Introduction to Machine Learning

Assoc. Prof. Karl Ezra Pilario, Ph.D.

Process Systems Engineering Laboratory

Department of Chemical Engineering

University of the Philippines Diliman

Outline

- What is Machine Learning?
  - Why only now?
  - Types of Learning Problems

- Intro to the Course (AI 221)
  - Course Delivery
  - Course Content
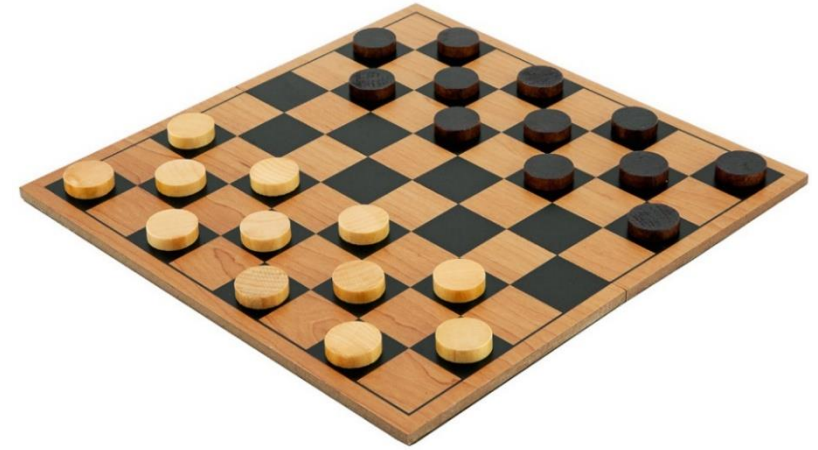 - Course Requirements
  - Software

What is Machine Learning?

A field of study concerned with giving computers the *ability to learn* without being explicitly programmed.
(Arthur Samuel, 1959)
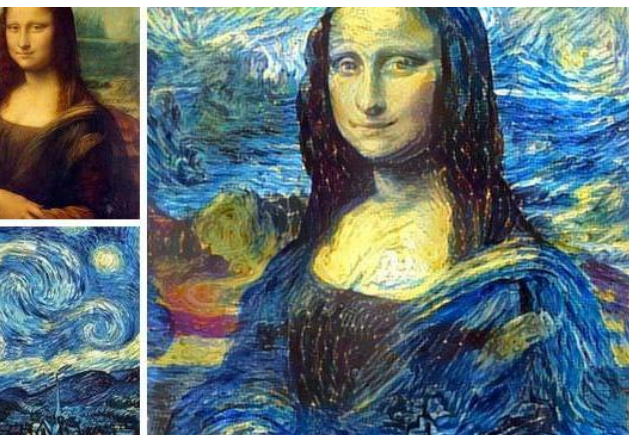
Arthur Samuel and the IBM 701 Computer

- Arthur Samuel was not a very good checkers player.

- But he was able to program a checkers bot that plays *better than a human*.

- The bot learned by playing millions of times against itself.

- In the process, it learns which moves lead to wins and which moves lead to losses.
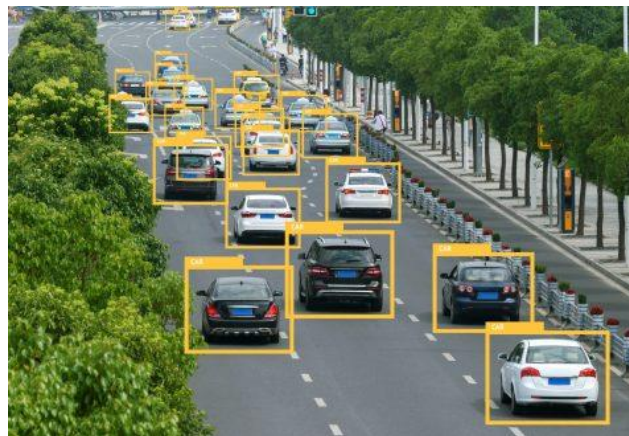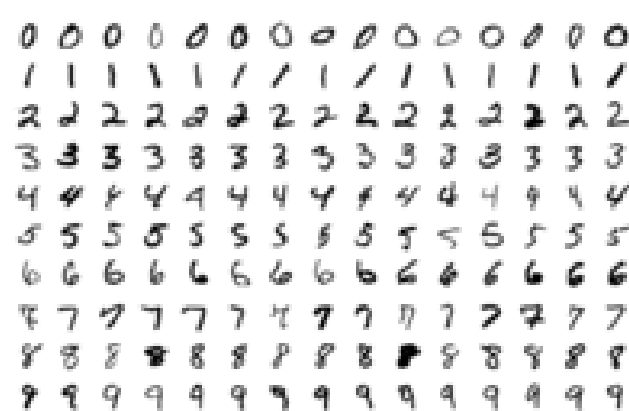
Typical ML Applications

- Filtering emails as spam / not spam

- Handwritten digits recognition

- Speech recognition, Natural Language Processing (NLP)

- Social media (Face recognition, News Feed Ranking, etc.)

- Image / Object Recognition, Image Segmentation

- Recommender systems (movies, products, videos, webpages, bookings)

- Finance (Stock market prediction, customer behaviour, etc.)

- Transportation (Self-driving cars, travel demand modelling)

- Healthcare (Early diagnostics, hospital demand forecasting)

- Bioinformatics (Protein folding and structure prediction, Gene function prediction, Biomedical image analysis)

- Chemometrics (GC-MS data analysis, drug discovery, compound classification, chemical property prediction)

IBM Watson
Jeopardy, **2011**
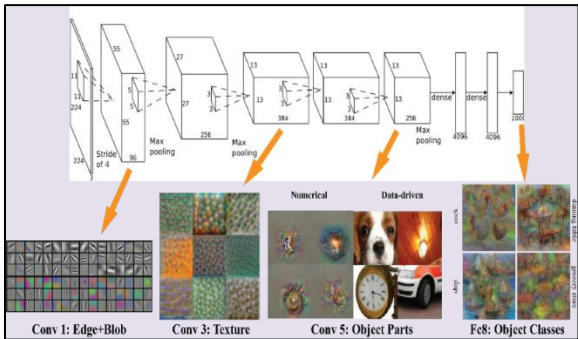
AlexNet
ImageNet Visual Recognition
Challenge, **2012**

AlphaGo
Game of Go, **2016**
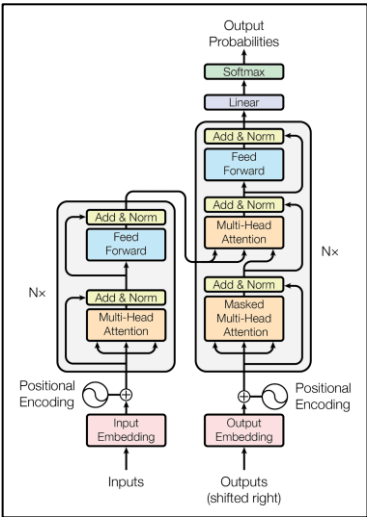
Transformers
2017

DALL-E
2021, 2022

IBM Deep Blue
Chess, **1997**

AlphaStar
StarCraft II, **2019**

T1O37 / 6vr4
9O.7 GDT
(RNA polymerase domain)
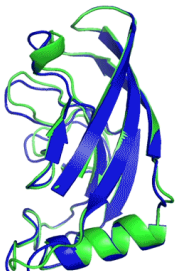
T1O49 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction

AlphaFold
Protein Structure Prediction,
2016, **2018**

ChatGPT
2022

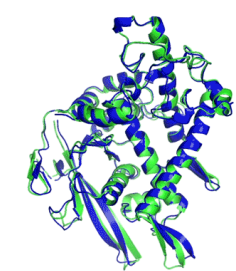**Machine Learning,
Data Science,
Data Analytics,**

…why only now?

Big Data

Computing Power

Data Storage

Machine Learning +
Practical Applications

Growth in **Data Volume** Generated worldwide *(in zetabytes)*

97

2022

Growth in **Data Storage** worldwide
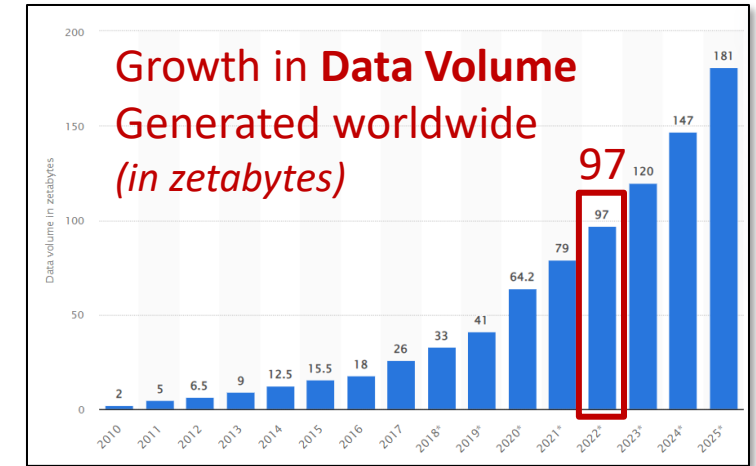
Growth in **Computing Power** worldwide
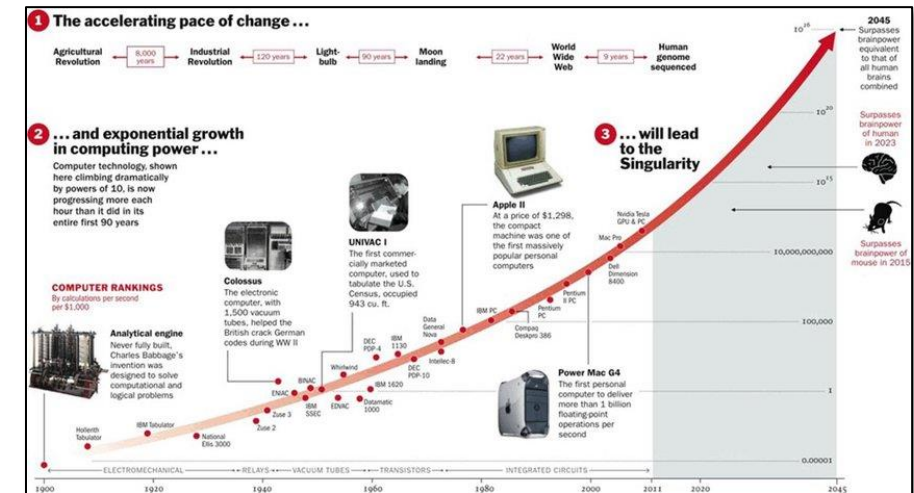
Machine Learning,

Data Science,

Data Analytics,

…why only now?

We are currently DROWNING[1] in data!

- There are about 1 trillion web pages.

- 1 hr of video is uploaded to Youtube every second.

- Human genomes have a length of 3.8×10^9 base pairs.

- Walmart handles more than 1 million transactions per hour.

- Etc…

Popular websites where we can get publicly available data:
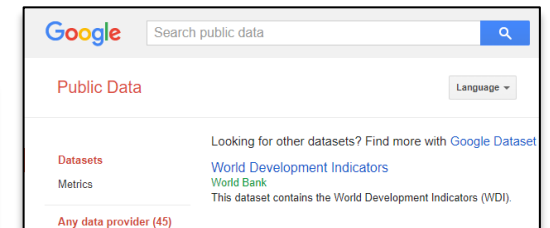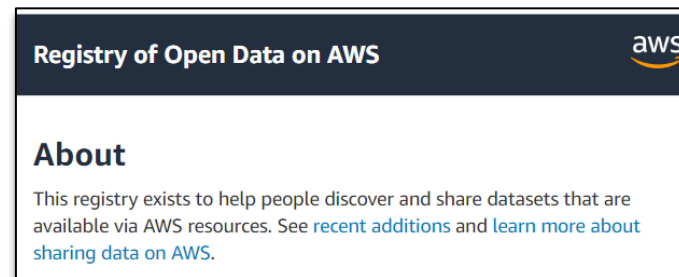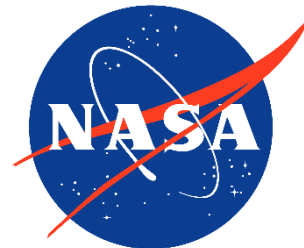
[1] Venkatasubramanian (2009). DROWNING IN DATA: Informatics and Modeling Challenges in a Data-Rich Networked World. *AIChE Journal.*
[2] Murphy (2012). Machine Learning: A Probabilistic Perspective. *MIT Press.*

Why use Machine Learning in your Industry?

Three approaches to engineering problems:

1. Physics-driven Methods
2. Knowledge-driven Methods
3. Data-driven Methods

Machine learning is a **data-driven approach**.

First-principles Modelling

System of Differential Equations

Physics-driven

Physical Laws

Digital Twins

Simulation Software

Knowledge-driven

Human Experience

Heuristics

Rules-of-Thumb

Expert Knowledge Base

Physics-informed A.I.

Grey-box models

Expert Systems, A.I., Data Mining

Black-box models

Machine Learning

Data-driven

Multivariate Statistics

Online Data Streams

How to turn data into decisions?

- Applying machine learning to your data is not enough.

- Ask yourself the purpose for using ML and AI.

- Don't just let your data speak, let it change the way you do things. The goal is prescriptive analytics!

- Getting through each stage of analytics requires more and more effort, but also more returns.

Prescriptive Analytics — Prescribe **what should be done**.

Predictive Analytics — Predict **what will happen** next

Diagnostic Analytics — Find out **why** it happened

Descriptive Analytics — Describe **what** happened

Value (y-axis) / Complexity (x-axis)

Outline

- What is Machine Learning?
    - Why only now?
    - **Types of Learning Problems**

- Intro to the Course (AI 221)
    - Course Delivery
    - Course Content
    - Course Requirements
    - Software

Types of Learning Problems

Supervised Learning

Learn a mapping or a function:

$$y = f(x)$$

from inputs (x) to outputs (y), *given a labelled set of input-output examples* (● or ✖).

Regression
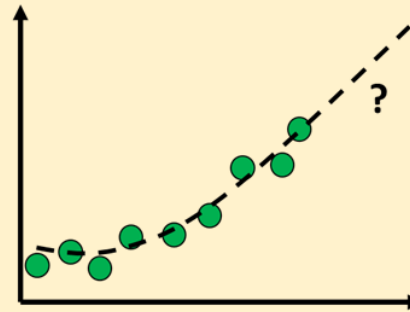
Classification
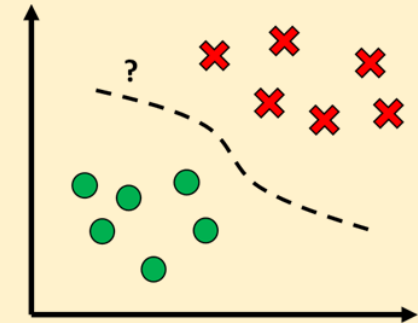
Unsupervised Learning

Discover *patterns or structure* from a data set (●) without any label information.

Dimensionality Reduction
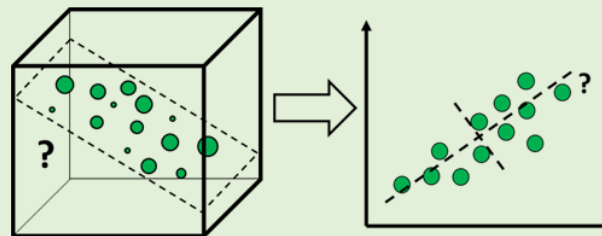
Clustering
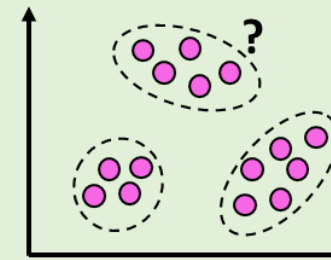
Density Estimation
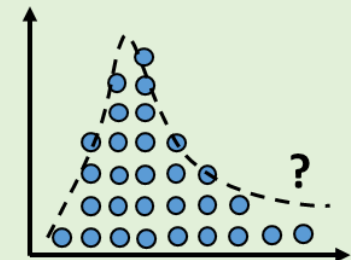
Types of Learning Problems

A simple example…

Types of Learning Problems

Semi-Supervised Learning

Goal: Make a computer learn from both labelled and unlabelled data.

Labelled Data
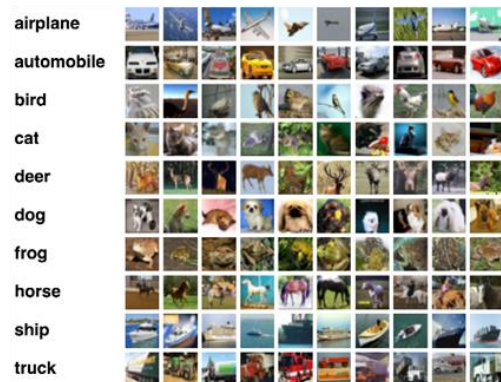
Unlabelled Data

Reinforcement Learning

Goal: Make a computer learn by letting it interact with the environment.

state s_t
reward r_t
action a_t
Agent
r_{t+1}
s_{t+1}
Environment

Supervised Learning

Learn a mapping or a function:

$$y = f(x)$$

from inputs (x) to outputs (y), given a labelled set of input-output examples (● or ✖).

Regression

- **Given:** Training Data $\{x_i, y_i\}_{i=1,2\dots,N}$

- Target y_i is a **continuous** variable.

- Examples:
  - Forecasting future stock price
  - Forecasting energy resources
  - Prediction of key performance indicators
  - Predicting the properties of molecules based on their structure
  - Predicting the environmental impact of pollutants
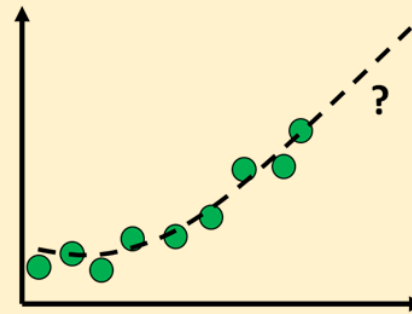
Classification

- **Given:** Training Data $\{x_i, y_i\}_{i=1,2\dots,N}$

- Target y_i is a **categorical** variable.

- Examples:
  - Classifying objects in images
  - Classifying chest X-ray images into COVID positive/negative
  - Handwritten digits recognition
  - Filter e-mails into spam/not spam
  - Classify critical equipment as to healthy or faulty
  - Activity recognition from wearable devices

Dimensionality Reduction

- **Given:** Data $\{x_i\}_{i=1,2...,N}$

- **Reduce features** but retain the most important information from the original data.

- Examples:
  - Feature Engineering
  - Image compression
  - Filtering noise from signals
  - Source separation in audio
  - Data visualization

Clustering

- **Given:** Data $\{x_i\}_{i=1,2...,N}$

- **Group** similar data points together.

- Examples:
  - Customer segmentation
  - Recommendation systems
  - Identifying fake news
  - Clustering documents, tweets, posts

Density Estimation

- **Given:** Data $\{x_i\}_{i=1,2...,N}$

- **Estimate** the distribution of your data.

- Examples:
  - Anomaly Detection
  - Novelty Detection
  - Generative Models
  - Finding distribution modes
  - Spatio-temporal analytics

Machine Learning Methods
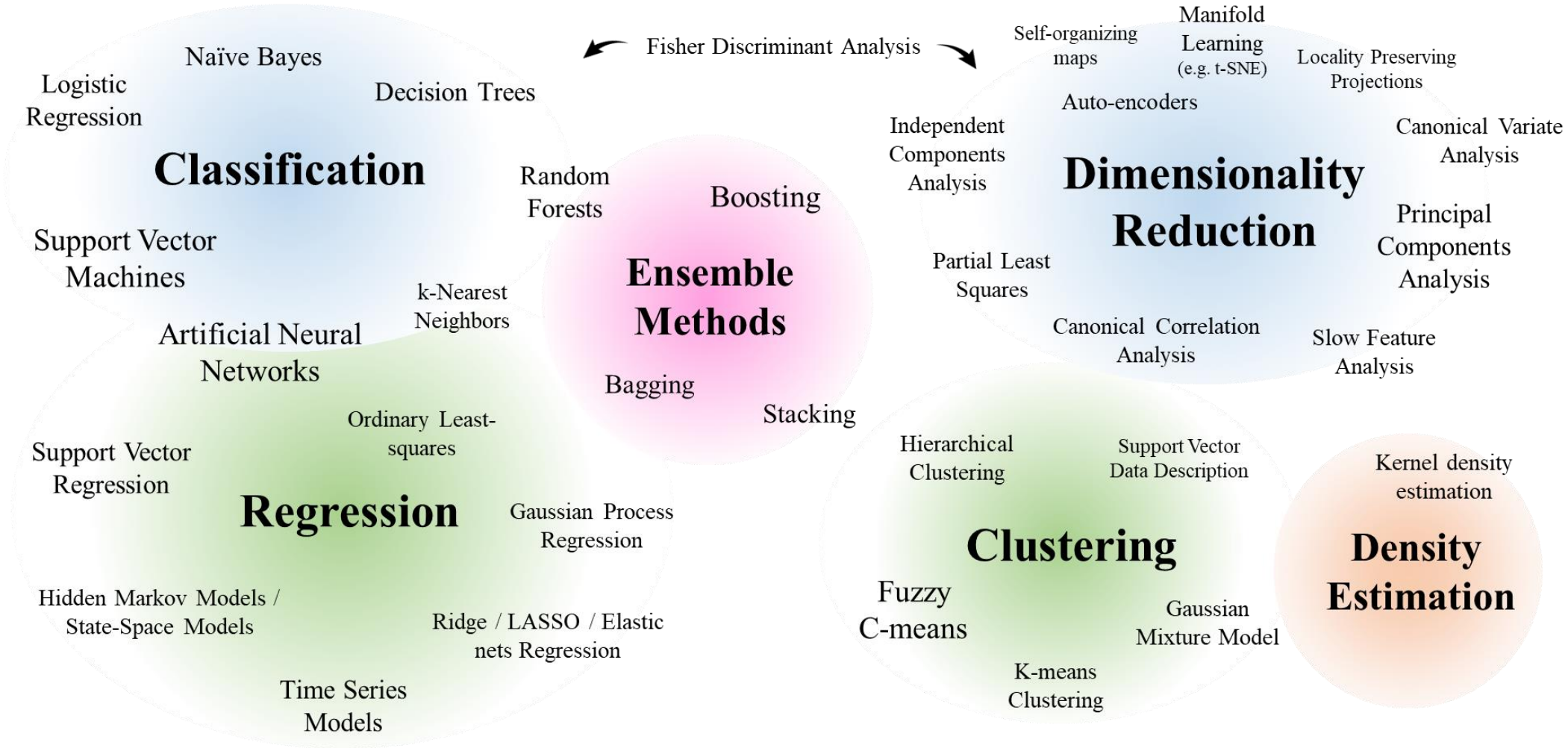
Reference: Pilario et al. (2020), *A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring.* MDPI: Processes, https://doi.org/10.3390/pr8010024

Outline

- What is Machine Learning?
    - Why only now?
    - Types of Learning Problems

- **Intro to the Course (AI 221)**
    - Course Delivery
    - Course Content
    - Course Requirements
    - Software

Introduction to the Course

COURSE NUMBER: AI 221

COURSE TITLE: Classical Machine Learning

COURSE DESCRIPTION: Linear Models. Kernel Methods. Neural Networks. Trees. Clustering. Dimensionality Reduction. Feature Engineering. Density Estimation. Ensemble Learning. Gaussian Processes. Bayesian Methods. Hyperparameter Search. AutoML. Explainability.

COURSE CREDIT: 3 units

3.0 hours/week

COURSE LMS*:   **UVLE Course Page: AI 221 [SQRU]**

Github: https://github.com/kspilario/AI221

*LMS = Learning Management System

Introduction to the Course

MEngg in Artificial Intelligence

Total Units: 31 Units
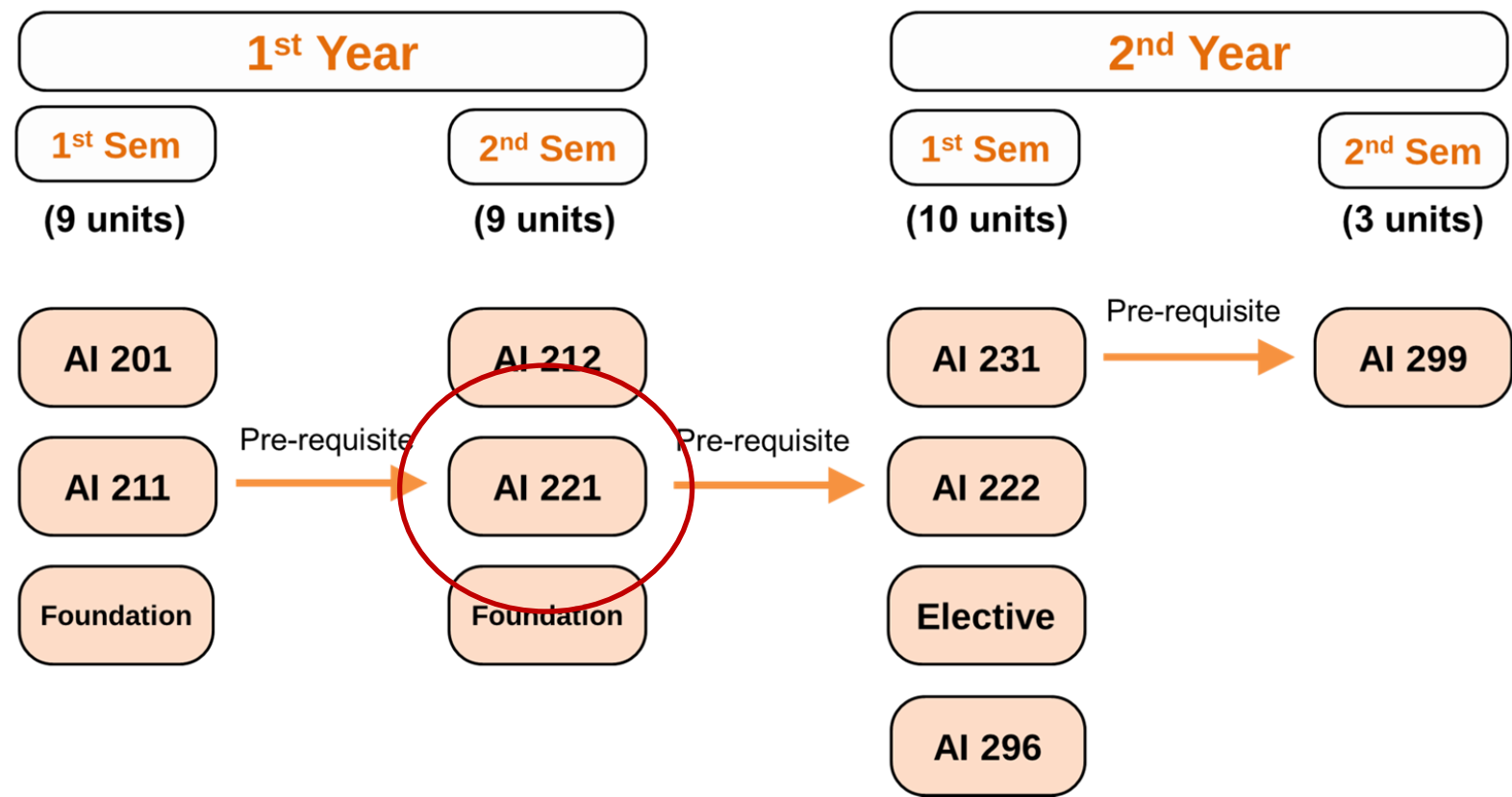
Introduction to the Course

PhD in Artificial Intelligence
Option A

Total Units: 62 Units
Requirement: 2 Publications
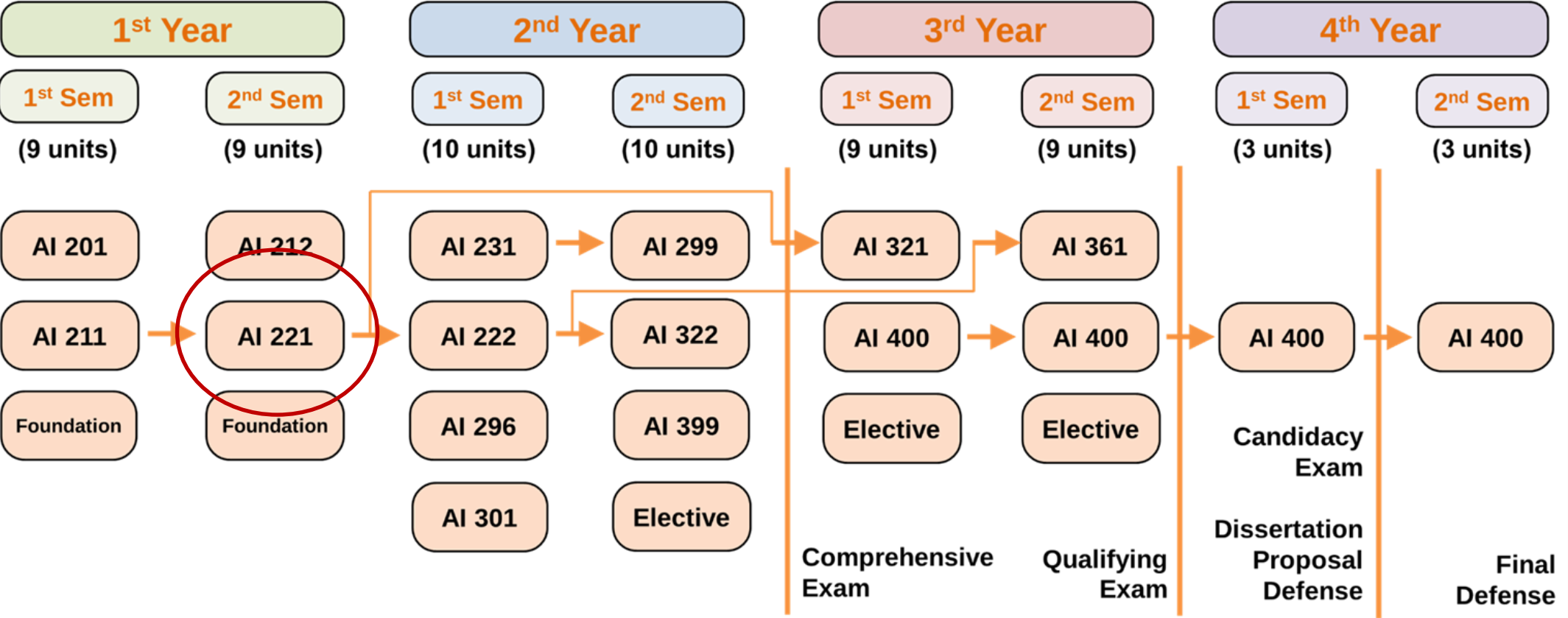
AI 221 Course Delivery

- **Meeting:**   Every Saturday online via Zoom, except for the *Team Project Presentation.*

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	**Saturday**

- **Course Requirements:**

Requirement	% of Final Grade	Mode
• Team Project • Oral Presentation (40%) • Written Report (60%)	**40%**	"Teams" of 1 to 3 members only, Face-to-face
• Machine Exercises	**40%**	Individual, Asynchronous
• Journal Critique	**20%**	Individual, Asynchronous

- **Grading System:**

[92,100]	[88,92)	[84,88)	[80,84)	[76,80)	[72,76)	[68,72)	[64,68)	[60,64)	[0,60)
1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**5.00**

AI 221 Course Content

Feb 18   Week 1.   Introduction to Machine Learning

Mar 4   Week 2.   Linear and Logistic Regression

Mar 11   Week 3.   Support Vector Machines and Kernel Methods

Mar 18   Week 4.   Cross-validation and Hyper-parameter Optimization

Mar 25   Week 5.   Linear Dimensionality Reduction + Discriminant Analysis

Apr 1   Week 6.   Nonlinear Dimensionality Reduction

Apr 22   Week 7.   Clustering, Density Estimation, and Anomaly Detection

Apr 29   Week 8.   Trees, Weak Learners, and Ensemble Learning (Boosting, Bagging, Stacking)

May 6   Week 9.   Neural Networks for Classification, Regression, and Dim. Reduction

May 13   Week 10. Neural Networks for Time Series

May 20   Week 11. Gaussian Processes and Bayesian Optimization

May 27   Week 12. AutoML and ML Explainability

Jun 3   Week 13. *Team Project Presentation*

AI 221 Course Requirements

Team Project (40%)

- A team should have at most 3 members only.

- Aims:
  - Find a problem + data set that requires an ML solution.
  - Solve the problem using the ML methods discussed in class.
  - Present your results face-to-face.

- NO two teams should have the same problem.

- Grading:
  - Oral Presentation (40%)
  - Written Report (60%)

Machine Exercises (40%)

- Mode: Individual, asynchronous
- To be given every lecture week (Weeks 2-12).
- Submission deadline for all MEX is at the end of classes, June 15, 2023.

Journal Critique (20%)

- Mode: Individual, asynchronous
- Find a paper from a reputable journal / conference proceedings related to your field.
  - Should have an impact factor of 2 or higher.
  - Should be published in the last 3 years.
  - Send me the paper for approval first.

- Critique the paper. More guidelines to follow.

AI 221 Required Software

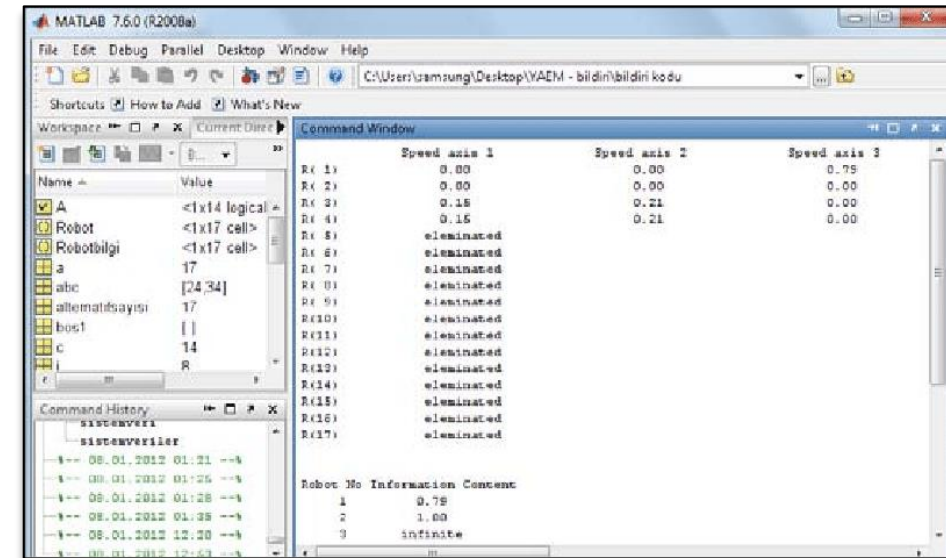- Anaconda >> Spyder
- Google Colab
- Jupyter Notebook
- PyCharm

or
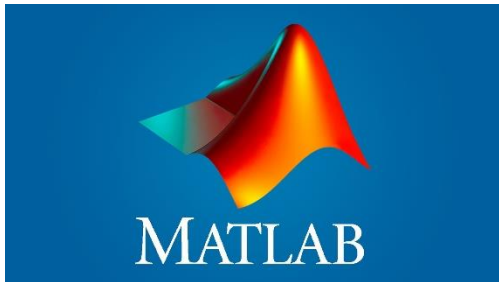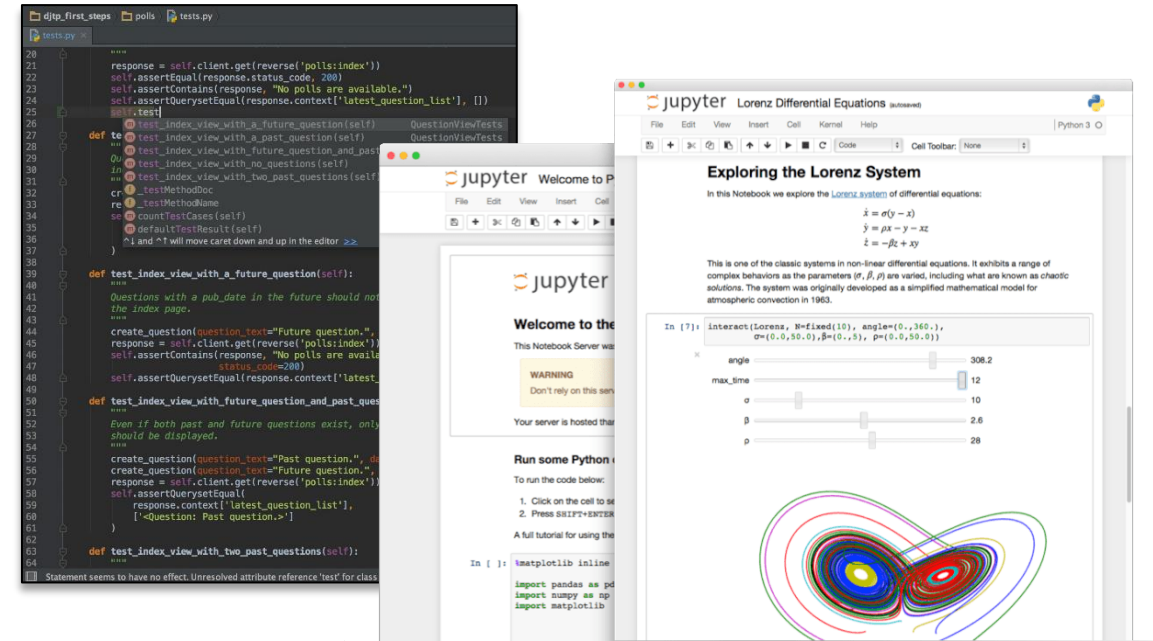
- You can download MATLAB by logging in to www.mathworks.com
  - Use your UP credentials!

- You can also use MATLAB online.

AI 221 Required Software
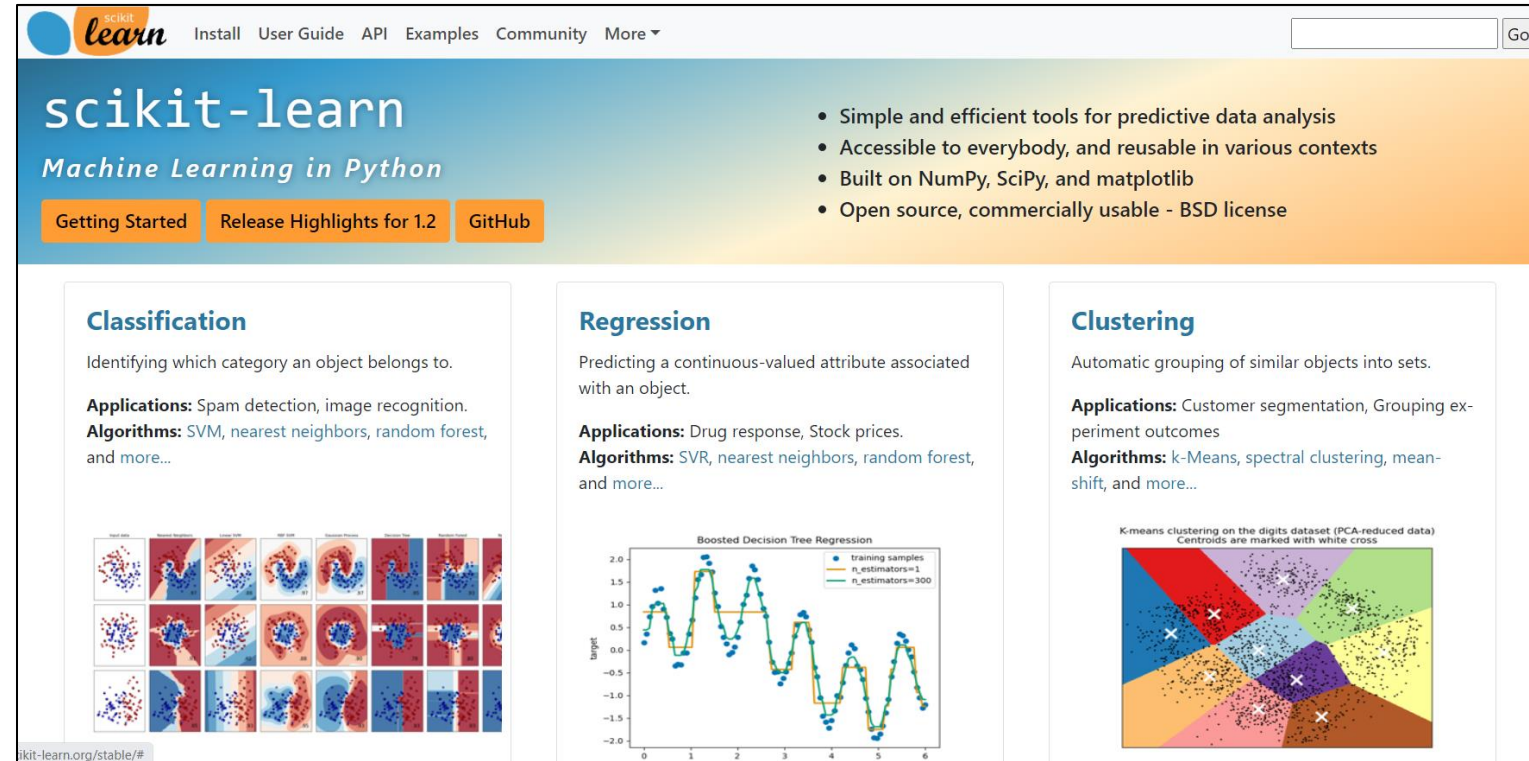

- Python 3
  - https://www.python.org
- Numpy
  - http://www.numpy.org/
- Scikit-Learn
  - https://scikit-learn.org/
- Juptyer Lab
  - https://jupyter.org/try-jupyter/lab/
  - https://nbviewer.org/

AI 221 Course Instructor

University of the Philippines, Diliman

Cranfield University, U.K.

Current Position

Karl Ezra S. Pilario

Associate Professor
Department of Chemical Engineering
University of the Philippines, Diliman

- Process Dynamics & Control
- Programming in MATLAB, Python, Aspen HYSYS
- Numerical Methods in Engineering
- Plant Design and Research
- Machine Learning and Artificial Intelligence

Research Lab

Head, Process Systems Engineering Laboratory (PSEL)
Department of Chemical Engineering
University of the Philippines - Diliman

Education

- **Bachelor's Degree:**
  Chemical Engineering, *SCL* (**2012**)
  University of the Philippines Diliman
- **Master's Degree:**
  Chemical Engineering (**2015**)
  University of the Philippines Diliman
- **PhD Degree:**
  PhD Energy and Power (**2020**)
  Cranfield University, United Kingdom

Research Interests

- Process Data Analytics
- Process Systems Engineering
- Industrial Process Monitoring and Predictive Maintenance
- Machine Learning for Energy, Water, and Environment
- Cheminformatics and Materials Informatics

Outline

- What is Machine Learning?
    - Why only now?
    - Types of Learning Problems

- Intro to the Course (AI 221)
    - Course Delivery
    - Course Content
    - Course Requirements
    - Software