
# BÀI GIẢNG NHẬP MÔN KHOA HỌC DỮ LIỆU

Trần Quang Quý* Hà Thị Thanh†

Thái Nguyên, tháng 06 năm 2023

# Mục lục



*Trường Đại học Công nghệ Thông tin & Truyền thông - ICTU, tqquy@ictu.edu.vn
†Trường Đại học Công nghệ Thông tin & Truyền thông - ICTU, httthanh@ictu.edu.vn

# 1 Mở đầu

**Data science** còn được biết đến với tên gọi là Khoa học dữ liệu. Đúng với tên gọi của nó, về mặt bản chất, đây chính là công việc thu thập và phân tích dữ liệu. Data science là một lĩnh vực liên ngành mà trong đó, những bộ dữ liệu được xử lý, sắp xếp và giải mã bằng các mô hình thống kê hay phương pháp toán học.

Các kỹ sư khoa học dữ liệu sẽ sử dụng tất cả các điểm dữ liệu khác nhau thu thập được, từ đó tạo ra một mô hình dữ liệu hoặc thuật toán để áp dụng cho từng mục đích cụ thể mang tính chiến lược. Có 2 mục đích chính:

- Sử dụng dữ liệu để phân tích chuyên sâu về một vấn đề nào đó và đưa ra các giải pháp, hoặc dự đoán cho tương lai.
- Xây dựng các mô hình dữ liệu để tạo ra các sản phẩm, hoặc tính năng công nghệ nào đó.

Có thể nói, ứng dụng của Data Science là vô cùng quan trọng khi số lượng dữ liệu được tạo ra mỗi ngày là rất lớn, tăng lên theo cấp số nhân, và ngày càng trở nên phức tạp. Điều đó khiến việc xử lý dữ liệu thông qua các công cụ như Excel không còn khả thi. Thay vào đó, các kỹ sư khoa học dữ liệu phải sử dụng đến ngôn ngữ lập trình cao cấp hơn, phổ biến nhất là Python và R, để hiểu dữ liệu.

Trong nội dung học phần "Nhập môn khoa học dữ liệu", tiếng Anh là Introduction to Data Science sẽ trình bày những hiểu biết cũng như những kiến thức về thống kê, phân tích dữ liệu và một số nội dung khác liên quan tới Khoa học dữ liệu. Hai ngôn ngữ được sử dụng chính trong quyển bài giảng là ngôn ngữ Python và ngôn ngữ R.

# 2 Giới thiệu chung

## 2.1 Tổng quan về Khoa học dữ liệu

### 2.1.1 Lịch sử

Chắc chắn bạn đã có kinh nghiệm về khoa học dữ liệu dưới nhiều hình thức. Điều này có thể khẳng định được! Ví dụ đơn giản là khi bạn đang tìm kiếm thông tin trên web bằng cách sử dụng công cụ tìm kiếm hoặc hỏi đường trên điện thoại di động của mình, bạn đang tương tác với các sản phẩm khoa học dữ liệu. Khoa học dữ liệu đã hỗ trợ giải quyết một số nhiệm vụ hàng ngày phổ biến nhất của chúng ta trong vài năm. Thuật ngữ Khoa học dữ liệu không phải là mới mẻ, mà nó là sản phẩm của Toán học và Thống kê từ xưa trong một thời gian dài. Thống kê là một ngành khoa học lâu đời đứng trên vai những người khổng lồ của thế kỷ 18 như Pierre Simon Laplace (1749–1827) và Thomas Bayes (1701–1761)[1]

### 2.1.2 Dữ liệu hóa

Tính mới của khoa học dữ liệu không bắt nguồn từ kiến thức khoa học mới nhất, mà bắt nguồn từ một sự thay đổi mang tính đột phá trong xã hội của chúng ta do sự phát triển của công nghệ gây ra: dữ liệu hóa - **Datification**. Dữ liệu hóa là quá trình kết xuất thành các khía cạnh dữ liệu của thế giới chưa từng được định lượng trước đây. Ở cấp độ một thực thể nào đó, danh sách các khái niệm được dữ liệu hóa rất dài và vẫn đang tiếp tục phát triển: mạng lưới kinh doanh, quyển sách chúng ta đang đọc, những bộ phim chúng ta thích, thức ăn chúng ta ăn, hoạt động thể chất, mua sắm, hành vi lái xe của chúng ta, v.v. Ở cấp độ kinh doanh, các công ty đang xác định dữ liệu bán cấu trúc (semi-structured) mà trước đây đã từng bị loại bỏ không được sử dụng như: nhật ký hoạt động web, hoạt động mạng máy tính, tín hiệu máy móc, v.v. Dữ liệu phi cấu trúc (unstructured), chẳng hạn như báo cáo bằng văn bản, e-mail hoặc bản ghi âm, hiện đang được lưu trữ không chỉ cho mục đích lưu trữ mà còn được phân tích[2].

### 2.1.3 Xu hướng tương lai

Khoa học dữ liệu (Data science) là ngành khoa học về việc khai phá, quản trị và phân tích dữ liệu để dự đoán các xu hướng trong tương lai và đưa ra các quyết định, chiến lược hành động. Khoa học dữ liệu (Data science) gồm ba phần chính: tạo và quản trị dữ liệu, phân tích dữ liệu, và áp dụng kết quả phân tích thành những hành động có giá trị. Việc phân tích và sử dụng dữ liệu dựa vào ba nguồn tri thức: toán học (thống kê toán học - Mathematical Statistics), công nghệ thông tin (máy học - Machine Learning) và tri thức của lĩnh vực ứng dụng cụ thể.

Số liệu lớn (Big Data) đã cách mạng hóa các công ty và đem lại cho họ một lợi thế cạnh tranh. Các công ty này cần những người chuyên môn, thành thạo trong việc xử lý, quản lý, phân tích và hiểu xu hướng trong dữ liệu. Chính vì thế mà ngành Khoa học dữ liệu (Data science) càng ngày càng trở thành xu hướng và được săn đón.

## 2.2 Ngành Khoa học dữ liệu

### 2.2.1 Khái niệm

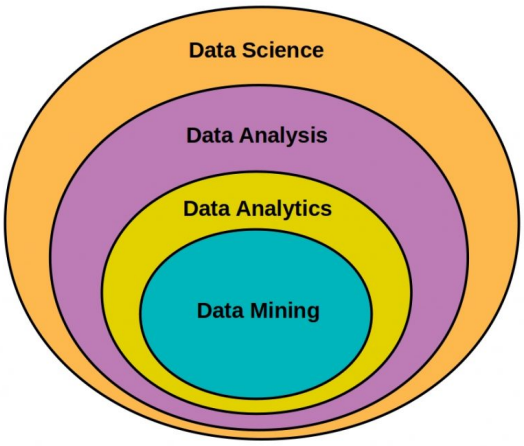
Có nhiều quan điểm khác nhau về ngành khoa học dữ liệu (data science) nhưng về cơ bản thì đây là lĩnh vực sử dụng các phương pháp khoa học, thuật toán và hệ thống công nghệ thông tin để giải mã 'thông tin' đằng sau các dữ liệu ngẫu nhiên, từ đó biến các báo cáo thô thành các thông tin có giá trị.

Những thông tin này có thể ứng dụng đa dạng vào nhiều lĩnh vực khác nhau như giúp ngành y tế phân tích hình ảnh y khoa, phân nhóm khách hàng trong ngành tài chính hay dự báo rủi ro cho nhóm ngành sản xuất. Dưới sự hỗ trợ của khoa học dữ liệu, nhiều vấn đề trước giờ vẫn 'tốn nhân công, tốn nguồn lực' để giải quyết, nay hoàn toàn có thể xử lý nhanh chóng, hiệu quả hơn, đặc biệt khi công nghệ ngày càng hoàn thiện và phổ biến..

### 2.2.2 Các cấp độ

Khoa học dữ liệu là một ngành lớn và có những phân cấp nhỏ hơn để giải quyết những vấn đề cụ thể, chi tiết trong các lĩnh vực khác nhau:

- Cấp thứ 1- data mining: khoanh vùng và tìm kiếm dữ liệu đầu vào dưới sự hỗ trợ của công nghệ và thuật toán thống kê.
- Cấp thứ 2- data analytics: Cách con người ứng dụng công nghệ phân tích để biến những dữ liệu đã khai thác thành các nhóm thông tin cụ thể.
- Ở cấp thứ 3- data analysis: hệ thống hóa và xây dựng cấu trúc cho những thông tin đầu ra của cấp thứ 2. Những thông tin này sẽ được đào sâu phân tích, kết nối và diễn giải thành các thông tin thông dụng, mang tính ứng dụng cao.

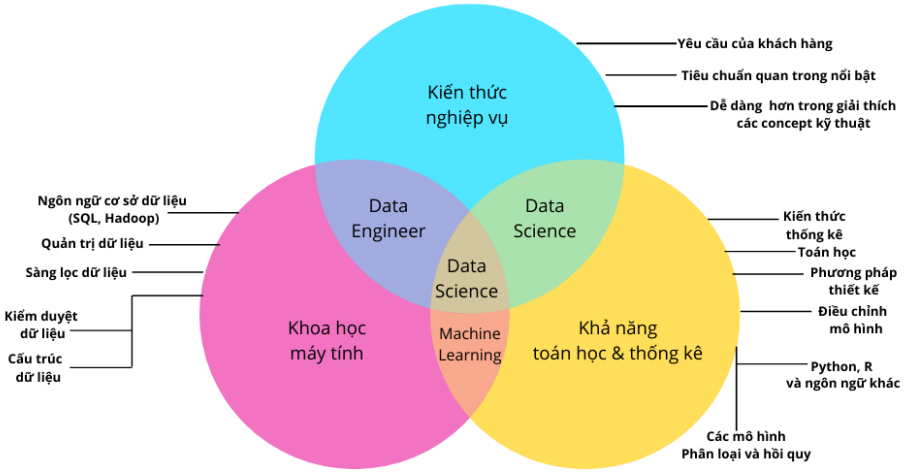



Hình 1: Các cấp độ trong Khoa học dữ liệu

Data science ngày càng được ứng dụng trong nhiều lĩnh vực khi công nghệ ngày càng cải tiến khiến cho tính chính xác của dữ liệu gia tăng. Trong đó tính ứng dụng của khoa học dữ liệu đặc biệt rõ nét trong kinh doanh vì các quyết định kinh doanh sẽ có được cơ sở chắc chắn và giúp sớm loại bỏ các quyết định nhiều rủi ro.

Các nhóm kỹ năng cần thiết của một nhà khoa học dữ liệu bao gồm Phân tích (Analytics), Lập trình (Programming), và Kiến thức chuyên ngành (Domain Knowledge). Chính vì thế, nếu theo học ngành Khoa học dữ liệu, chúng ta sẽ được học một số các môn chuyên ngành như:

- Thống kê ứng dụng (Applied Statistics)
- Nhập môn Khoa học máy tính (Introduction to Computer Science)
- Lập trình cùng Python, R hay SQL (Programming with Python/R/SQL)
- Trực quan hóa dữ liệu (Data Visualization)
- Xác suất (Probability)
- Khai phá dữ liệu (Data Mining)
- Công cụ trực quan hóa dữ liệu - Tableau, Power BI

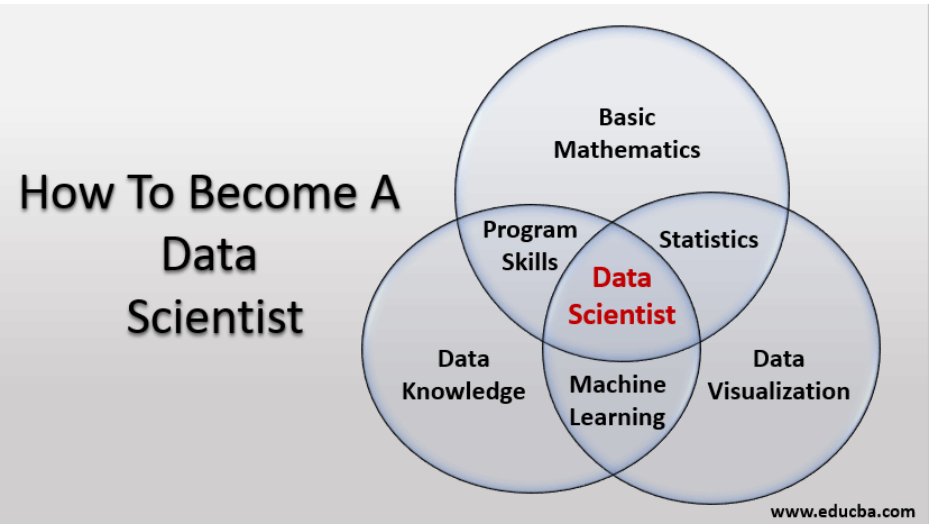



Hình 2: Tổng quan chung về Data Science

Để trở thành một nhà Khoa học dữ liệu - Data Scientist, chúng ta cần có kiến thức tổng quan như miêu tả ở Hình 2 và Hình 3.

Nếu theo đuổi ngành khoa học dữ liệu trong trường học, sau đây là những kỹ năng về kỹ thuật mà chúng ta cần học. Nói cách khác, để có thể tìm kiếm cơ hội việc làm và phát triển trong ngành data science thì chúng ta không thể bỏ qua những kỹ năng quan trọng này:

- Phân tích thống kê và tính toán
- Xử lý tập dữ liệu lớn
- Sắp xếp dữ liệu
- Machine learning
- Deep learning




Hình 3: Kỹ năng để trở thành Data Scientist

- Trực quan hoá dữ liệu
- Toán học
- Lập trình (các ngôn ngữ lập trình Python, Java, C/C++, v.v.)
- Big data
- Thống kê

Ngoài ra, ngành khoa học dữ liệu nói chung và một data scientist nói riêng không thể thiếu những công cụ hỗ trợ phân tích dữ liệu như SQL, R, Spark, SAS, Hive, v.v.

## 2.3 Ngôn ngữ lập trình

Có rất nhiều ngôn ngữ (công cụ) giúp chúng ta thao tác với dữ liệu và khám phá dữ liệu. Gần đây, hai ngôn ngữ như Python và R đã chứng minh được sức mạnh như những ngôn ngữ mang tính kịch bản (script) để xử lý dữ liệu.

### 2.3.1 Ngôn ngữ Python

Python là ngôn ngữ lập trình máy tính bậc cao thường được sử dụng để xây dựng trang web và phần mềm, tự động hóa các tác vụ và tiến hành phân tích dữ liệu. Python là ngôn ngữ có mục đích chung, nghĩa là nó có thể được sử dụng để tạo nhiều chương trình khác nhau và không chuyên biệt cho bất kỳ vấn đề cụ thể nào.

**Một vài sự thật thú vị về Python:**

- Python được phát triển vào cuối những năm 1980 bởi Guido van Rossum tại Viện Nghiên cứu Quốc gia về Toán học và Khoa học Máy tính ở Hà Lan với tư cách là người kế thừa ngôn ngữ ABC có khả năng xử lý và giao tiếp ngoại lệ.
- Python có nguồn gốc từ các ngôn ngữ lập trình như ABC, Modula 3, small talk, Algol-68.

- Van Rossum đã chọn tên Python cho ngôn ngữ mới từ một chương trình truyền hình, Monty Python's Flying Circus.
- Trang Python là một tệp có phần mở rộng .py chứa có thể là sự kết hợp của Thẻ HTML và tập lệnh Python.
- Vào tháng 12 năm 1989, người sáng tạo đã phát triển trình thông dịch python đầu tiên như một sở thích, và sau đó vào ngày 16 tháng 10 năm 2000, Python 2.0 được phát hành với nhiều tính năng mới.
- Vào ngày 3 tháng 12 năm 2008, Python 3.0 được phát hành với nhiều thử nghiệm hơn và bao gồm các tính năng mới.
- Python là một ngôn ngữ kịch bản mã nguồn mở.
- Python là mã nguồn mở, có nghĩa là bất kỳ ai cũng có thể tải xuống miễn phí từ trang chủ và sử dụng nó để phát triển các chương trình. Mã nguồn của nó có thể được truy cập và sửa đổi theo yêu cầu trong dự án.
- Python là một trong những ngôn ngữ chính thức tại Google.

**Ứng dụng của Python:**

Python thường được sử dụng để phát triển trang web và phần mềm, tự động hóa tác vụ, phân tích dữ liệu và trực quan hóa dữ liệu. Vì tương đối dễ học, Python đã được nhiều người không phải là lập trình viên như kế toán và nhà khoa học áp dụng cho nhiều công việc hàng ngày, chẳng hạn như tổ chức tài chính.

**Phân tích dữ liệu và học máy:**

Python đã trở thành một yếu tố chính trong khoa học dữ liệu, cho phép các nhà phân tích dữ liệu và các chuyên gia khác sử dụng ngôn ngữ này để thực hiện các phép tính thống kê phức tạp, tạo trực quan hóa dữ liệu, xây dựng thuật toán học máy, thao tác và phân tích dữ liệu cũng như hoàn thành các nhiệm vụ khác liên quan đến dữ liệu.

Python có thể xây dựng nhiều dạng trực quan hóa dữ liệu khác nhau, chẳng hạn như biểu đồ đường và thanh, biểu đồ hình tròn, biểu đồ 3D. Python cũng có một số thư viện cho phép các lập trình viên viết chương trình để phân tích dữ liệu và học máy nhanh hơn và hiệu quả hơn, như TensorFlow và Keras.

**Phát triển web:**

Python thường được sử dụng để phát triển back-end của trang web hoặc ứng dụng—những phần mà người dùng không nhìn thấy. Vai trò của Python trong phát triển web có thể bao gồm gửi dữ liệu đến và đi từ máy chủ, xử lý dữ liệu và giao tiếp với cơ sở dữ liệu, định tuyến URL và đảm bảo tính bảo mật. Python cung cấp một số khuôn khổ để phát triển web. Những cái thường được sử dụng bao gồm Django và Flask.

Một số công việc phát triển web sử dụng Python bao gồm kỹ sư phụ trợ, nhà phát triển Python, kỹ sư phần mềm và kỹ sư DevOps.