**ROADMAP DATA SCIENCE**

Người viết: Bùi Minh Cường

Ngày Viết: 30/09/2023

**GIỚI THIỆU**

**Tổng quan**

Lộ trình này dành cho những ai quan tâm đến chuyên ngành Khoa Học Dữ Liệu và muốn đột nhập vào chuyên ngành nhằm mục đích tìm hiểu.

Khoa học dữ liệu là lĩnh vực nghiên cứu kết hợp chuyên môn về miền, kỹ năng lập trình và kiến thức về toán học và thống kê để trích xuất những hiểu biết có ý nghĩa từ dữ liệu. Các học viện khoa học dữ liệu áp dụng các thuật toán học máy cho các con số, văn bản, hình ảnh, video, âm thanh, v.v. để tạo ra các hệ thống trí tuệ nhân tạo (AI) để thực hiện các nhiệm vụ thường đòi hỏi trí thông minh của con người. Đổi lại, các hệ thống này tạo ra những hiểu biết sâu sắc mà các nhà phân tích và người dùng doanh nghiệp có thể chuyển thành giá trị kinh doanh hữu hình.

Dữ liệu có giá trị, và khoa học trong việc giải mã nó cũng vậy. Hàng triệu byte dữ liệu đang được tạo ra, và bây giờ giá trị của nó cũng đã vượt qua dầu. Vai trò của một nhà khoa học dữ liệu đang và sẽ có tầm quan trọng tối cao đối với các tổ chức trên nhiều ngành dọc.

Dữ liệu mà không có khoa học thì không là gì cả. Dữ liệu cần được đọc và phân tích. Điều này đòi hỏi phải có chất lượng dữ liệu và hiểu cách đọc nó và thực hiện các khám phá theo hướng dữ liệu.

Dữ liệu sẽ giúp tạo ra trải nghiệm khách hàng tốt hơn. Đối với hàng hóa và sản phẩm, khoa học dữ liệu sẽ tận dụng sức mạnh của học máy để cho phép các công ty tạo và sản xuất các sản phẩm mà khách hàng sẽ yêu thích. Ví dụ: đối với một công ty Thương mại điện tử, một hệ thống đề xuất tuyệt vời có thể giúp họ khám phá chân dung khách hàng của mình bằng cách xem lịch sử mua hàng của họ.

Dữ liệu sẽ được sử dụng trên các ngành dọc. Khoa học dữ liệu không chỉ giới hạn ở hàng tiêu dùng hoặc công nghệ hoặc chăm sóc sức khỏe. Sẽ có nhu cầu cao để tối ưu hóa các quy trình kinh doanh bằng cách sử dụng khoa học dữ liệu từ ngân hàng và vận tải đến sản xuất. Vì vậy, bất cứ ai muốn trở thành một nhà khoa học dữ liệu sẽ có một thế giới cơ hội hoàn toàn mới mở ra ngoài kia. Tương lai là dữ liệu.

**Chúng ta sẽ học những gì ?**

1. Khoa học cơ bản
2. Phân tích dữ liệu
3. Máy học
4. And More - More

**Work**

1. Học về lập trình hoặc kỹ thuật lập trình - Ước tính thời gian cần: 2 – 3 tháng

Đầu tiên, hãy đảm bảo rằng bạn có kỹ năng lập trình. Mọi công việc liên quan đến data science đều yêu cầu kiến thức về lập trình trong ít nhất một ngôn ngữ.

Các chủ đề lập trình cụ thể cần biết

Cấu trúc dữ liệu phổ biến (data types, lists, dictionaries, sets, tuples), viết hàm, logic, luồng điều khiển, thuật toán tìm kiếm và sắp xếp, lập trình hướng đối tượng và làm việc với thư viện bên ngoài.

Tập lệnh SQL: Truy vấn cơ sở dữ liệu bằng cách sử dụng các phép nối, tổng hợp và truy vấn con (joins, aggregations, and sub-queries)

Sử dụng tốt Terminal, công cụ kiểm soát phiên bản trong Git và sử dụng GitHub

Các nguồn tài liệu và khóa học về Python

learnpython.org (miễn phí) – Tài nguyên miễn phí để học Python (và các ngôn ngữ khác) cho người mới bắt đầu. Nó bao gồm tất cả các chủ đề lập trình cơ bản từ đầu. Bạn có thể thực hành các chủ đề ngay trên đó.

Kaggle (miễn phí) – hướng dẫn tương tác học python. Đây là một hướng dẫn ngắn bao gồm tất cả các chủ đề quan trọng của data science.

Python certifications on freeCodeCamp (miễn phí) – freeCodeCamp cung cấp một số chứng chỉ dựa trên Python, chẳng hạn như tính toán khoa học, phân tích dữ liệu và học máy.

Python Course by freecodecamp on YouTube (miễn phí) –  Khóa học kéo dài 5 giờ mà bạn có thể theo học để thực hành các khái niệm cơ bản.

Intermediate python (miễn phí) – một khóa học miễn phí liên quan đến data science khác của Patrick trên freecodecamp.org.

Coursera Python for Everybody Specialization (có trả phí) – Khóa học này  bao gồm các khái niệm cấp độ mới bắt đầu, cấu trúc dữ liệu python, thu thập dữ liệu từ web và sử dụng cơ sở dữ liệu với python.

Bắt đầu học Python với Hướng dẫn Python (Get into Python). Công cụ này dành cho tất cả những ai muốn học ngôn ngữ lập trình Python, cho dù bạn là người mới bắt đầu hay đã là chuyên nghiệp.

Tài liệu học về Git và GitHub

Hướng dẫn Git và GitHub (miễn phí): hoàn thành các hướng dẫn và labs để phát triển khả năng kiểm soát phiên bản một cách vững chắc. Nó sẽ có ích cho bạn trong việc đóng góp cho các dự án nguồn mở.

Khóa học Git and GitHub crash course trên freeCodeCamp YouTube channel.

Tài liệu học SQL

Intro to SQL và Advanced SQL trên Kaggle.

Datacamp cũng có nhiều khóa học về SQL.

Khóa học course on SQL and Databases trên freeCodeCamp YouTube channel

2. Học về thu thập và sắp xếp dữ liệu (Data Collection & Wrangling) - Ước tính thời gian cần: 2 tháng

Một phần quan trọng của công việc data science là tập trung vào việc tìm kiếm dữ liệu phù hợp có thể giúp bạn giải quyết vấn đề của mình. Bạn có thể thu thập dữ liệu từ các nguồn hợp pháp khác nhau – cóp nhặt (nếu trang web cho phép), API, Cơ sở dữ liệu và các kho lưu trữ có sẵn công khai.

Sau khi bạn có dữ liệu trong tay, nhà phân tích thường sẽ tự tìm cách làm sạch các tập dữ liệu (dataframes), làm việc với các mảng đa chiều, sử dụng tính toán mô tả / khoa học và thao tác với dataframes để tổng hợp dữ liệu.

Dữ liệu hiếm khi sạch và được định dạng để sử dụng trong “thế giới thực”. Pandas và NumPy là hai thư viện cho phép bạn chuyển từ dữ liệu bẩn sang dữ liệu sẵn sàng phân tích.

Khi bạn bắt đầu cảm thấy thoải mái khi viết các chương trình Python, hãy bắt đầu tham gia các bài học về cách sử dụng các thư viện như pandas và numpy.

Ý tưởng về dự án thu thập dữ liệu:

Thu thập dữ liệu từ một trang web / API (mở cho sử dụng công khai) mà bạn chọn và chuyển đổi dữ liệu để lưu trữ từ các nguồn khác nhau thành một tệp hoặc bảng tổng hợp (DB). Các API mẫu bao gồm TMDB, quandl, Twitter API, v.v.

Chọn bất kỳ tập dữ liệu nào có sẵn công khai và xác định một bộ câu hỏi mà bạn muốn theo đuổi sau khi xem tập dữ liệu và domain. Thu thập dữ liệu để tìm ra câu trả lời cho những câu hỏi đó bằng cách sử dụng Pandas và NumPy.

3. Học về Exploratory Data Analysis, Business Acumen, Storytelling - Ước tính thời gian cần: 2-3 tháng

Một số thuật ngữ

Exploratory Data Analysis : Phân tích dữ liệu thăm dò, là phương pháp phân tích dữ liệu chủ yếu sử dụng các kỹ thuật về biểu đồ, hình vẽ, cho phép bạn phát triển ý chính về dữ liệu của bạn trông như thế nào và những loại câu hỏi nào có thể được chúng trả lời

Business Acumen: Nhạy bén kinh doanh

Data Storytelling: Kể chuyện bằng dữ liệu

Giai đoạn tiếp theo cần nắm vững là phân tích dữ liệu và kể chuyện. Rút ra thông tin chi tiết từ dữ liệu và sau đó truyền đạt thông tin tương tự bằng các thuật ngữ và hình ảnh hóa đơn giản là trách nhiệm cốt lõi của Nhà phân tích dữ liệu.

4. Tìm hiểu về Kỹ thuật Dữ liệu - Ước tính thời gian cần: 4-5 tháng

Kỹ thuật dữ liệu (data engineering) hỗ trợ các nhóm R&D bằng cách cung cấp dữ liệu sạch cho các kỹ sư nghiên cứu và nhà khoa học tại các công ty dựa trên dữ liệu lớn. Bản thân nó là một lĩnh vực và bạn có thể quyết định bỏ qua phần này nếu bạn muốn chỉ tập trung vào khía cạnh thuật toán thống kê của các vấn đề.

Các trách nhiệm của một kỹ sư dữ liệu bao gồm xây dựng một kiến trúc dữ liệu hiệu quả, hợp lý hóa việc xử lý dữ liệu và duy trì các hệ thống dữ liệu quy mô lớn.

Các kỹ sư sử dụng Shell (CLI), SQL và Python / Scala để tạo ETL pipelines, tự động hóa các tác vụ hệ thống tệp và tối ưu hóa các hoạt động cơ sở dữ liệu để làm cho chúng có hiệu suất cao.

Một kỹ năng quan trọng khác là triển khai các kiến trúc dữ liệu này đòi hỏi sự thành thạo về các nhà cung cấp dịch vụ đám mây như AWS, Google Cloud Platform, Microsoft Azure, v.v.

5. Học về thống kê ứng dụng và toán học - Ước tính thời gian cần: 4-5 tháng

Phương pháp thống kê là một phần trung tâm của data science. Hầu hết tất cả các cuộc phỏng vấn khoa học dữ liệu chủ yếu tập trung vào thống kê mô tả và suy luận.

Đa phần mọi người thường bắt đầu viết code về các giải thuật machine learning mà không hiểu rõ về các phương pháp thống kê và toán học cơ bản giải thích hoạt động của các thuật toán đó. Tất nhiên, đây không phải là điều nên khuyến khích.

Các chủ đề chính về Thống kê ứng dụng (Applied Statistics) và toán học

Thống kê mô tả (Descriptive Statistics) – Tìm hiểu về các ước tính của vị trí (trung bình, trung vị, chế độ, thống kê có trọng số, thống kê được cắt bớt) và khả năng thay đổi để mô tả dữ liệu.

Thống kê tham chiếu (Inferential statistics) – thiết kế kiểm tra giả thuyết, kiểm tra A / B, xác định số liệu kinh doanh, phân tích dữ liệu thu thập và kết quả thử nghiệm bằng cách sử dụng khoảng tin cậy, giá trị p và giá trị alpha.

Đại số tuyến tính, phép tính đơn và đa biến (Linear Algebra, Single and multivariate calculus) để hiểu các hàm mất mát, gradient và các trình tối ưu hóa trong học máy.

6. Học về Machine Learning và AI - Ước tính thời gian cần: 4-5 tháng

Sau khi tự tìm hiểu và xem qua tất cả các khái niệm chính nói trên, bây giờ bạn nên sẵn sàng để bắt đầu với các thuật toán Machine Learning ưa thích.

Có ba kiểu học chính:

Học có giám sát (Supervised Learning ) – bao gồm các bài toán hồi quy và phân loại. Nghiên cứu hồi quy tuyến tính đơn giản, hồi quy bội, hồi quy đa thức, thuật toán Naive Bayes , hồi quy logistic, KNNs, mô hình cây, mô hình tổng hợp. Tìm hiểu về các chỉ số đánh giá.

Học không giám sát (Unsupervised Learning ) – Phân cụm (Clustering) và giảm chiều dữ liệu ( dimensionality reduction) là hai ứng dụng được sử dụng rộng rãi của học không giám sát. Đi sâu vào PCA (Principal component Analysis), phân cụm K-mean, tạo cây phân cấp (hierarchical clustering) và mô hình gaussian hỗn hợp (gaussian mixtures).

Học tập tăng cường (Reinforcement learning) – giúp bạn xây dựng hệ thống tự thưởng. Tìm hiểu cách tối ưu hóa phần thưởng, sử dụng thư viện TF-Agents, mô hình Deep Q-networks, v.v. Tuy nhiên bạn có thể bỏ qua phần này.

**New-bie**

1. Thống kê mô tả [Giới thiệu về thống kê](https://www.udacity.com/course/intro-to-descriptive-statistics--ud827)

Khóa học tương tự về các nguyên tắc cơ bản về thống kê trên [YouTube](https://www.youtube.com/playlist?list=PLAwxTw4SYaPn22DmaF6x8JtG4TeWOJk_1)

[StatQuest - Youtube](https://www.youtube.com/playlist?list=PLblh5JKOoLUK0FLuzwntyYI10UQFUhsY9)

Thống kê [trực tuyến](http://onlinestatbook.com/Online_Statistics_Education.pdf)

Giới thiệu về thống kê mô tả [Article 1](https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fdescriptive-statistics-f2beeaf7a8df) & [Article 2](https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fintro-to-descriptive-statistics-252e9c464ac9)

[Khóa học tiếng Ả Rập](https://www.youtube.com/watch?v=d5jh5mmwcKI&list=PLY99ZSsxRyJiu6kb4WRRpeEFqK1pAr-EO)

[Giới thiệu về Thống kê suy luận](https://www.udacity.com/course/intro-to-inferential-statistics--ud201) ++

[Thống kê thực tế dành cho nhà khoa học dữ liệu](https://github.com/Moataz-Elmesmary/Data-Science-Roadmap/blob/main/Practical%20Statistics%20for%20Data%20Scientists.pdf)

2. Xác suất

[Học viện Khan](https://www.khanacademy.org/math/statistics-probability/probability-library)

[Khóa học tiếng Ả Rập](https://www.youtube.com/playlist?list=PL158D091D26F47358)

[Giới thiệu về Xác Suất](https://drive.google.com/file/d/15Y0oFNHQRls1qvQNvO3DFLJVhIZvUjTD/view?usp=sharing)

3. Ngôn ngữ lập trình

🔹R - công cụ tốt để trực quan hóa và phân tích thống kê.

[Giới thiệu về R (Datacamp)](https://www.datacamp.com/courses/free-introduction-to-r)

[Chuyên ngành Khoa học Dữ liệu - Coursera](https://www.coursera.org/specializations/jhu-data-science)

[Giới thiệu về R](https://cran.r-project.org/doc/manuals/R-intro.pdf)

[R cho Khoa học dữ liệu](https://r4ds.had.co.nz/)

🔹Python💯

[Giới thiệu về lập trình Python](https://www.udacity.com/course/introduction-to-python--ud1110)

[OOP](https://learn.datacamp.com/courses/object-oriented-programming-in-python)

Tiếng Ả Rập - [Hassouna](https://www.youtube.com/watch?v=MxYLqE3Ils8&list=PLHIfW1KZRIfnM9y0sQRwjVz2-IwvnEJep) | [Elzero](https://www.youtube.com/watch?v=mvZHDpCHphk&list=PLDoPjvoNmBAyE_gei5d18qkfIe-Z8mocs)

[Khóa học đầy đủ về Python - FreeCodeCamp trên YouTube](https://www.youtube.com/watch?v=rfscVS0vtbw)

[Giới thiệu về Python cho CS và Khoa học dữ liệu](https://drive.google.com/file/d/1rXkYFjw1iKbXCra_B4Ykm0AMRgo6v93w/view?fbclid=IwAR2lg9omGaAsG3g1ZhHQHja8_uxkZ7QddnOUSxfoceRXShU1V_bl4V63xCQ)

[Nhiều hơn nữa trong OOP](https://www.futurelearn.com/courses/object-oriented-principles)

4. Pandas

[Corey Schafer-Youtube](https://www.youtube.com/watch?v=ZyhVh-qRZPA&list=PL-osiE80TeTsWmV9i9c58mdDCSskIFdDS)

[Kaggle](https://www.kaggle.com/learn/pandas)

[Tài liệu](https://pandas.pydata.org/pandas-docs/version/0.15/tutorials.html)

[Trường dữ liệu-Youtube](https://www.youtube.com/watch?v=yzIMircGU5I&list=PL5-da3qGB5ICCsgW1MxlZ0Hq8LL5U3u9y&index=1)

[Khóa học tiếng Ả Rập](https://www.youtube.com/watch?v=3ISW655DemU&list=PLvLvlVqNQGHCb2_ygmr1DQOMOv0yXp84F)

5. Numpy

[Kaggle](https://www.kaggle.com/legendadnan/numpy-tutorial-for-beginners-data-science)

[Khóa học tiếng Ả Rập](https://www.youtube.com/watch?v=5-5CrLmf2vk&list=PLIA_seGogbkGDYq-dnVCsELEIq_7HK7Ca)

[Hướng dẫn](http://cs231n.github.io/python-numpy-tutorial/)

[Tài liệu](https://numpy.org/doc/1.18/user/quickstart.html)

6. Scipy

[Hướng dẫn](https://cs231n.github.io/python-numpy-tutorial/#scipy)

[Tài liệu](https://docs.scipy.org/doc/scipy/reference/tutorial/general.html)

7. Làm sạch dữ liệu: Một trong những kỹ năng quan trọng NHẤT mà bạn cần thành thạo để trở thành một nhà khoa học dữ liệu giỏi, bạn cần thực hành trên nhiều bộ dữ liệu để thành thạo nó.

[Đọc này](https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4)

[Khóa 1](https://www.datacamp.com/courses/cleaning-data-in-python)

[Sổ tay 1](https://www.kaggle.com/bandiatindra/telecom-churn-prediction)

[Sổ tay 2](https://drive.google.com/drive/folders/1OQAEQ8rC4j6oBP7AyDU4bKpPr8sSStJI?fbclid=IwAR2dSrbyoZLM-Wm57yEYy8L8PmpPV9hqXdkNf-pURJC5C5xCz7UJB4YpJ7M)

[Sổ tay 3](https://www.kaggle.com/ashishg21/data-cleaning-and-some-analysis-shoe-prices)

[Làm sạch dữ liệu Kaggle](https://www.kaggle.com/learn/data-cleaning)

8. Trực quan hóa dữ liệu 📊

[Giới thiệu về Trực quan hóa dữ liệu bằng Matplotlib](https://app.datacamp.com/learn/courses/introduction-to-data-visualization-with-matplotlib?fbclid=IwAR1OrJSdZ2LVD_c1o3d-_1I7Nhq8OZ3pzTu4010E_XWEmMc0KYsTosz8CIU)

[Corey Schafer - Danh sách phát trên Youtube](https://www.youtube.com/watch?v=UO98lJQ3QGI&list=PL-osiE80TeTvipOqomVEeZ1HRrcEvtZB_)

[Sentdex - Danh sách phát trên YouTube](https://www.youtube.com/watch?v=q7Bo_J8x_dw&list=PLQVvvaa0QuDfefDfXb9Yf0la1fPDKluPF)

[Kaggle trực quan hóa dữ liệu với Seaborn](https://www.kaggle.com/learn/data-visualization)

[Danh sách phát-Youtube](https://www.youtube.com/watch?v=z7ZINBk8EUk&list=PL998lXKj66MpNd0_XkEXwzTGPxY2jYM2d)

[Khóa 1: Giới thiệu về trực quan hóa dữ liệu với Seaborn](https://learn.datacamp.com/courses/introduction-to-data-visualization-with-seaborn)

[Khóa 2: Trực quan hóa dữ liệu trung cấp với Seaborn](https://learn.datacamp.com/courses/intermediate-data-visualization-with-seaborn)

[Khóa 3: Tìm hiểu và trực quan hóa bằng Python](https://www.coursera.org/learn/understanding-visualization-data)

9. EDA Lưu ý: nó đã được đề cập trong khóa học xác suất ở trên

[DataCamp-EDA bằng Python](https://learn.datacamp.com/courses/exploratory-data-analysis-in-python)

[IBM-EDA dành cho học máy](https://www.coursera.org/learn/ibm-exploratory-data-analysis-for-machine-learning)

10. Bảng điều khiển

Power BI

[Power BI - Youtube (Alex)](https://youtube.com/playlist?list=PLUaB-1hjhk8HqnmK0gQhfmIdCbxwoAoys&si=pR4VSrR1P2O-AaBJ)

[Đào tạo Power BI](https://powerbi.microsoft.com/en-us/learning/)

[Tiếng Ả Rập - Youtube (Zanoon)](https://www.youtube.com/watch?v=ykvAWKML9Gk&list=PLof3yw6ZFPFhV75Ptf-5Q88bgUtLOBvOw)

[Tiếng Ả Rập - Youtube](https://www.youtube.com/watch?v=ykvAWKML9Gk&list=PLof3yw6ZFPFhV75Ptf-5Q88bgUtLOBvOw)

Tableau

[Hướng dẫn](https://www.datacamp.com/community/tutorials/data-visualisation-tableau)

[Tài liệu](https://www.tableau.com/learn/training/20201)

K[hóa học - datacamp](https://learn.datacamp.com/courses/introduction-to-tableau)

[Đơn giản hóa - Youtube](https://learn.datacamp.com/courses/introduction-to-tableau)

11. SQL và cơ sở dữ liệu

SQL để phân tích dữ liệu ([Udacity](https://www.udacity.com/course/sql-for-data-analysis--ud198)-notes [l📋l](https://github.com/julianjohannesen/Udacity-SQL-Notes/tree/main)  hoặc [simplilearn](https://www.simplilearn.com/free-online-course-to-learn-sql-basics-skillup))

[Giới thiệu về SQL](https://learn.datacamp.com/courses/introduction-to-sql) hoặc [IBM (SQL dành cho khoa học dữ liệu)](https://www.coursera.org/learn/sql-data-science)

[Giới thiệu về Cơ sở dữ liệu quan hệ trong SQL](https://learn.datacamp.com/courses/introduction-to-relational-databases-in-sql)

[Khóa học tiếng Ả Rập](https://www.youtube.com/watch?v=B7evUQGmN6M&list=PLfM2wZNebA2zROxUcAbGxNrpVZncsF3oD)

[Tiếng Ả Rập -ITI của Eng.Ramy](https://www.youtube.com/watch?v=QJHy89zMvj8&list=PLYpJKvLDuJhgMzOXRwUJ2_ZlVt3zSh8PA&index=1) Advanced - [[Tài liệu khóa học]](https://drive.google.com/drive/folders/1P_uUWl6pBPC3MFrqR_ovoJQNCZhi9rAS?fbclid=IwAR1EyxPR3pLXABIcK2bkgh4_HYNSowb01Kij8OiHFh_x5maiqsiwvJNHvXg)

[Khoa học dữ liệu 365 - SQL](https://mega.nz/folder/wswGEIhb#tsqUggTZyfy5HyRWUkV9sg/folder/R1AxXCxB)

Thực hành [HackerRank](https://www.hackerrank.com/domains/sql) & [Data Lemur](https://datalemur.com/)

12. Biểu thức chính quy Python

[Hướng dẫn](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial)

13. Phân tích chuỗi thời gian

[Bài hát](https://learn.datacamp.com/skill-tracks/time-series-with-python)

[Sách](https://www.oreilly.com/library/view/practical-time-series/9781492041641/?fbclid=IwAR20cq7hAdWf6voOd61u-pNzZCHvB0rZhT_BUoGTAXxPBhhi82p8BhxLEsI)

[fbprophet](https://facebook.github.io/prophet/docs/quick_start.html)

Nguồn tiếng Ả Rập [Video 1](https://www.youtube.com/watch?v=TvhaHPq6xLU&list=TLPQMjYwNzIwMjEPGXX6392WJA&index=1) & [Video 2](https://www.youtube.com/watch?v=mipF7mRVpk0&list=TLPQMjYwNzIwMjEPGXX6392WJA&index=2)

Khi kết thúc giai đoạn Sơ cấp, hãy áp dụng tất cả những gì bạn đã học được vào một dự án.

**Trung cấp**

1. Toán học ML: gồm Đại số tuyến tính, Giải tích và PCA.

[Toán học cho máy học và khoa học dữ liệu - Andrew Ng](https://www.coursera.org/specializations/mathematics-for-machine-learning-and-data-science?irclickid=zzy33K1O0xyNUAmxqWUjDwedUkAUtSWUJXKyTY0&irgwc=1&utm_medium=partners&utm_source=impact&utm_campaign=3117765&utm_content=b2c#courses)

[Chuyên môn](https://www.coursera.org/specializations/mathematics-machine-learning)

[Toán học cho Machine Learning - Hầu hết các kiến thức cơ bản cần thiết](https://www.youtube.com/watch?v=vLJcduC4lBM&list=PLcQCwsZDEzFmlSc6levE3UV9rZ8yY-D_7)

🔹Đại số tuyến tính

[Khan Academy - Đại số tuyến tính](https://www.khanacademy.org/math/linear-algebra)

[Toán học cho máy học: Đại số tuyến tính](https://www.coursera.org/learn/linear-algebra-machine-learning)

[3Blue1Brown - Tinh hoa của đại số tuyến tính](https://www.3blue1brown.com/topics/linear-algebra)

🔹Tính toán

[Phép tính đa biến - Coursera](https://www.coursera.org/learn/multivariate-calculus-machine-learning?fbclid=IwAR243aoz0jxs4iUn539pnjSQliXtr7Y5QAsvgeRTietZT_tkyoRU3b6Sq1o)

[Bản chất của phép tính - Youtube](https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr)

🔹PCA

[PCA - Coursera](https://www.coursera.org/learn/pca-machine-learning)

2. Học máy

[Coursera - Khóa học cũ của Andrew Ng (Octave/Matlab)](https://www.coursera.org/learn/machine-learning)

[Chuyên ngành ML mới của Coursera Andrew (Python)](https://www.coursera.org/specializations/machine-learning-introduction?_hsenc=p2ANqtz-_R9x3Nm07uCw6YAw9VpCpdjRdfaUFyxdOcvgDljRt7j_NXiahN1plnI_Ob9jn0jSNipuE_Y08llrfPSt_1P7EBvj4LuImpBTKG3bsR6Z9bzjzBoRY&_hsmi=216611333&action=enroll&utm_campaign=mls-launch-2022&utm_content=216613012&utm_medium=email&utm_source=hs_email#courses)

[Học máy - StatQuest - YouTube](https://www.youtube.com/playlist?list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF)

[Khóa học đầy đủ về Machine Learning Stanford trên YouTube của Andrew](https://www.youtube.com/watch?v=PPLop4L2eGk&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN)

[CS480/680 Giới thiệu về Machine Learning - Mùa xuân 2019 - Đại học Waterloo](https://www.youtube.com/playlist?list=PLdAoL1zKcqTW-uzoSVBNEecKHsnug_M0k)

[SYDE 522 – Trí tuệ máy (Mùa đông 2018, Đại học Waterloo)](https://www.youtube.com/playlist?list=PL4upCU5bnihwCX93Gv6AQnKmVMwx4AZoT)

[Giới thiệu khóa học Machine Learning - Udacity](https://www.udacity.com/course/intro-to-machine-learning--ud120)

[Hesham Asem - nội dung tiếng Ả Rập](https://www.youtube.com/c/HeshamAsem/playlists)

[IBM ML với Python](https://www.coursera.org/learn/machine-learning-with-python)

[Học máy từ đầu - YouTube (Kỹ sư Python)](https://www.youtube.com/watch?v=ngLyX54e1LU&list=PLqnslRFeH2Upcrywf-u2etjdxxkL8nl7E)

Thực hành Phiên bản ML ([1](https://drive.google.com/file/d/1uro1p6SlYolSkF0fbFKau0pOQ9ENZqny/view?usp=sharing) & [2](https://drive.google.com/file/d/1rS95FTNfiVG4WjGnPjd73GqrmEKey4N1/view?usp=sharing) & [3](https://drive.google.com/file/d/11VeqPJw8s9SC9Ru7IVeQhiTyV_9TliOE/view?usp=sharing)) | mã ví dụ ['Sổ tay'](https://github.com/ageron/handson-ml?fbclid=IwAR3s31KlwkLKyrEwuEd4UMOcvHN1Q9Z2LLGzPg5vP4UKSwjriHxU0uO405c)

[Thuật toán ML trong thực tế](https://www.coursera.org/specializations/machine-learning-algorithms-real-world?utm_medium=email&utm_source=marketing&utm_campaign=A39CcMUuEempyReieZALEQ)

[Nhà khoa học ML](https://learn.datacamp.com/career-tracks/machine-learning-scientist-with-python?version=1)

[Dự án](https://www.coursera.org/learn/applied-data-science-capstone)

3. Quét web/API

[Khóa học](https://learn.datacamp.com/courses/web-scraping-with-python)

[Giới thiệu 2](https://www.dataquest.io/blog/web-scraping-tutorial-python/)

[Hướng dẫn](https://realpython.com/beautiful-soup-web-scraper-python/)

[Sách cho cả 2 chủ đề](https://b-ok.africa/book/3515980/5d50aa)

API

[Hướng dẫn](https://www.dataquest.io/blog/python-api-tutorial/)

[Bài viết](https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fhow-to-pull-data-from-an-api-using-python-requests-edcc8d6441b1)

[Hướng dẫn](https://rapidapi.com/blog/how-to-use-an-api-with-python/)

4. Số liệu thống kê.

[Số liệu thống kê này - Sách](https://greenteapress.com/thinkstats/thinkstats.pdf)

[Think Bayes - Sách](https://www.greenteapress.com/thinkbayes/thinkbayes.pdf)

5. SQL nâng cao

[Nối dữ liệu trong SQL - DataCamp](https://learn.datacamp.com/courses/joining-data-in-postgresql)

[SQL trung cấp - DataCamp](https://learn.datacamp.com/courses/joining-data-in-postgresql)

[SQL nâng cao hơn](https://www.coursera.org/lecture/data-driven-astronomy/more-advanced-sql-GDmo5)

6. Kỹ thuật tính năng

[Hướng dẫn](https://www.kaggle.com/learn/feature-engineering)

[Bài viết](https://www.medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Ffeature-engineering-for-machine-learning-3a5e293a5114)

[Sách](https://drive.google.com/file/d/1BkJYO0tqMYptTWUDQ7X0vd2aygohHRm8/view?usp=sharing)

7. Interpet giải thích dựa trên Shapley về các mô hình ML.

[SHAP](https://www.kaggle.com/learn/feature-engineering)

[Khả năng giải thích của Kaggle ML](https://www.kaggle.com/learn/machine-learning-explainability)

Sau khi hoàn thành cấp độ này, hãy áp dụng cho 2 hoặc 3 dự án có quy mô tốt.

Vui lòng đọc cuốn sách này [📖 Giới thiệu về Học thống kê với các ứng dụng trong R](https://github.com/Moataz-Elmesmary/Data-Science-Roadmap/blob/main/Introduction%20to%20Statistical%20Learning%20with%20Applications%20in%20R.pdf)

**Nâng cao**

1. Học sâu

[Nguyên tắc cơ bản về học tập sâu](https://www.youtube.com/playlist?list=PLZbbT5o_s2xq7LwI2y8_QtvuXZedL6tQU)

[Giới thiệu về Học sâu - MIT](http://introtodeeplearning.com/?fbclid=IwAR35rIygYlCn84DV7mlHvdvs4sMUm2D6RLYVwFpp2nT2t1Zj1GGy3QAWQvQ)

[Chuyên môn](https://www.coursera.org/specializations/deep-learning)

[Đi sâu vào Học sâu (En)](https://d2l.ai/d2l-en.pdf?fbclid=IwAR0sVdA8VFYpNZCpYZHgo_kl_HYrjcjDfjEka26D8xRWAhbhh6mmSNIXg3U) | (Ar) phiên bản [➡️Phần 1](https://drive.google.com/file/d/1SrmT_r8dNK42IqyS0gwXtbLCZbk5G8eu/view?fbclid=IwAR1Xcf8PNKkPJMg0uHRE1QyIW4_BMxISIdoB8pPaepw38njhaIf04MYM218) & [Phần 2](https://drive.google.com/file/d/1UqEu0amRfAvJD0L1HosIn3UJi0FkNemU/view?fbclid=IwAR1og8pkWr1gT3jdUwqikCZVrOCpyrm0x6ZRL63Kitwhki35pazHdo_ScJI)

[Học sâu UC Berkeley](https://www.youtube.com/playlist?list=PLZSO_6-bSqHQHBCoGaObUljoXAyyqhpFW)

[Github của Đi sâu vào DL](https://github.com/d2l-ai/d2l-en?fbclid=IwAR0QN35b-NHHWq_zKISA1cbI063aRqqoKqR_0e3cpnT5h58GkcNbCIJs3iw)

[Bài giảng của Stanford - Mạng thần kinh tích chập để nhận dạng hình ảnh](https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv)

[Đại học Waterloo - ML/DL](https://www.youtube.com/playlist?list=PLdAoL1zKcqTW-uzoSVBNEecKHsnug_M0k)

2. Tensorflow

[Chuyên môn](https://www.coursera.org/specializations/tensorflow-in-practice)

[Youtube](https://www.youtube.com/playlist?list=PLZbbT5o_s2xrwRnXk_yCPtnqqo4_u2YGL)

[Các khóa học sâu của fast.ai](https://www.fast.ai/)

TensorFlow đánh bại PyTorch về khả năng trực quan hóa và triển khai các mô hình đã được đào tạo. Hãy sử dụng PyTorch nếu bạn muốn tính linh hoạt, khả năng sửa lỗi và thời gian đào tạo ngắn.

3. PyTorch

[PyTorch (UC Berkeley - Youtube) - Lec3 (5 phần)](https://m.youtube.com/watch?v=AOypIa_8RXg&list=PL_iWQOsE6TfXxKgI1GgyV1B_Xa0DxE5eH&index=11)

[PyTorch - Tiến sĩ Khoa học Dữ liệu - Youtube](https://www.youtube.com/watch?v=vVQs4h6HUvA&list=PLLeO8f6PhlKb_FAC7qxOBtxT9-8EPDAqk)

[Hướng dẫn sử dụng Pytorch - Aladdin - Youtube](https://www.youtube.com/watch?v=2S1dgHpqCdk&list=PLhhyoLH6IjfxeoooqP9rhU3HJIAVAJ3Vz)

[Khóa học PyTorch (2022) - Youtube](https://www.youtube.com/watch?v=v43SlgBcZ5Y&list=PLkdGijFCNuVk9fO1IMfdV1Igob0FUHhkB)

[Học sâu với Pytorch](https://drive.google.com/file/d/1-KG_ufeg7zw2iLgG5RrJSFpyonLwulgF/view?usp=sharing)

[Học máy với PyTorch và Scikit-Learn -2022](https://dl.ebooksworld.ir/books/Machine.Learning.with.PyTorch.and.Scikit-Learn.Sebastian.Raschka.Packt.9781801819312.EBooksWorld.ir.pdf)

4. Khoa học dữ liệu nâng cao

[Khoa học dữ liệu nâng cao với chuyên ngành IBM](https://www.coursera.org/specializations/advanced-data-science-ibm)

5. NLP

[Chuyên ngành - Coursera](https://www.coursera.org/specializations/natural-language-processing)

[Tiếng Ả Rập - Ahmed El Sallab](https://www.youtube.com/playlist?list=PLxmZ0b-n395VxzEUL8-Dy257zSqYZe4yU)

[Stanford CS224N - YouTube](https://www.youtube.com/playlist?app=desktop&list=PLoROMvodv4rMFqRtEuo6SGjY4XbRIVRd4&si=k91y-bepIiPjHMrj&fbclid=IwAR2h6KcYboHCjG9YBIEB08srgYSesqZ5UHXr0ni8yxOqrxNV3-_TGxq0Csg)

[Giới thiệu về Xử lý ngôn ngữ tự nhiên trong Python](https://www.datacamp.com/courses/natural-language-processing-fundamentals-in-python)

🔸LLMS - [Mô hình ngôn ngữ lớn là gì?](https://www.snowflake.com/guides/what-large-language-model-and-what-can-llms-do-data-science)

[AI sáng tạo với LLM](https://www.deeplearning.ai/courses/generative-ai-with-llms/)

6. Thống kê suy luận

[Chuyên ngành, khóa 2 & 3](https://www.coursera.org/specializations/statistics-with-python)

[Khóa học](https://www.coursera.org/learn/statistical-inferences)

7. Thống kê Bayes

[1 - Từ ý tưởng đến phân tích dữ liệu](https://www.coursera.org/learn/bayesian-statistics)

[2 - Kỹ thuật và mô hình](https://www.coursera.org/learn/mcmc-bayesian-statistics)

[3 - Mô hình hỗn hợp](https://www.coursera.org/learn/mixture-models)

8. Triển khai mô hình

[Hướng dẫn sử dụng Flask](https://towardsdatascience.com/deploying-a-deep-learning-model-using-flask-3ec166ef59fb)

[TensorFlow: Chuyên môn về dữ liệu và triển khai](https://www.coursera.org/specializations/tensorflow-data-and-deployment)

[Triển khai các mô hình với TensorFlow Serve và Flask](https://www.coursera.org/projects/deploy-models-tensorflow-serving-flask)

[Cách triển khai Mô hình học máy lên Google Cloud - Daniel Bourke](https://www.youtube.com/watch?v=fw6NMQrYc6w)

Nếu bạn quan tâm đến nhiều phương pháp triển khai hơn, hãy tìm kiếm (FastAPI - Heroku - chitra)

9. MLOps : là sự kết hợp giữa Triển khai mô hình, Phục vụ mô hình, Giám sát mô hình và Bảo trì mô hình.

[MLOps-zoomcamp](https://github.com/DataTalksClub/mlops-zoomcamp)

[Hướng dẫn MLOps](https://github.com/Nyandwi/machine_learning_complete/blob/main/010_mlops/1_mlops_guide.md)

[MLO thực tế](https://drive.google.com/file/d/17RhXQ2ix6rFMaas3HI7bnM_GL8lS7u3f/view?usp=sharing)

10. Mô hình đồ họa xác suất

[Chuyên môn](https://www.coursera.org/specializations/probabilistic-graphical-models)

🌟 Hãy đọc những cuốn sách này, chúng sẽ có ích cho bạn.

[Lý luận Bayes và học máy](https://drive.google.com/file/d/18fh0orqSNAaIyhLkVwh9cGuWBywCBbuw/view?usp=sharing)

[Các yếu tố của việc học thống kê](https://drive.google.com/file/d/1ePRkuB9Zm5Fkw-1-VG8prQXfj8pI6dWX/view?usp=sharing)

[Nhận dạng mẫu và học máy - Bishop (Nâng cao)](https://drive.google.com/file/d/1QkQj_azL6O7qUzshB8lPzueYWj0TRwEu/view?usp=sharing)

**Dự Án**

[Deena Gergis - Dự án cuối cùng](https://www.youtube.com/playlist?list=PLatl6hdtJ0RnbkReSAuel6PeCPO155FpG)

[Dự án học máy - Youtube](https://www.youtube.com/watch?v=fiz1ORTBGpY&list=PLfFghEzKVmjvuSA67LszN1dZ-Dd_pkus6)

[10 dự án khoa học dữ liệu hàng đầu dành cho người mới bắt đầu](https://www.kdnuggets.com/2021/06/top-10-data-science-projects-beginners.html)

[12 Dự án Khoa học Dữ liệu dành cho Người mới bắt đầu và Chuyên gia](https://builtin.com/data-science/data-science-projects)

[Ý tưởng và dự án khoa học dữ liệu](https://nevonprojects.com/data-science-projects-solutions/)

[Hơn 310 dự án học máy hàng đầu năm 2023](https://data-flair.training/blogs/machine-learning-project-ideas/)

[10 Dự án khoa học dữ liệu có hướng dẫn từ đầu đến cuối](https://pub.towardsai.net/10-end-to-end-guided-data-science-projects-to-build-your-portfolio-b7b9047fe6c9)

[Hướng dẫn ML trong thế giới thực với Scikit Learn](https://www.youtube.com/watch?v=M9Itm95JzL0)

[Kết thúc dự án ML với Docker, hành động và triển khai Github](https://www.youtube.com/watch?v=MJ1vWb1rGwM)

[Mã Python trong Khoa học dữ liệu](https://github.com/RubensZimbres/Repo-2017/)

[12 dự án Khoa học dữ liệu miễn phí để thực hành Python và Pandas (giải quyết tương tác trực tuyến)](https://www.datawars.io/articles/12-free-data-science-projects-to-practice-python-and-pandas)

**More - More - More**

[65 cuốn sách quan trọng miễn phí](https://techgrabyte.com/springer-released-65-machine-learning-data-science-books-free/?fbclid=IwAR3x9Mfd4FSrBo_y2F92cAghhZ8851qv3wQ1R_u87s70MZTDtKeG5yn7Hts')

[Toán học cho máy học](https://mml-book.github.io/')

[Giới thiệu về Học Thống kê](https://www.statlearning.com/)

[Tìm hiểu về Machine Learning: Từ lý thuyết đến thuật toán](https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/copy.html)

[Học máy xác suất: Giới thiệu](https://probml.github.io/pml-book/book1.html)

[Kể chuyện bằng dữ liệu](https://drive.google.com/file/d/1OQu6ZWImGnHbuI_WJOLPdSvKWCABSWMH/view?usp=sharing) Hướng dẫn trực quan hóa dữ liệu quan trọng.

**Tuyển tập những cheatsheet hay nhất**

1. [Importing Data](https://lnkd.in/e3jnyTEi)

2. Pandas  [- (1)](https://lnkd.in/eiXuBbWh_)    [- (2)](https://lnkd.in/e8PKwQQQ)    [- (3)](https://lnkd.in/ewQfqe8q)

3. [Matplotlib](https://lnkd.in/ejxbW8ak)

4. [Seaborn](https://lnkd.in/ejhxUp2K)

5. [Probability](https://lnkd.in/e4Jxx6xP)

6. [Supervised Learning](https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/cheatsheet-supervised-learning.pdf)

7. [Unsupervised Learning](https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/cheatsheet-unsupervised-learning.pdf)

8. [Deep Learning](https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/cheatsheet-deep-learning.pdf)

9. [Machine Learning Tips and Tricks](https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/cheatsheet-machine-learning-tips-and-tricks.pdf)

10. [Probabilities and Statistics](https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/refresher-probabilities-statistics.pdf')

11. [Comprehensive Stanford Master Cheat Sheet](https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/super-cheatsheet-machine-learning.pdf)

12. [Linear Algebra and Calculus](https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/refresher-algebra-calculus.pdf)

13. [Data Science Cheat Sheet](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PythonForDataScience.pdf)

14. [Keras Cheat Sheet](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Keras_Cheat_Sheet_Python.pdf)

15. [Deep Learning with Keras Cheat Sheet](https://github.com/rstudio/cheatsheets/raw/master/keras.pdf)

16. [Visual Guide to Neural Network Infrastructures](http://www.asimovinstitute.org/wp-content/uploads/2016/09/neuralnetworks.png)

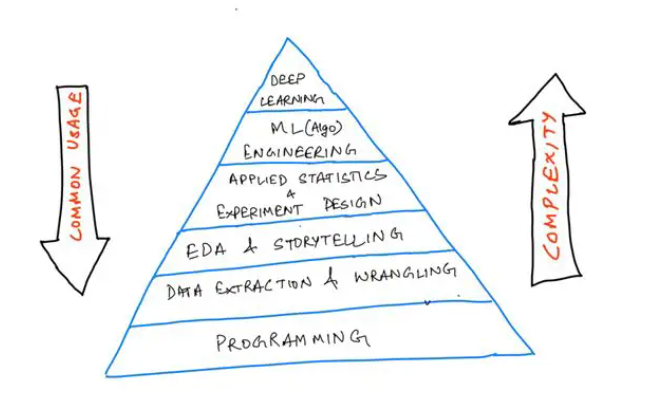
17. [Skicit-Learn Python Cheat Sheet](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Scikit_Learn_Cheat_Sheet_Python.pdf)

18. [Scikit-learn Cheat Sheet: Choosing the Right Estimator](https://scikit-learn.org/stable/tutorial/machine_learning_map/)

19. [Tensorflow Cheat Sheet](https://github.com/kailashahirwar/cheatsheets-ai/blob/master/PDFs/Tensorflow.pdf)

20. [Machine Learning Test Cheat Sheet](https://www.cheatography.com/lulu-0012/cheat-sheets/test-ml/pdf/)

21. [Machine Learning Cheat Sheets (Recommended Guide)](https://drive.google.com/file/d/1rQRJvWk5s9rUCesri0apxutbF4eDHR69/view?usp=sharing)



- Biểu đồ mô tả lộ trình dành cho Data Science

Hoặc bạn có thể tham khảo lộ trình biểu đồ này.

