7].

### **1.3. Vì sao cần nghiên cứu Khoa học dữ liệu**

Thế giới đang bước vào kỷ nguyên dữ liệu lớn. Trong thời đại dữ liệu rất lớn thì nhu cầu xử lý và hiểu sâu hơn dữ liệu cũng sẽ tăng. Dữ liệu phát sinh không ngừng và rất lớn theo thời gian thực đòi hỏi sự khám phá và tìm hiểu để tận dụng tối đa giá trị hữu ích tiềm năng có được từ dữ liệu [8]. Điều này thôi thúc người ta nghĩ đến một hướng nghiên cứu mới nhằm phân tích và có được những tri thức hữu ích từ khối dữ liệu lớn và ngày càng đa dạng này.

Dữ liệu phát sinh rất lớn và mang bốn

tính chất quan trọng, đó là khối lượng rất lớn (volume), tính đa dạng (variety), tốc độ thu thập dữ liệu (velocity) ngày càng nhanh hơn và mang tính xác thực hay còn gọi là tính chính xác (veracity) [2]. Đó cũng là bốn thách thức đến từ dữ liệu lớn đối với con người cũng như rất nhiều hệ thống khai thác và phân tích dữ liệu trên khắp thế giới. Và đây cũng chính là nhiệm vụ mà KHDH cần giải quyết trong nhiều thập kỷ tới.

Rất nhiều hệ thống rút trích tri thức có được từ dữ liệu đang được triển khai, rất nhiều tiềm năng của dữ liệu lớn chưa được khai thác hết. Vẫn còn nhiều nhiệm vụ đặt ra cho các nhà KHDH trong tương lai. Như giáo sư Hal Varian- một chuyên gia của Google đã nói: “Công việc hấp dẫn trong 10 năm tới sẽ là: Nhà KHDH”. Điều này đề cập đến vai trò ngày càng quan trọng của lĩnh vực nghiên cứu mới này. Rất nhiều lĩnh vực cần các tri thức hỗ trợ có từ dữ liệu: y học, sinh – tin học, điều khiển học, thiết bị thông minh, chống tội phạm, đưa ra quyết định kinh doanh, dự đoán... Theo nghiên cứu của công ty tư vấn quản lý toàn cầu McKinsey, đến năm 2018, Mỹ sẽ cần 140.000 đến 190.000 người có kỹ năng phân tích chuyên sâu cũng như 1,5 triệu nhà quản lý và phân tích trong lĩnh vực “dữ liệu lớn” (big data) [6]. Điều này cho thấy mức độ quan trọng và ngày càng hấp dẫn của ngành phân tích dữ liệu trong tương lai.

Khả năng hiểu, xử lý, trích xuất giá trị, trực quan hóa, giao tiếp với dữ liệu sẽ là một kỹ năng cực kỳ quan trọng trong những thập kỷ tiếp theo khi mà dữ liệu ngày càng lớn mà con người như đang chìm đắm trong biển dữ liệu nhưng lại thiếu thông tin hữu ích. KHDH sẽ phần nào giúp cho con người giải quyết các vấn đề quan trọng trên.

#### 1.4. Một số ví dụ tiêu biểu của Khoa học dữ liệu

Một ví dụ tiêu biểu là đội ngũ KHDL của các hãng công nghệ lớn như: Google, Amazon, Facebook, Microsoft, Youtube, LinkedIn, Instagram, Twitter... Họ không ngừng tham gia vào quá trình thu thập, phân tích, dự đoán các tri thức có từ dữ liệu để hỗ trợ cho doanh nghiệp của mình. Ngoài ra còn có các ứng dụng đã tạo nên ý nghĩa lớn của ngành KHDL hiện nay. Sau đây sẽ là một số ví dụ tiêu biểu nhất.

Ứng dụng KHDL tiêu biểu đầu tiên có thể kể đến là giải thưởng Netflix cho các thuật toán tốt nhất để dự đoán xếp hạng phim và sở thích xem phim của khách hàng. Ý tưởng chính là thông qua khảo sát dữ liệu có được từ các phim sắp chiếu, sở thích xem phim của khách hàng để đưa ra các thuật toán dự đoán tốt. Một ứng dụng khác cũng khá nổi tiếng đó là Google flu trends của Google giúp phân tích số lượng lớn các dữ liệu truy vấn tìm kiếm của khách hàng để cảnh báo bệnh cúm trong một quần thể. Ứng dụng này giúp ngăn ngừa dịch bệnh cúm một cách hiệu quả [3].

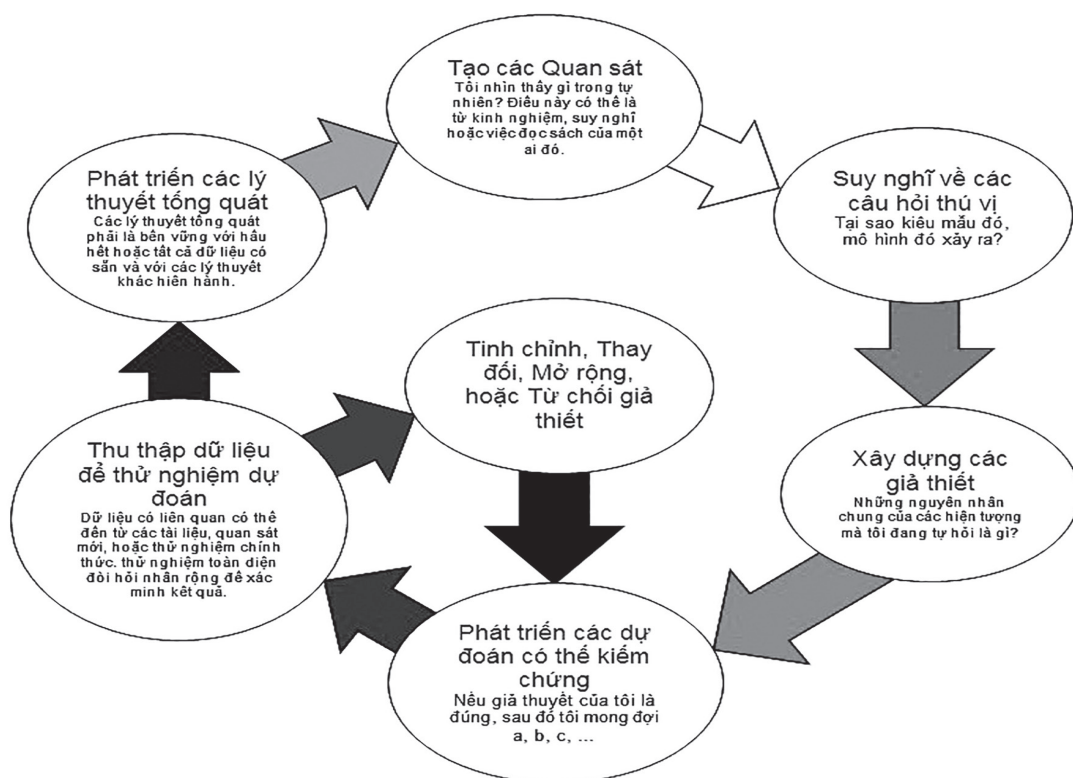
Một ví dụ khác là ứng dụng hỗ trợ tranh cử tổng của ông Obama. Các nhà KHDL đã tập trung để phát triển ứng dụng hỗ trợ chiến dịch tái tranh cử tổng thống Mỹ năm 2012 của ông Barack Obama góp phần giúp ông đắc cử tổng thống nhiệm kỳ thứ 2 liên tiếp [3]. Ngoài ra, Moneyball, một bộ phim về vai trò của phân tích hiệu suất trong bóng chày và bài học về định hướng dữ liệu (data-driven). Ứng dụng tìm hiểu lý do vì sao các cầu thủ trong đội bóng rời bỏ câu lạc bộ để đề xuất các biện pháp tốt nhất nhằm giữ họ ở lại. Một ứng dụng khác là Minard map, đây là bản đồ của Charles Joseph Minard trong lĩnh vực đồ họa thông tin về các công trình dân dụng và thống kê, địa lý, tiêu biểu cho việc tổ chức và phân tích dữ liệu.

#### 1.5. Các nhược điểm của Khoa học dữ liệu

Nhược điểm đầu tiên và dễ thấy nhất đó là sự cường điệu hóa (hype) hay còn gọi là cơn sốt KHDL. Hậu quả của cơn sốt này sẽ dẫn đến một tư duy sai lầm rằng KHDL có thể giải quyết được bất kỳ vấn đề nào trong thế giới thực. Tuy nhiên, trong thực tế thì không như vậy [1].

Một số nhược điểm khác có thể kể đến đó là Netflix prize failure và Google flu trends failure. Netflix không chạy thuật toán mới. Họ đã đánh giá offline và sử dụng lại hai thuật toán đã đoạt giải đầu tiên để xếp hạng phim và họ không có kế hoạch phát triển mới trong tương lai. Về Google flu trends failure thì các con số dự đoán bệnh cúm của Google đã bắt đầu sai lệch dần theo thời gian [2]. Khi dữ liệu lớn dần, những dự báo của Google đã bắt đầu sai lệch và đôi khi dẫn đến những dự đoán thiếu chính xác khiến cho nhiều bệnh nhân phải mất thời gian hơn để thăm khám bác sĩ.

Một số lỗi chung thường mắc phải khi bắt đầu nghiên cứu KHDL có thể kể đến như: bắt đầu phân tích mà không đặt câu hỏi, sử dụng dữ liệu chất lượng kém, chỉ tập trung vào công nghệ mà không quan tâm đến cơ sở lý thuyết và kiến thức chuyên môn, nhầm lẫn sự tương quan (correlation - same time) và quan hệ nhân quả (causation - trước sau), thất bại trong việc truyền đạt (communicate) các kết quả, làm phức tạp việc phân tích quá nhiều, thất bại để duy trì việc học kiến thức chuyên môn (failing to always keep learning). Dữ liệu rất lộn xộn và phức tạp, do vậy, KHDL là lĩnh vực không dễ dàng. Không quá khó để lập trình, nhưng điều đó sẽ dẫn đến sự thiếu chính xác. Cụm từ “nhà KHDL” ngụ ý chúng ta cần làm việc theo phương pháp khoa học như sơ đồ bên dưới, bắt đầu từ bước “Tạo các quan sát” [2].

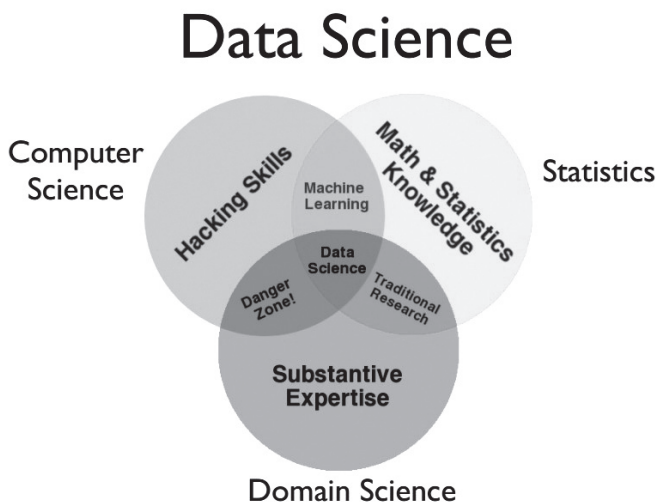


Hình 1. Mô phỏng quy trình làm việc một cách khoa học

## 2. Các chủ điểm chính trong Khoa học dữ liệu

Khoa học dữ liệu là nơi giao thoa của nhiều lĩnh vực khác nhau. Những trụ cột của KHDL bao gồm: tính toán, thống kê, toán học, các môn học định lượng và kiến thức khoa học chuyên ngành kết hợp để phân tích dữ liệu và cho ra quyết định tốt

hơn. Theo sơ đồ Venn về KHDL của Drew Conway, chúng ta có thể thấy được ba nội dung chính của KHDL là: các kỹ năng thâm nhập (hacking skills), các kiến thức toán học và thống kê (math & statistics knowledge) và các khoa học chuyên ngành (domain science).



Hình 2. Sơ đồ Venn về Khoa học dữ liệu (Drew Conway) [3]

Trong sơ đồ Venn trên, phần giao thoa giữa ba nội dung chính là KHDL. Nếu chỉ có các kỹ năng thâm nhập và kiến thức toán thống kê, chúng ta đang thực hiện các phương pháp máy học, nếu chỉ có các kỹ năng thâm nhập và kiến thức chuyên ngành thì đó là khu vực nguy hiểm (danger zone!). Sở dĩ nói như vậy là bởi vì khi không có kiến thức thống kê và toán học thì chúng ta sẽ có xu hướng chỉ cần chọn các thuật toán từ một thư viện và sử dụng chúng mà không thực sự hiểu những gì chúng ta đang làm, đây là một sự nguy hiểm có thể dẫn đến việc vận hành ứng dụng sai. Nếu chỉ có các kiến thức chuyên môn kết hợp với kiến thức toán và thống kê thì đó chính là cách nghiên cứu khoa học truyền thống.

Các kỹ năng thâm nhập (hacking skills) bao gồm tập hợp những kỹ năng cần có để thâm nhập sâu hơn vào dữ liệu, trong đó có kỹ năng tính toán. Kỹ năng tính toán đề cập đến khả năng xây dựng và tìm giải pháp thông minh cho các vấn đề. Một nhà KHDL sẽ thường xuyên sử dụng kỹ thuật, công nghệ để truy cập, thu thập, làm sạch và lưu trữ dữ liệu. Bên cạnh đó, họ cũng cần có hiểu biết về các cơ sở dữ liệu thường được sử dụng phổ biến là MySQL, PostgreSQL, Cassandra, MongoDB và CouchDB. Kiến thức về cơ sở dữ liệu là rất quan trọng, đặc biệt là các công nghệ dữ liệu lớn. Các công nghệ dữ liệu lớn (Big Data) thường được dùng như Hadoop, MapReduce và Spark. Một nội dung khác khá quan trọng trong kỹ năng thâm nhập đó là trực quan hóa và báo cáo. Đây là quá trình hiển thị kết quả và kết luận của các phân tích một cách trực quan dễ hiểu. Một số công cụ trực quan hóa thường được sử dụng, như: D3, Tableau, Qlikview và R Markdown.

Kiến thức toán học và thống kê (math & statistics knowledge) là một nội dung rất quan trọng của KHDL. Bản chất KHDL là chủ yếu dựa trên nền tảng thống kê. Do vậy, có thể nói thống kê là cốt lõi. Đại số tuyến tính (linear algebra) và tính toán định lượng sẽ là nền tảng để hiểu và áp dụng

các thuật toán máy học. Khi giỏi toán, bạn sẽ hiểu sâu hơn và biết khi nào sẽ cần sử dụng các thuật toán. Trong đó, được sử dụng nhiều nhất là ma trận đại số xuất hiện trong rất nhiều khái niệm máy học. Ngoài ra, các tính toán (calculus) bao gồm tính toán đa biến cũng thường được sử dụng trong một số khái niệm thống kê và các thuật toán máy học. Thống kê mô tả (descriptive statistics) cũng được sử dụng khá nhiều vì phương pháp tiếp cận đầu tiên đến một tập dữ liệu mới liên quan đến việc áp dụng phân tích mô tả. Trong khi đó, suy luận thống kê (statistical inference) là cần thiết trong kỹ thuật dự đoán để tạo ra kiến thức mới, không chỉ mô tả các dữ liệu thực tế.

Các kiến thức chuyên môn (domain science) cho phép bắt đầu dự án dữ liệu với mục tiêu rõ ràng. Nó cho phép nhà KHDL có thể định hướng rõ ràng hơn với các dạng dữ liệu sắp tiếp cận và lựa chọn phương pháp nghiên cứu tốt hơn. Khả năng đặt câu hỏi tốt cũng đòi hỏi sự hiểu biết chuyên ngành. Trong khi đó, kiến thức chuyên ngành thường rất rộng lớn, đến từ mọi lĩnh vực đời sống, từ khoa học tự nhiên đến khoa học xã hội. Đây cũng là một khó khăn đối với các nhà KHDL khi tiếp cận phân tích dữ liệu. Lĩnh vực chuyên ngành cũng bao gồm khả năng truyền đạt (communicate) rõ ràng và hiệu quả về các mẫu được tìm thấy trong dữ liệu.

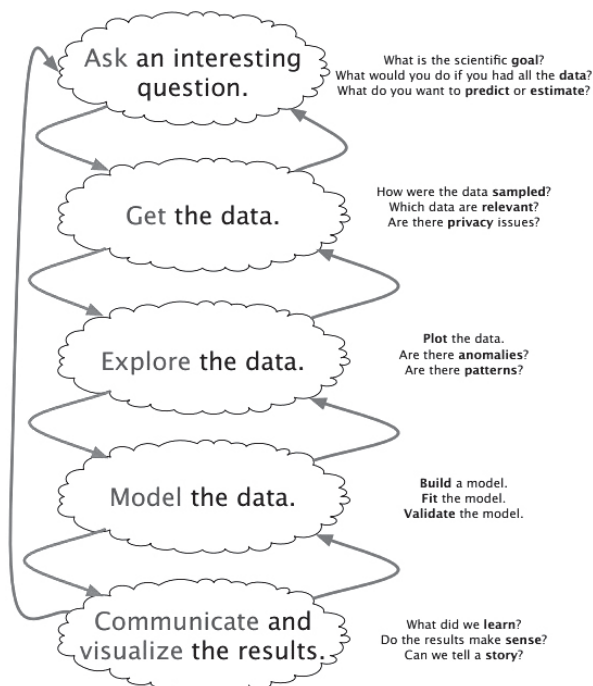
Theo Brandon Rohrer, chỉ có năm câu hỏi KHDL có thể trả lời:

- Đây là A hay là B? (phân nhóm 2 lớp) hoặc đây là A, hay B, hay C hay là D? (phân nhóm nhiều lớp).
- Đây có phải là điều kỳ lạ? (anomaly detection - phát hiện bất thường).
- Bao nhiêu? (regression - hồi quy).
- Nó được tổ chức như thế nào? (clustering - gom cụm).
- Tôi nên làm gì tiếp theo? (reinforcement learning - học tăng cường).

Kiến thức về học máy là cần thiết và sẽ luôn là nền tảng cơ bản giúp cho quy trình KHDH được thực hiện tốt hơn và đạt được hiệu quả như mong muốn.

### 3. Các phương pháp nghiên cứu KHDH

KHDH là sự tổng hợp kinh nghiệm về tri thức thao tác từ dữ liệu thô thông qua quy trình vòng đời dữ liệu hoàn chỉnh. Quy trình KHDH có thể được mô phỏng thông qua sơ đồ dưới đây:



**Hình 3.** Quy trình KHDH (Hanspeter Pfister, tài liệu khóa học CS109, 2013)

Bước vào giai đoạn đầu, chúng ta chưa biết mình sẽ làm gì tiếp theo với bộ dữ liệu. Điều này tương tự bước đầu của quá trình khám phá tri thức trong cơ sở dữ liệu. Sau khi có dữ liệu, chúng ta nên đặt các câu hỏi thú vị và xem xét các mục tiêu có thể có. Sau đó là quá trình lấy mẫu (sampling, get the data). Tiếp theo, chúng ta bắt đầu quá trình thăm dò dữ liệu, phân tích và phát hiện. Sau khi có kết quả, chúng ta mô hình hóa chúng bằng việc xây dựng các mô hình, chỉnh sửa cho phù hợp và xác nhận các giá trị của mô hình. Cuối cùng là trực quan hóa kết quả nghiên cứu và giao tiếp với kết quả thông qua các công cụ trực quan, từ đó có được các hiểu biết từ dữ liệu.

Các phương pháp nghiên cứu KHDH có thể bao gồm [2]:

- Thao tác với dữ liệu để có được thông tin của bộ dữ liệu quản lý (data munging/scraping/sampling/cleaning).
- Lưu trữ và quản lý dữ liệu để có thể truy cập dữ liệu - đặc biệt là dữ liệu lớn - một cách nhanh chóng và đáng tin cậy trong quá trình phân tích tiếp theo.
- Phân tích dữ liệu thăm dò để tạo ra các giả thuyết và trực giác về dữ liệu.
- Dự đoán dựa trên các công cụ thống kê như hồi quy, phân lớp, và gom cụm.
- Truyền đạt kết quả (communication) thông qua trực quan hóa, những câu chuyện, và các tóm lược có thể phiên dịch được.

Nhóm các phương pháp này có thể được chia thành các nhóm hoạt động như sau [2]:

+ Dự đoán (sử dụng quy trình khoa học dữ liệu, các kiểu dữ liệu và data “munging”, các kiến thức xác suất thống kê (probability review), phân lớp & hồi quy và gom cụm (classification & regression & clustering), trực quan hóa và kể chuyện (visualization & story telling).

+ Các kiến thức cần thiết khác: Bayesian Thinking & Computation, Monte Carlo Methods, Machine Learning Methods, Databases.

+ Các phân tích mạng: Network Visualization, Network Sampling, Network Models.

### 4. Ngôn ngữ và công cụ hỗ trợ nghiên cứu Khoa học dữ liệu

Ngôn ngữ được sử dụng phổ biến nhất là Python. Đây cũng là một ngôn ngữ mới xuất hiện trong thời gian gần đây. Python là một ngôn ngữ lập trình thông dịch, mã nguồn mở đa mục đích có hỗ trợ các phong cách lập trình cốt lõi khác nhau, được sử dụng tốt để phân tích dữ liệu. Python tạo kiểu động và dùng cơ chế cấp phát bộ nhớ tự động. Chúng ta có thể sử dụng Pandas

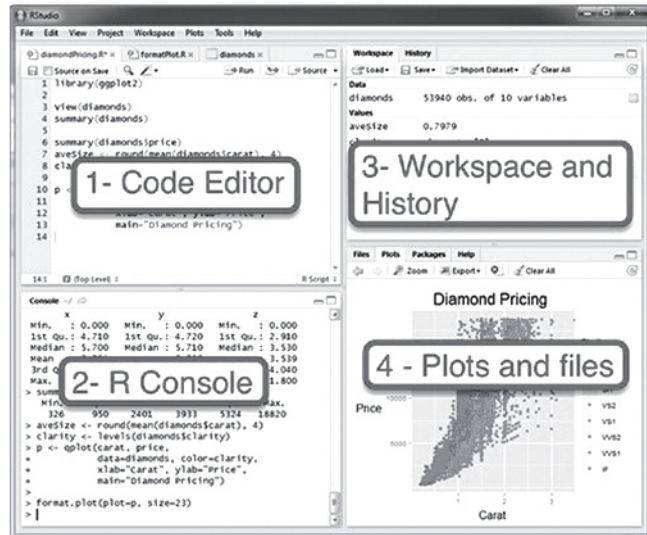
để thao tác dữ liệu trong Python, NumPy để thao tác mảng dữ liệu, scikits learn để khai thác dữ liệu trong Python...

Ngoài ra, R là một môi trường phần mềm dành cho tính toán và đồ họa thống kê. R hỗ trợ tốt cho lĩnh vực KHDH thông qua

các tính toán và biểu đồ thống kê chuyên ngành. Môi trường ứng dụng R vẫn đang tiếp tục phát triển nhanh chóng và mạnh mẽ trong nhiều ngành và lĩnh vực khác nhau, trong đó có KHDH.



RConsole  
RCommander  
RStudio  
RMarkdown



Hình 4. Môi trường R hỗ trợ khoa học dữ liệu

## Kết luận

KHDH sẽ phát triển mạnh mẽ hơn nữa trong tương lai và giúp cho con người tận dụng được tối đa tiềm năng của các nguồn dữ liệu lớn. KHDH sẽ làm cho các hệ thống trở nên thông minh hơn và góp phần xây dựng nên một thế giới thông minh trong kỷ nguyên số. Trong đó, con người sẽ có nhiều cơ hội thâm nhập sâu hơn vào dữ liệu, hỗ trợ đưa ra các quyết định tốt hơn trong nhiều lĩnh vực của cuộc sống. Tương lai phát triển của KHDH sẽ hứa hẹn nhiều điều thú vị và có ý nghĩa lớn đối với cuộc sống con người.

## TÀI LIỆU THAM KHẢO

1. Rachel Schutt, Cathy O'Neil,. (2013). *Doing Data Science*. O'Reilly Media, Inc. 405 pages.
2. CS109 Data Science Course by Harvard University,  
Truy cập từ: <http://cs109.github.io/2015>, ngày 4/11/2017
3. CS194-16 Introduction to Data Science Fall 2014,

Truy cập từ: <https://bcourses.berkeley.edu/courses/1267848/wiki>, ngày 4/11/2017

4. Data science,

Truy cập từ: [https://en.wikipedia.org/wiki/Data\\_science](https://en.wikipedia.org/wiki/Data_science), ngày 2/11/2017

5. KHDH,

Truy cập từ: [https://vi.wikipedia.org/wiki/Khoa\\_học\\_dữ\\_liệu](https://vi.wikipedia.org/wiki/Khoa_học_dữ_liệu), ngày 2/11/2017

6. KHDH là ngành thu hút nhất thế kỷ 21,

Truy cập từ: <http://www.giaoducdaotao.org/home/khoa-hoc-du-lieu-la-nganh-thu-hut-nhat-the-ky-21>, ngày 4/11/2017

7. Làm sao để trở thành một nhà KHDH,

Truy cập từ: <https://bigsonata.wordpress.com/2014/11/15/lam-sao-de-tro-thanh-nha-khoa-hoc-du-lieu>, ngày 4/11/2017

8. Tổng quan về Dữ liệu lớn,

Truy cập từ: <http://vienthongke.vn/attachments/article/2290/Bai4.So5.2016.pdf>, ngày 4/11/2017.

(Ngày Tòa soạn nhận được bài: 6-5-2018;  
Ngày phản biện đánh giá: 15-9-2018; Ngày chấp nhận đăng: 15-10-2018).