



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
 - **Thời gian:** 03/10/2020 -> 31/10/2020
 - **Lịch học :** Thứ Bảy & Chủ Nhật (07h30 – 12h00) + bù 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:** Làm project cuối khóa
- **Phụ trách:**
 - **Ms. Phương**
 - Email: tubirona@gmail.com

BIG DATA IN MACHINE LEARNING



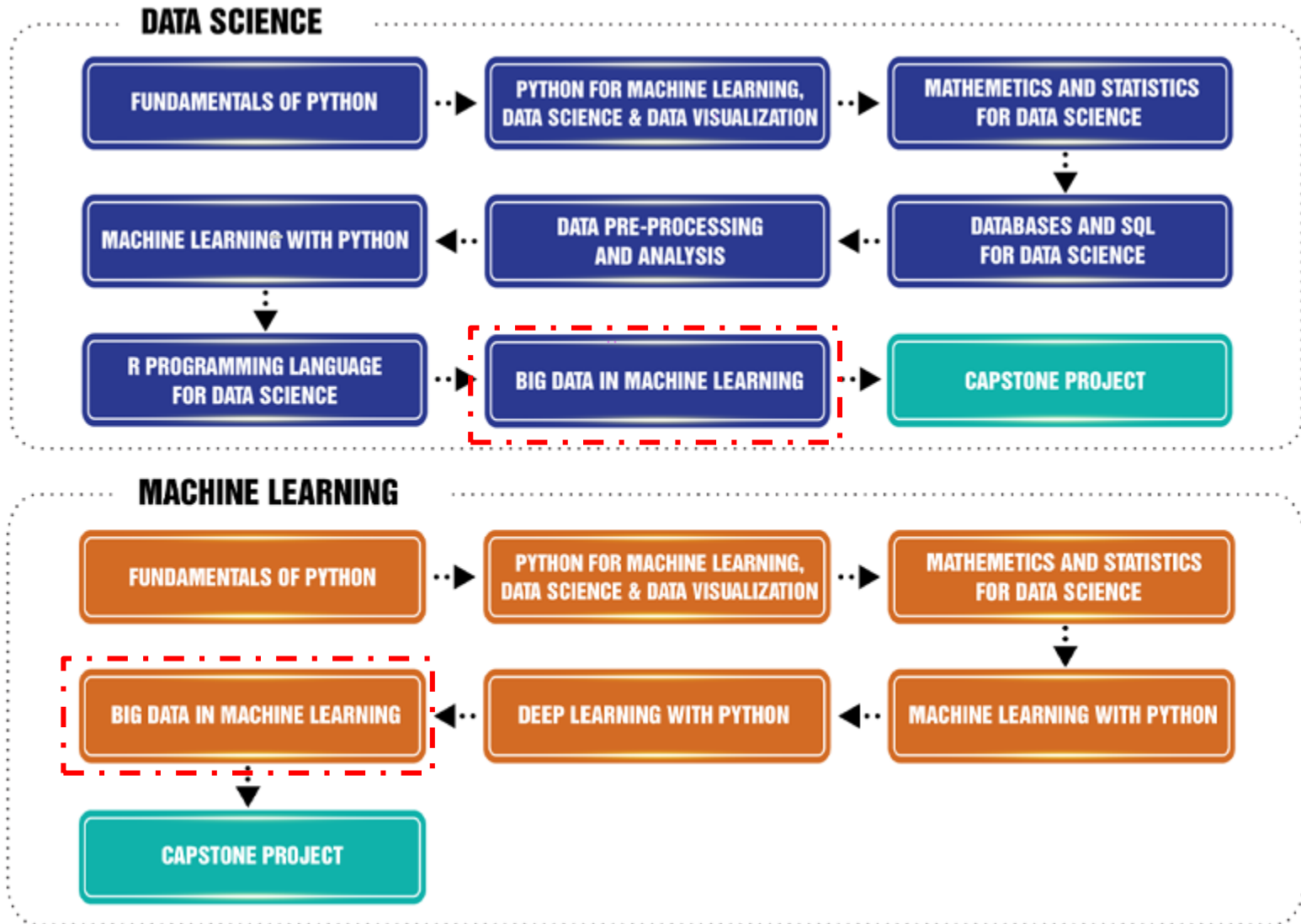
KẾ HOẠCH HỌC VÀ THI LỚP LDS9 - k260

	Tuần 1		Tuần 2		Tuần 3		Tuần 4				Tuần 5
	Buổi 1	Buổi 2	Buổi 3	Buổi 4	Buổi 5	Buổi 6	Buổi 7(học bù buổi tối thứ 3 hoặc tối thứ 4) - HV chọn 1 trong 2 buổi		Buổi 8	Buổi 9	Buổi 10
Lịch học	3/10	4/10	10/10	11/10	17/10	18/10	20/10	21/10	24/10	25/10	Hoàn chỉnh và nộp bài thi
Hạn nộp bài	8/10/2020		15/10/2020		22/10/2020		29/10/2020				31/10
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





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



❑ 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



❑ 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



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

