

# Python Programming for Science and Machine Learning

서울대학교 자연과학대학 물리천문학부

2018 가을학기

전산물리 (884.310) & 응용전산물리 (3342.618)

# 1 주차 - Introduction to PPSML

- 강좌 소개
- 실습반 지정
- 평가 및 과제 제출 방법
- 실습 및 정규 과제 소개
- 전산물리소개
- 리눅스에서의 프로그래밍 환경 갖추기

# 강좌 개요

- 담당교수 : 조원상 (wscho@snu.ac.kr)
- 소립자 물리학 전공 (입자가속기현상론)
- 담당조교:(연습반,메일):

곽영진(월, jinjin299@gmail.com)

고상현 (화, koman1111@snu.ac.kr)





이용선 (수, leeynsn@gmail.com)

박찬주 (목, chanju6537@hanmail.net)





# 강좌 개요

### 연습반 신청

- 9/9(일)까지 이메일로 신청 후 2 주차 (9/10) 부터 연습반 출석
- 시간은 연습반 조교와 월 목 (저녁 5-9 시) 사이에 탄력적으로 결정 후 공지

### 홈페이지

- ETL [전산물리 / 응용전산물리]
- GitHub
- Zoom Meeting Room (<a href="https://zoom.us/j/326917405">https://zoom.us/j/326917405</a>)

### 평가

- 출석 (10, 임의지정좌석, 매달초 새로운 좌석배치 ETL 공고)
- 과제:
  - 실습과제 (10): (+a 독창성)
  - 정규과제 (20): (+a 독창성)
- 중간 (30)
- 기말(30)

# 강좌 개요

### 과제 제출 방법

- 수강생별 구글 드라이브 과제 제출함을 이용
- 구글 이메일 계정 생성 후, 2 주차 (9.10 14) 까지 연습반 조교와 제출함 생성 및 공유 완료

### 실습 / 과제파일 제출 형식

- 실습, 정규과제 파일 제출시 제출파일이름의 형식을 꼭 (!) 지킬 것 (if not: 누락 가능성 += 1.0)
- 형식: {'과제종류': ['exercise', 'homework']}\_{' 회차': ['01', '02',...]}\_{' 과목': {'대학원': 'c1', '학부': 'c2'}}\_{(\*student\_id'}\_{(\*name')}.{'file extension'}

예)

실습과제 1 차: "exercise\_n01\_c1\_2017-1234\_ 황이조 .ipynb / .py"

정규과제 11 차: "homework\_n11\_c2\_2018-1234\_ 조현우 .ipynb / .py"

# 1 주차 과제

### 개별과제 (1)

- 연습반 신청 마무리
- 연습반 조교와 개인과제 제출함 연결 완료
- 개인과제 제출함 들어있는 개인별 소개 / 설문 페이지 완성

### ● 개별과제 (2)

- 연습반조교와 Python & Jupyter-notebook/lab 환경준비완료 (개별 랩탑, ETL 공지 참조)
- Linux 명령어 & vi 편집기 익혀보기

### ● 연습반 실습과제

- 주피터 노트북을 활용한 간단한 파이썬 프로그램 연습
- (9/8 일 ETL 공지)
- 정규과제 (2 주차 (9/14) 부터 제시)

# 인공지능과 과학 (전산물리 ?) Augmentation/Evolution of Science with AI

- (인공) 지능: 어떤 상황 (x) 에 대처할 수 있는 능력 (f(x)), 그리고 새로운 경험 (x') 으로부터 학습을 거쳐 새로운 능력 (f->f') 을 습득할 수 있는 능력
  - 1 ex) Human Intelligence, Machine Intelligence (AI)
- 과학: 관찰데이터 x 대한 이론과 모형 (f(x), ~ 인공지능의 능력)을 체계적으로 가설을 세워, 반복 검증을 통해 확립해나가는 활동
- 과학과 인공지능의 공통적인 목표
  - 1 데이터를 설명하거나 예측할 수 있는 모형의 건설.
  - 2 핵심 키워드: '데이터 (data)' & '모형 (model)'

### 과학 모형과 물리 모형의 건설:

1. 과학모형의 건설:

경험 [실험데이터] → 경험법칙 → 실험데이터 검증 반복 (ex: 케플러 법칙)

2. 물리모형의 건설:

과학모형 → 보다 근본적인 이론에 기반을 둔 보다 일반적이고 강력한 예측력을 갖는, 수학적 언어로 쓰여진 물리 모형으로 발전

(ex: 뉴튼의 중력이론)

3. 물리모형의 확장:

기존의 물리모형

- → 연역과 유추 / 수학적 , 물리적 대칭성 고려 / 실험적 이상현상
- → 물리모형의 확장
- → 새로운 이론 / 모형으로 새로운 현상 예측 / 이상현상설명
- → 검증 데이터 수집 및 새로운 이론의 검증

(ex: 아인슈타인의 중력이론, 소립자 물리학의 표준모형 & 힉스, 초대칭 모형, 중력파,...)

### 과학발전의 역사:

- 1. 자연으로부터의 최대한의 경험적 지식에 대한 수학적 연역체계의 완성과정
- 2. 기본적인 경험법칙 [데이터] 로부터, 고차원적인 추론이 가능한 물리모형으로 진화해오고 있으며, (실험적 이상 현상없이도) 연역과 유추에 기반을 둔 새로운 물리현상의 탐색도 활발히 이루어지고 있다.
  - ex) 대통일장 이론 (4 가지 힘의 통일), 초끈이론,…

### 머신러닝을 통한 인공지능 모형의 건설

- 1950 년대 AI(Artificial Intelligence) 라는 용어의 탄생 (J. McCarthy, ...) 과 더불 어 인공지능의 건설을 위한 다양한 방법론이 등장하기 시작하였다.
- 2000 년대 초기까지 주류에 속한 방법론은, 인간의 언어 개념과 논리에 기반한 프로그래밍 기법이라 할 수 있다. ex) Expert System (전문가 시스템)
- 가용 데이터량과 전산 성능의 폭발적인 향상에 힘입어, 현재는 데이터를 기반으로 스스로 모형을 학습하는 기계학습 (Machine Learing) 법이 인공지능의 발전을 이끌고 있음.

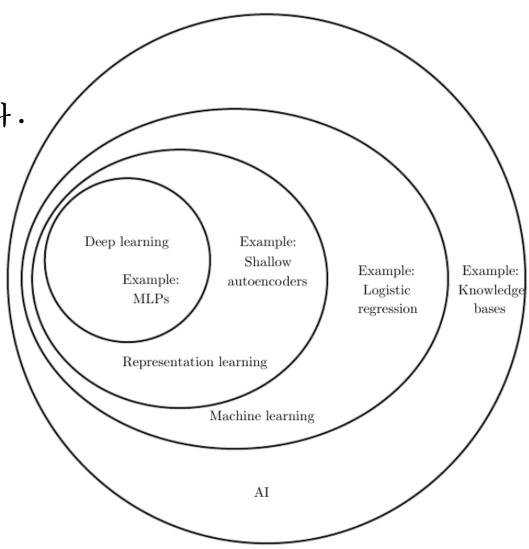
### Machine Learning (기계학습):

인공적인 기계 시스템이, 인간이 프로그램하지 않고도, 데이터를 통해 스스로 주요한 패턴과 법칙을 습득하는 것 / 습득할 수 있도록 만드는 것

[Giving computer systems the ability to learn the important patterns and rules based on data inputs, without being explicitly programmed]

- ".. we can know more than we can tell."
  - M. Polanyi, 'The Tacit Dimension'

머신러닝 / 딥러닝은 데이터를 기반으로 한 인공지능구현의 한 방법이다.



• [Ian Goodfellow, Yoshua Bengio and Aaron Courville] 'Deep Learning'

# 머신러닝의 예) 데이터 분류 모형의 건설

### CLASSIFICATION

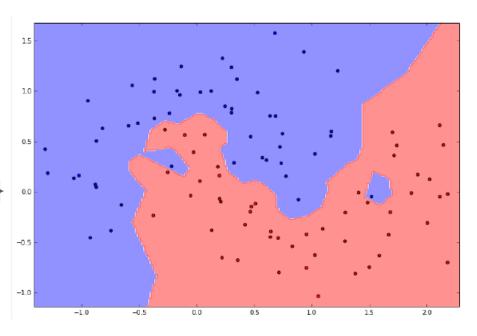
\* Mapping of an event (i) with features  $(x_i)$  to a  $(z_i)$  of a finite label set :

Vector of features, available for event (i)  $i:x_i o z_i$ 

Example of a binary classification using Machine Learning algorithm (kNN)

$$x_i \in \mathcal{R}^2$$
  
 $z_i \in \{0, 1\} \text{ for } \{\text{blue, red}\}$ 

=> {signal, background} discrimination



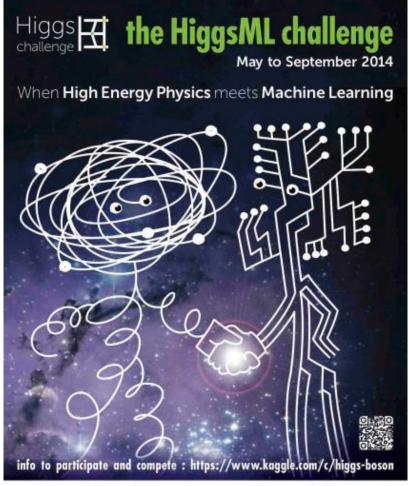
# Fun with Classification

• playground.tensorflow.org: 간단한 2Dim 이진 분류 문제에 대한 인공신경망모델이 구축되는 과정을 웹에서 직접 살펴볼 수 있다

- <a href="http://cs.stanford.edu/people/karpathy/convnetjs/">http://cs.stanford.edu/people/karpathy/convnetjs/</a>
  [ConvNetJS]: 간단한 2Dim 이진분류부터, CNN을 사용한 MNIST set (0,…,9 손글씨) 분류문제, CIFAR-10 Image 분류, 등등 여러가지 신경망들을 사용한 기계학습을 웹브라우저를 통하여 체험해 볼 수 있다.
- Vision recognition: <a href="http://cs.stanford.edu/people/karpathy/convnetjs/demo/cifar10.html">http://cs.stanford.edu/people/karpathy/convnetjs/demo/cifar10.html</a>

# Machine Learning in Science

- Higgs searches (Kaggle competition)
- --> Classification using ML!
- ...many (will be introduced)

















### 정리 및 전망

- 자연관찰 데이터를 학습해낸 머신러닝 모형은 과학모형이다.
- 머신러닝 모형 건설을 통해, 강력한 성능의 과학모형의 학습을 이룰 수 있을까?
  - → (현재 매우 활발히 이루어지고 있음) → (머신러닝@전산물리)
- 더 나아가, 머신러닝 모형 건설을 통해 연역과 유추가 가능한 고차원의 물리모형을 탐색하고 건설할 수 있을까?
  - → ( 여러분들의 몫!)
- 이를 위해서는 머신과의 원활한 의사소통능력이 필수 (외국어 습득하듯이)
  - →(프로그래밍 언어습득의 필요성)→(파이썬@ 리눅스 for 전산물리
  - = 권장언어@환경)

- 인간 vs (?) 기계 → 공포
- 우리의 상상과 실행에 따라서 ...
  - → 인간 과 기계
  - → ' 증강'된 인류 (Augmented Human)
  - → 달리는 호랑이의 등에 올라타있다.

# 인공지능과 과학

• ML to be continued ( 중간고사 이후 ) ···

- ML 관련 학기내 참고도서 (중간고사 이후)
  - 1 물리학과 첨단기술 '머신러닝' 특집
  - 2 Pattern Recognition and Machine Learning [C. Bishop]
  - 3 Hands on Machine Learning with Scikit-Learn and Tensorflow [Aurélien Géron, 박해선 역]
  - 4 Deep Learning [I. Goodfellow, et al]

# LINUX 에서의 프로그래밍 환경

### 하드웨어 아키텍쳐 Hardward architecture:

- CPU(x86\_64(intel,amd), ARM, PPC(ibm), ···)
- GPU(nvidia, AMD, ···)

### 운영체제 (Operating System):

- Unix 계열 (Linux, MacOSX, AIX, Solaris, FreeBSD, …)
- Windows ...

### 프로그래밍 언어 (Languages):

• Fortran77/90/95, C/C++, **Python**, R, Java, Javascript, Julia, Haskell, Mathematica, Matlab ...

프로그램 컴파일 & 빌드 도구: Make, Autotools, CMake, GDB

### 라이브러리의 사용:

컴퓨터 이용의 효율성을 높이기 위하여, 유용한 하부 프로그램 및 루틴들을 모듈화하여 모아놓은 도서관으로서, 보통 컴파일된 목적파일 (object file) 형태로 존재 (소스>목적>실행)

### 모듈:

어떤 기능에 대한 알고리즘이 있을 때, 프로그래밍 언어로써 그 기능이 구현되고 구조화된 단위 ~ 함수. 범용 프로그램에 링크하여 사용. 모듈의 집합 = 라이브러리

### 각각의 프로그래밍 언어에 따라 방대한 라이브러리가 존재:

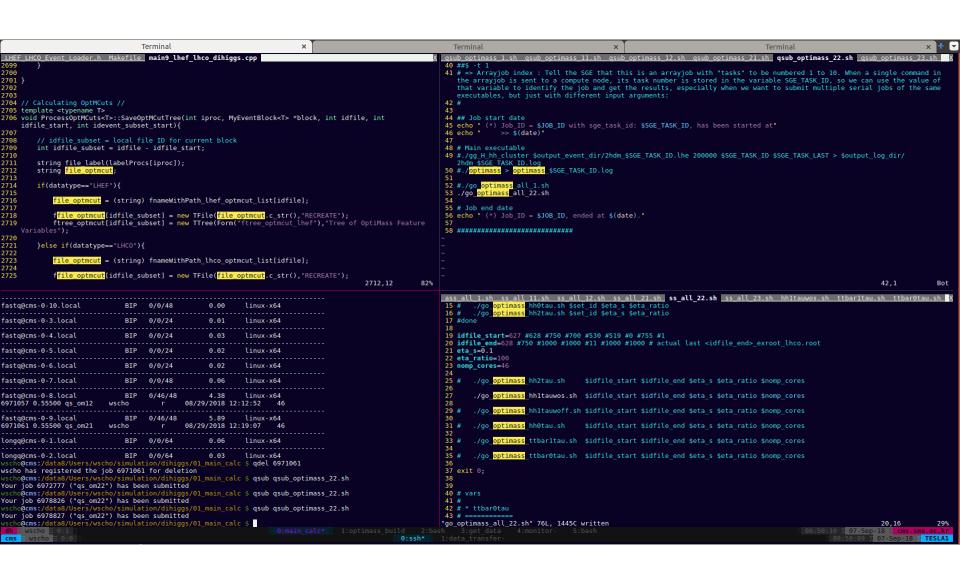
GSL(C/C++), STDLIB, Boost(C++), LaPack & BLAS, Intel MKL, (SciPy, NumPy, SymPy, MatPlotLib, ...) for Python

## 편리한 프로그래밍(코딩) 환경구성:

- 강력한 텍스트 편집기 (vi/vim, emacs / …)
  - <u>vim cheat sheet</u>
  - vim tutorial

- 작업창 / 터미널 관리자 (screen / tmux / …)
  - <u>tmux home</u>
  - <u>tmux tutorial</u>
  - tmux cheat sheet

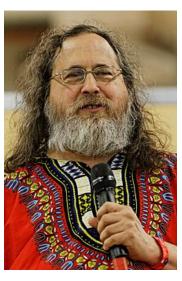
### 예 ) C++ & Shell scripts @ 리눅스 터미널 + vi + tmux



# 리눅스

### 리눅스:

- 'Let's make a free and open-source Unix-like OS!'
  - 1 <u>Unix OS</u> ( 멀티태스킹과 멀티유저가 가능한 OS, 상용, 1970~)
- = GNU project (1983~ , by Richard Stallman) 의 모토
  - 1 <u>GNU Operating System</u>
- Linus Torvalds 가 1991 년 GNU project 의 철학에 걸맞는 OS 의 커널 (Linux kernel) 을 독립적으로 제작, Linux OS 로 발전



Richard Stallman



Linus Torvalds

# 리눅스

### 여러가지 리눅스 배포판 (Linux distributions):

리눅스 커널을 바탕으로 각기 다양한 특성을 발전시킨 여러 배포판 (distribution) 존재 (상용 / 비상용). 약 300 여가지.

Red-Hat Enterprise Linux (RHEL, 상용)

- CentOS: stable version of RHEL
- Fedora: progressive version of RHEL

### Debian

• **Ubuntu**: Debian 으로부터 파생. 현재 많이 쓰이는 배포판들중 하나.

Mint

Arch

Gentoo

...

여러가지 리눅스 배포판 (Linux distributions) 랭킹 및 최신정보 확인

https://distrowatch.com/

# 리눅스

- 리눅스 설치 및 환경설정:
  - 전산물리반 [<u>링크</u>]
  - 응용전산물리반 [링크]
- 리눅스 기본명령어:
  - 전산물리반 [<u>링크</u>]
  - 응용전산물리반 [링크]