NHẬP MÔN KHOA HỌC DỮ LIỆU

GIỚI THIỆU MÔN HỌC

ThS. Vũ Hoài Thư

Nội dung

Thông tin môn học

Thông tin liên quan

- Giảng viên:
  - ThS. Vũ Hoài Thư, Bộ môn Khoa học máy tính, Khoa CNTT1
  - Email: thuvh@ptit.edu.vn
- Thời lượng môn học:
  - 03 tín chỉ = 45 giờ
  - 34 giờ lý thuyết, 10 giờ bài tập, 1 giờ thảo luận

Tài liệu tham khảo

[1]. Joel Grus. Data Science from Scratch: First Principles with Python. O'Reilly Media, 2nd edition, 2019.

[2]. Jake VanderPlas. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, 2017.

Một số quy định môn học

- Không nói chuyện riêng trong lớp, ảnh hưởng tới Giáo viên và không khí lớp học.
- Khi kiểm tra, thi cử:
  - Không gian lận, copy bài nhau. Nếu vi phạm sẽ nhận 0 điểm cho tất cả các bài kiểm tra liên quan.
  - Nếu bài thi/ bài kiểm tra cho phép sử dụng tài liệu, tra cứu Internet, KHÔNG ĐƯỢC PHÉP sử dụng các phần mềm chat, truyền tải văn bản, hình ảnh, video.
  - **Không xin điểm**, nộp bài trễ deadline.
  - Không phát tán bài kiểm tra cho các lớp khác.
  - Không đăng tải hình ảnh của lớp học lên mạng xã hội.
  - *Khi vi phạm các lỗi trên sẽ nhận 0 điểm hoặc chia đôi TẤT CẢ các đầu điểm thành phần.*
- Thường xuyên theo dõi thông tin trên nhóm lớp và các thông báo từ lớp trưởng.

Đánh giá môn học

Cách đánh giá:

- Chuyên cần (10%): Điểm danh 10 buổi
  - Vắng 1 buổi = -1 điểm
  - Muộn = Vắng
  - Vắng >= 5 buổi = 0 điểm chuyên cần.
- Bài tập (20%): Làm nhóm khi kết thúc chương (Lập trình + Thuyết trình)
- Kiểm tra giữa kỳ (10%): Trắc nghiệm, làm cá nhân
- Thi cuối kỳ (60%): Tự luận.

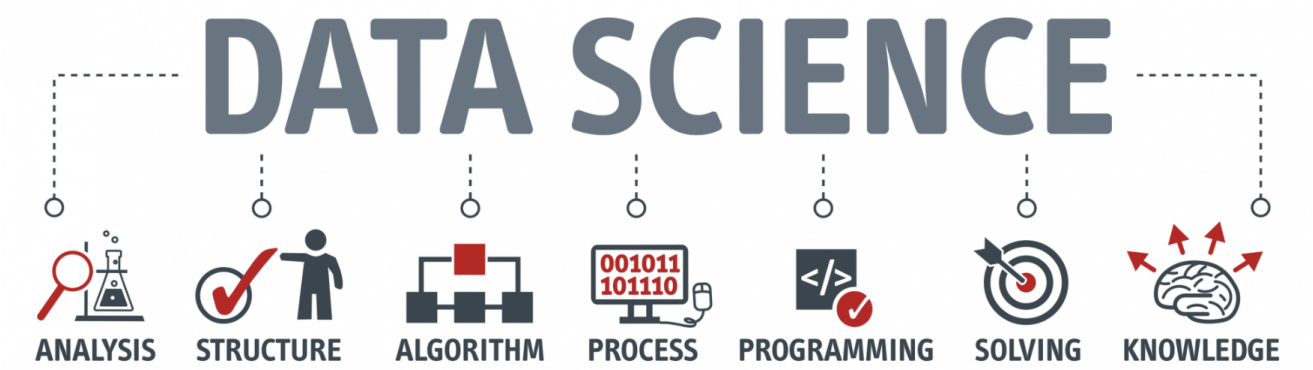
Lưu ý: Thiếu điểm thành phần hoặc nghỉ quá 20% số buổi sẽ không được thi hết môn!!!

Tổng quan về khoa học dữ liệu

Tổng quan về khoa học dữ liệu

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) bao gồm ba phần chính:

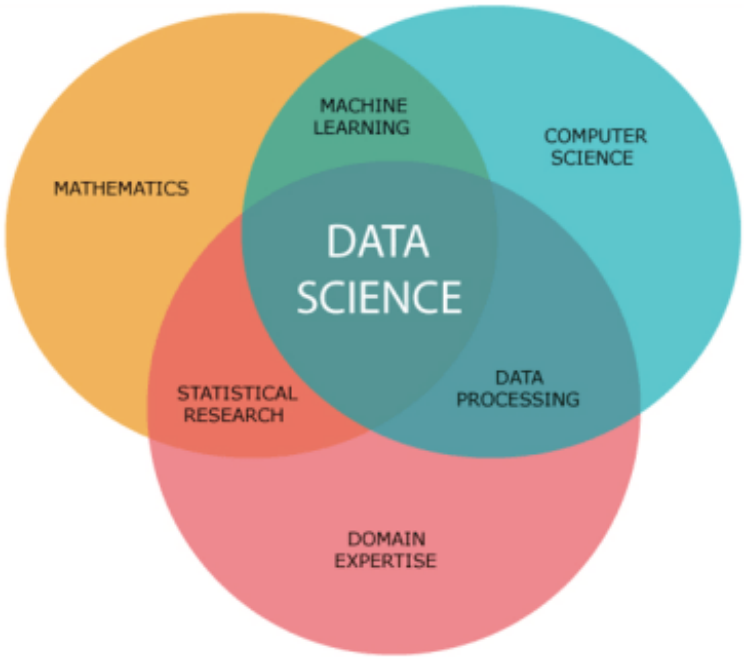
- Tạo và quản trị dữ liệu,
- Phân tích dữ liệu,
- Áp dụng kết quả phân tích cho các hành động có giá trị.



Tổng quan về khoa học dữ liệu

Ngành Khoa học dữ liệu dựa trên ba nguồn tri thức:

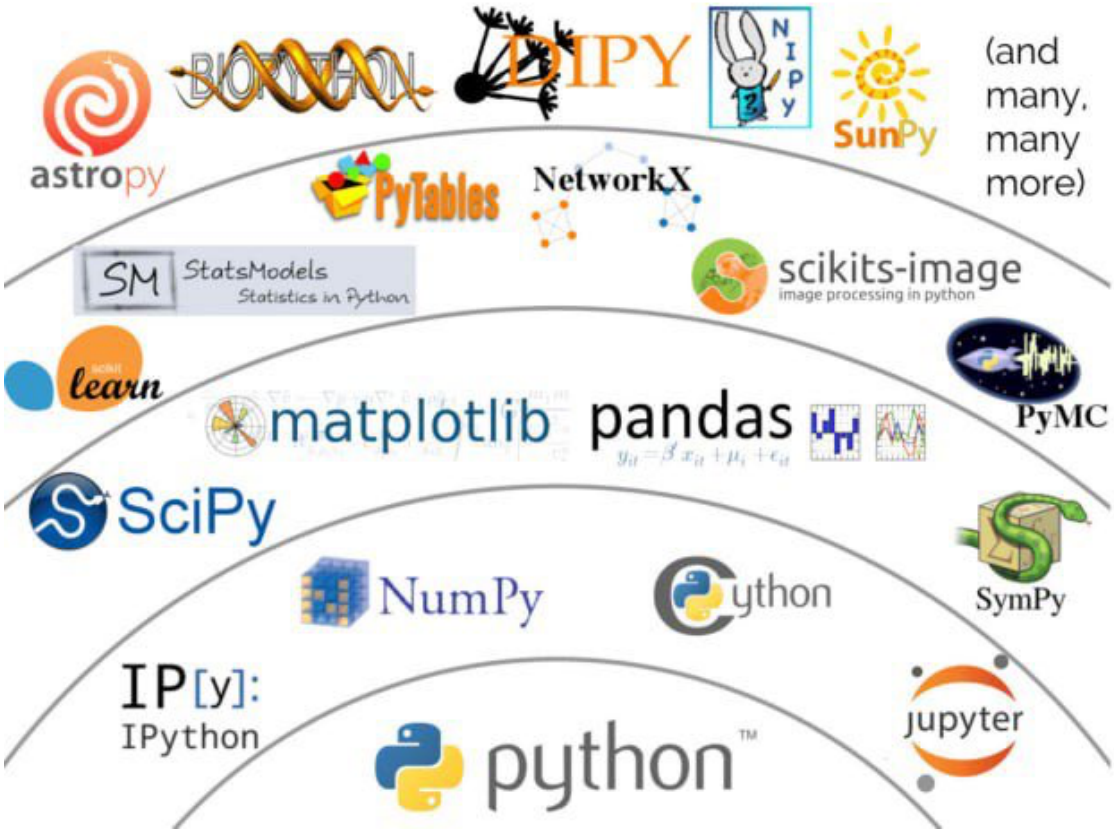
- Toán học (Thống kê toán học - Mathematical Statistics);
- Công nghệ thông tin (Học máy- Machine Learning);
- Tri thức của lĩnh vực ứng dụng cụ thể (Domain Expertise).



Kiến thức yêu cầu

- Biết sử dụng một trong các ngôn ngữ python, C++, Java, C (Trong môn học sẽ sử dụng chủ yếu python)
- Cấu trúc dữ liệu: SQL và NoSQL
- Làm việc với file
- Có kiến thức về các định dạng dữ liệu thường dùng như văn bản, ảnh, âm thanh, phim,..)
- Có kiến thức cơ bản về đại số tuyến tính và xác suất thống kê

Công cụ học tập

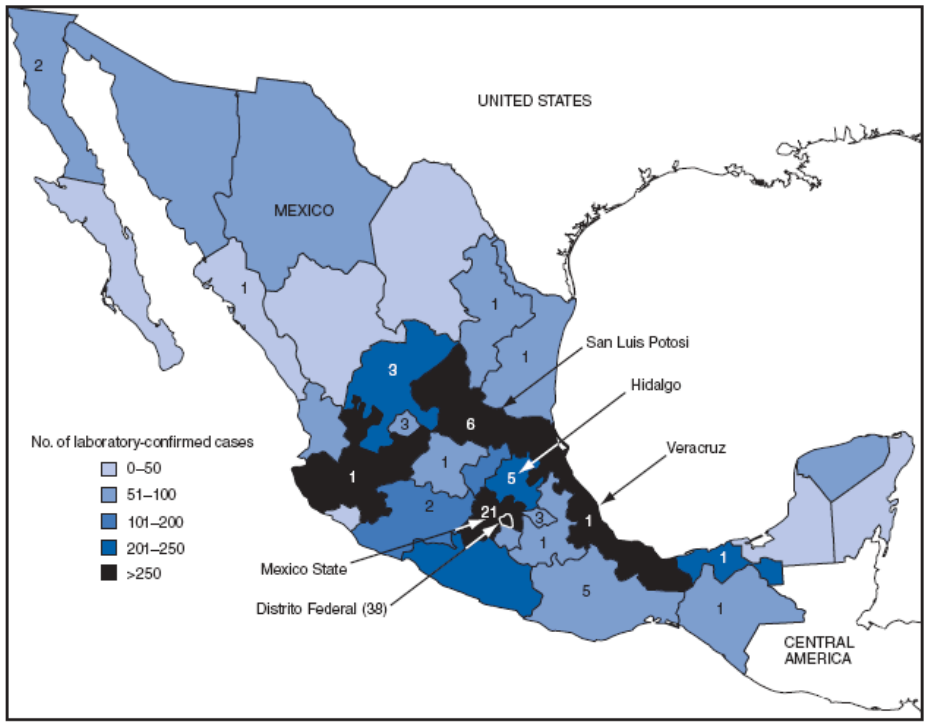


Động lực học tập

- Có kiến thức về khoa học dữ liệu
- Khả năng phân tích dữ liệu
- Hiểu biết về công việc của người làm khoa học dữ liệu và các bài toán liên quan
- Hiểu biết về cách thức áp dụng của KHDL vào các bài toán thực tế
- Có thêm lựa chọn cho nghề nghiệp sau này
- Khả năng học lên cao hơn sau khi tốt nghiệp

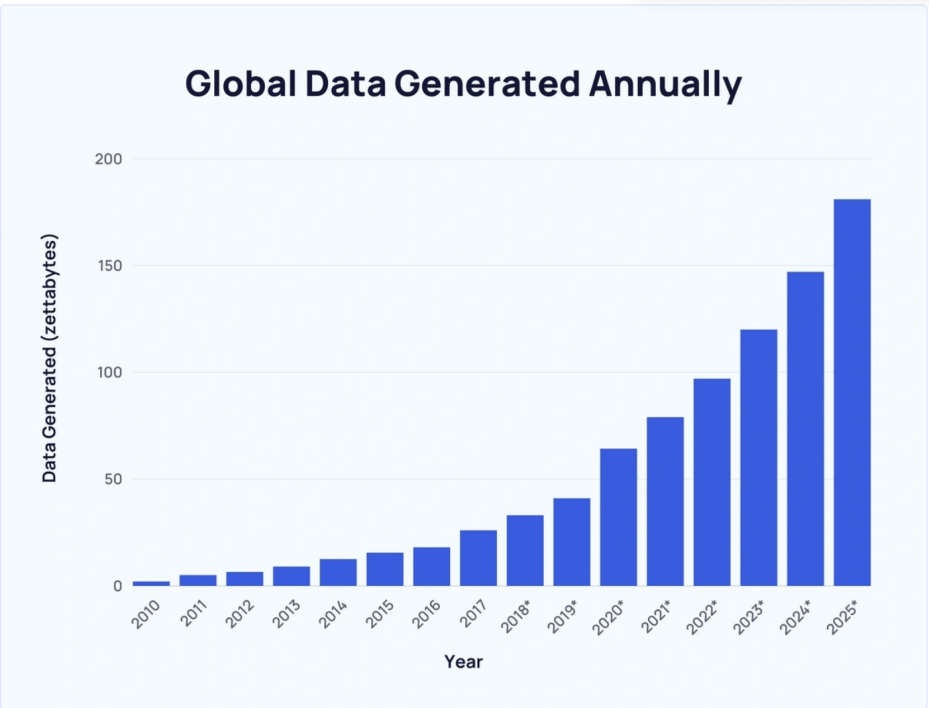
Ý nghĩa của Khoa học dữ liệu

- Google đã dự đoán về sự bùng nổ của bệnh cúm trước 2 tuần.
- Dự báo thành phố nào đang có nguy cơ lây lan virus Ebola



Ý nghĩa của Khoa học dữ liệu

- Phần lớn nỗ lực là chuẩn bị dữ liệu
- Theo kết quả phỏng vấn của các chuyên gia, từ 50% - 80% thời gian của họ trong công việc là thu thập và chuẩn bị dữ liệu trước khi sử dụng các công cụ phân tích.



Nội dung môn học

Nội dung môn học (1/6)

CHƯƠNG 1: KIẾN THỨC CƠ SỞ

- **Ôn tập về đại số tuyến tính**
  - Ma trận và Định thức
  - Ma trận nghịch đảo
- **Ôn tập về xác suất**
  - Sự kiện ngẫu nhiên và quan hệ giữa các sự kiện
  - Biến ngẫu nhiên và phân loại biến ngẫu nhiên
  - Phân phối chuẩn
- **Ôn tập về thống kê**
  - Khoảng tin cậy
  - Kiểm định giả thuyết thống kê

Nội dung môn học (2/6)

CHƯƠNG 2: CHUẨN BỊ DỮ LIỆU

- **Thu thập dữ liệu**
- **Làm sạch dữ liệu**
  - Xử lý giá trị bị thiếu
  - Xử lý giá trị sai hoặc không nhất quán
- **Co giãn và chuẩn hóa dữ liệu**
- **Giảm chiều và biến đổi dữ liệu**
  - Lấy mẫu
  - Lựa chọn đặc trưng
  - Giảm chiều dữ liệu
  - Biến đổi dữ liệu

Nội dung môn học (3/6)

CHƯƠNG 3: TRỰC QUAN HOÁ DỮ LIỆU

- Đồ thị dạng đường thẳng
- Đồ thị điểm rời rạc
- Trực quan hóa lỗi
- Đồ thị đường viền
- Histograms và mật độ
- Văn bản và chú thích
- Đồ thị ba chiều
- Dữ liệu địa lý

Nội dung môn học (4/6)

CHƯƠNG 4: HỌC MÁY

- **Các khái niệm cơ bản**
  - Học và suy diễn
  - Đánh giá mô hình
  - Quá vừa và dưới vừa
  - Độ lệch và phương sai
- **Biến đổi và trích chọn đặc trưng**
  - Đặc trưng phân loại
  - Đặc trưng văn bản
  - Đặc trưng ảnh
- **Các dạng học máy**
  - Học có giám sát
  - Học không giám sát
  - Học bán giám sát
  - Học kết hợp

Nội dung môn học (5/6)

CHƯƠNG 5: CƠ SỞ DỮ LIỆU VÀ SQL

- **Cơ sở dữ liệu cơ bản**
  - CREATE TABLE và INSERT
  - UPDATE
  - DELETE
  - SELECT
  - GROUP BY
  - ORDER BY
  - JOIN
- **Cơ sở dữ liệu nâng cao**
  - Truy vấn con
  - Tối ưu truy vấn
  - NoSQL

Nội dung môn học (6/6)

CHƯƠNG 6: HỆ KHUYẾN NGHỊ

- Giới thiệu
- Lọc dựa trên nội dung
- Lọc cộng tác
- Các hệ khuyến nghị lai
- Các hệ khuyến nghị dựa trên ngữ cảnh - Khuyến nghị theo phiên