Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

BIG DATA IN MACHINE LEARNING

Giảng viên: Khuất Thùy Phương

2020

BIG DATA IN MACHINE LEARNING

- **Hình thức học:** học 2 buổi/tuần
  - **Lịch học : Thứ Bảy & Chủ Nhật** (07h30 – 12h00) + 1 buổi tối
- **Đánh giá môn học: (tổng: 10 điểm)**
  - **1 điểm:** chuyên cần tham gia các buổi học
  - **5 điểm:** hoàn thành bài tập hàng tuần - Upload bài nộp vào thư mục share trên Google Drive: LDS9_HoTen (trong đó có 4 folder Week 1, 2, 3, 4– hạn chót nộp bài mỗi tuần là Thứ Năm hàng tuần)
  - **5 điểm: L**àm project cuối khóa (làm trong 1 tuần)
- **Phụ trách:**
  - **Ms. Phương**
  - Email: tubirona@gmail.com

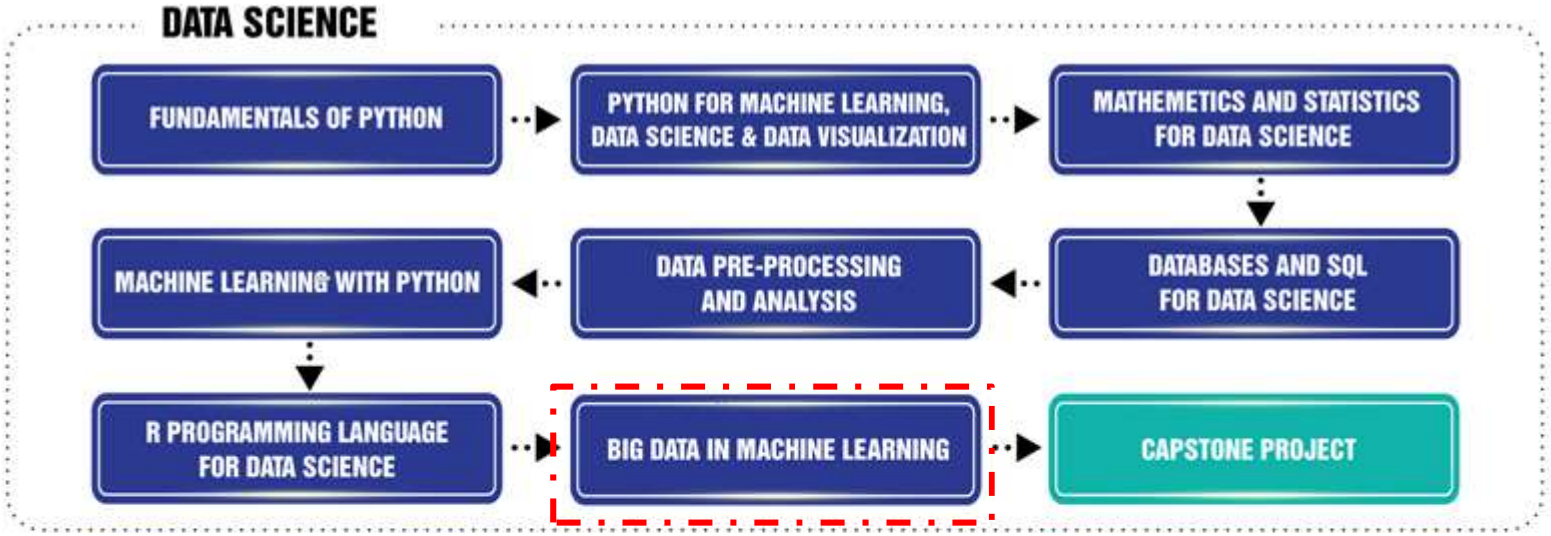
BIG DATA IN MACHINE LEARNING

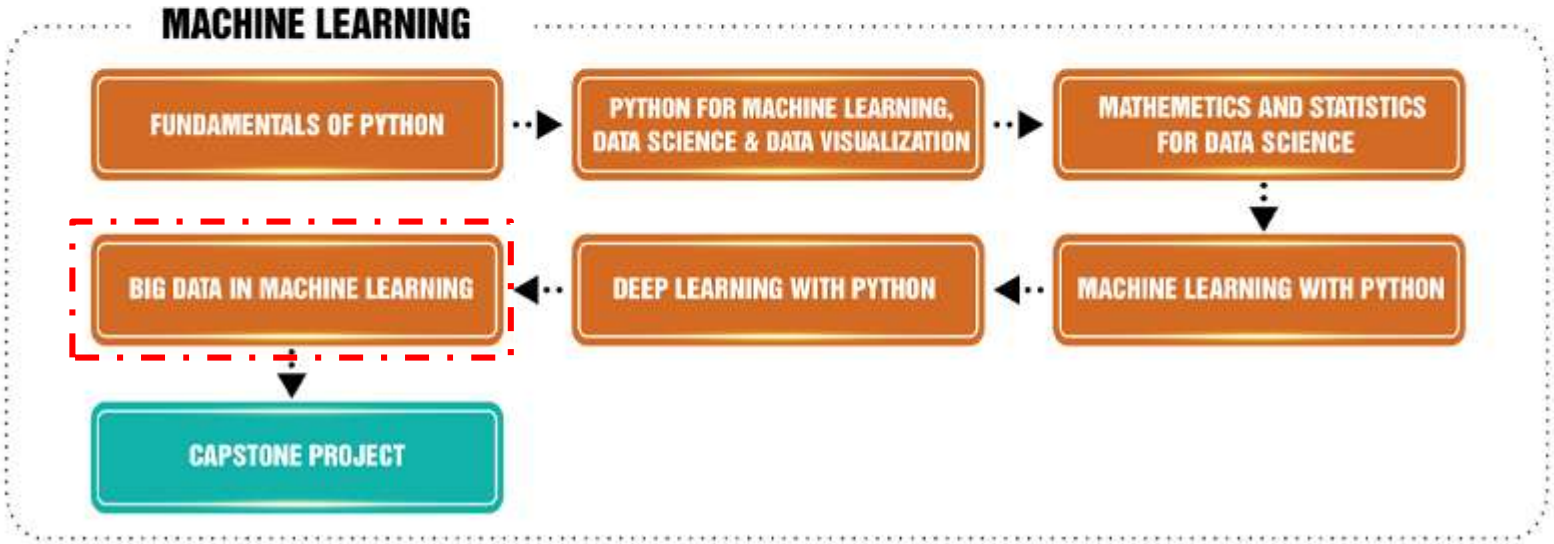
KẾ HOẠCH HỌC VÀ THI LỚP LDS9 - k262

	Tuần 1		Tuần 2		Tuần 3		Tuần 4		Tuần 5		
	Buổi 1	Buổi 2	Buổi 3	Buổi 4	Buổi 5	Buổi 6	Buổi 7	Buổi 8	Buổi 9 (HV chọn học 1 trong 2 buổi **tối thứ 3** hoặc **tối thứ 4**)		**Buổi 10**
Lịch học	5/12/20	6/12/20	12/12/20	13/12/20	19/12/20	20/12/20	26/12/20	27/12/20	29/12/20	30/12/20	**Hoàn chỉnh và nộp bài thi**
Hạn nộp bài	**10/12/2020**		**17/12/2020**		**24/12/2020**		**31/12/2020**				**3/1/2021**
Điểm	**0,5**		**1,5**		**1,5**		**1,5**				**5**
Điểm chuyên cần	1										
Điểm cộng	0 -> 1,5đ										

Note: HV tạo tài khoản developer trên twitter (và tạo app) sớm, link https://developer.twitter.com/en để thực hành lấy dữ liệu

BIG DATA IN MACHINE LEARNING





Nội dung

- Overview of Big Data
- Overview of PySpark
- PySpark RDDs
- PySpark SQL and DataFrame
- Data Preprocessing & Analysis
- Overview of PySpark Mllib

Nội dung

- **Supervised Learning (Classification & Regression)**
  - Linear Regression (pyspark.ml.regression)
  - Logistic Regression (pyspark.ml.classification)
  - Decision Tree (pyspark.ml.classification)
  - Random forest (pyspark.ml.classification)
  - Gradient-Boosted Trees (pyspark.ml.classification)
  - Pipeline

Nội dung

- **Unsupervised Learning (Clustering & Recommender System)**
  - Clustering with K-means (pyspark.ml.clustering)
  - Recommender System (pyspark.ml.recommendation)
  - Association rules – FPGrowth (pyspark.ml.fpm.FPGrowth)
  - Principal Component Analysis PCA

Nội dung

- PySpark Streaming
- Natural Language Processing (NLP)
- PySpark GraphX (cơ bản)
- Apache Spark standalone cluster