Introduction to Machine Learning

and Introduction to the AI 221 Course

Assoc. Prof. Karl Ezra Pilario, Ph.D.
Process Systems Engineering Laboratory
Department of Chemical Engineering
University of the Philippines Diliman

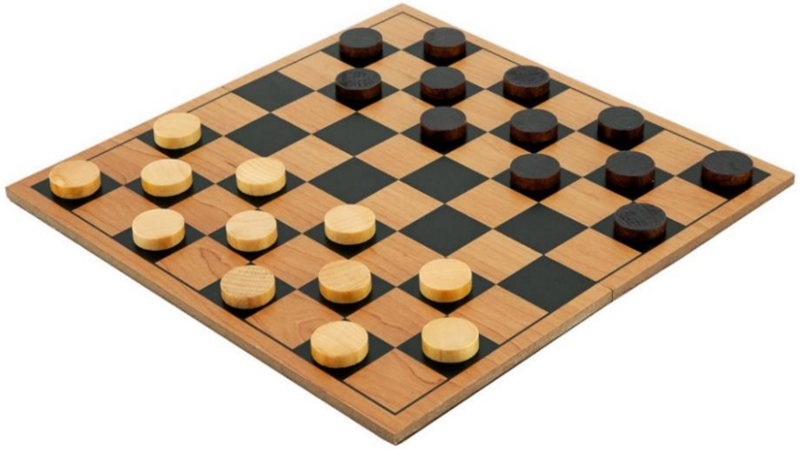
Outline

- What is Machine Learning?
  - Why only now?
  - Types of Learning Problems
- Intro to the Course (AI 221)
  - Course Delivery
  - Course Content
  - Course Requirements
  - Software

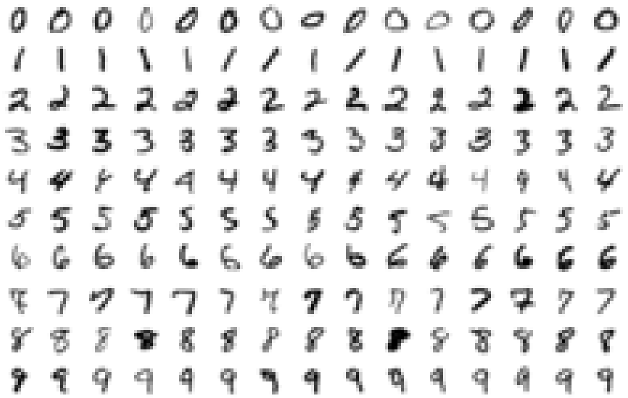
What is Machine Learning?

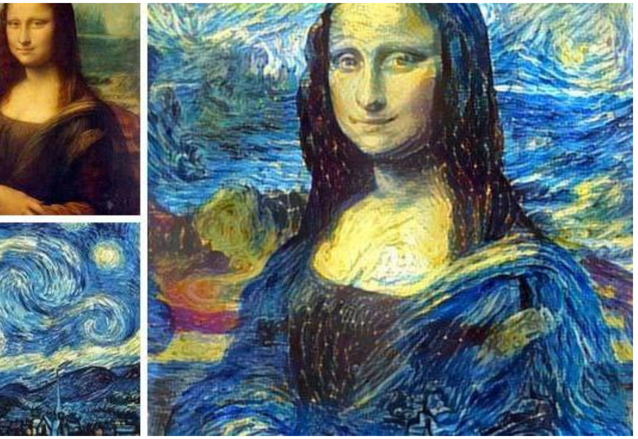
A field of study concerned with giving computers the *ability to learn* without being explicitly programmed.
(Arthur Samuel, 1959)

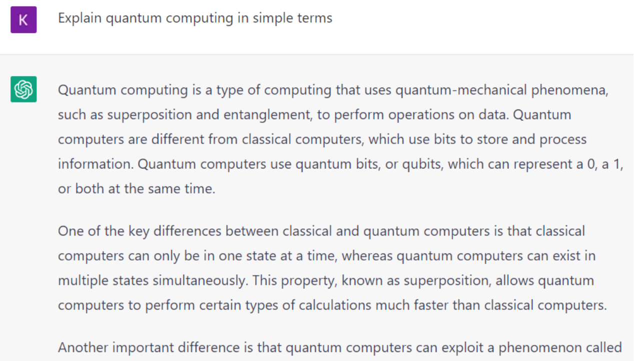
Arthur Samuel and the IBM 701 Computer



- 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.





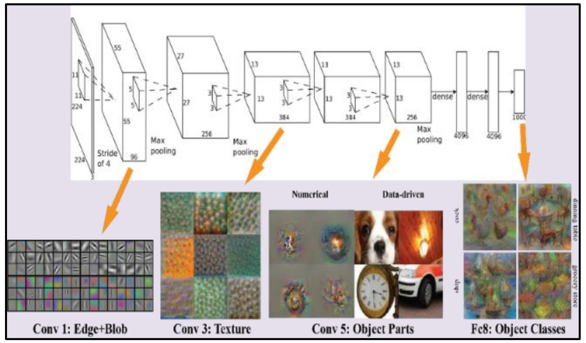


alexa

Typical ML Applications

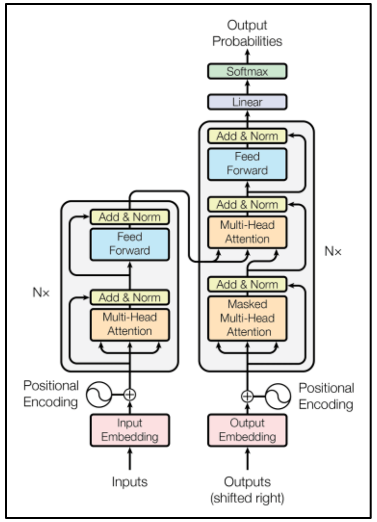
- Filtering emails as spam / not spam
- Handwritten digits recognition
- Speech recognition, Natural Language Processing (NLP)
- Social media (Face recognition, News Feed Ranking, etc.)
- Image / Object Recognition, Image Segmentation
- Recommender systems (movies, products, videos, webpages, bookings)
- 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**

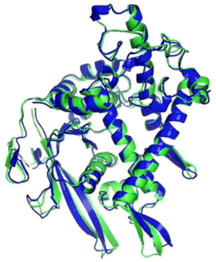


Transformers
2017

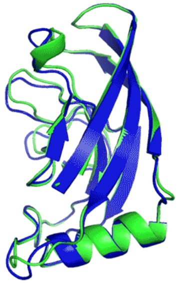
DALL-E
2021, 2022

IBM Deep Blue
Chess, **1997**

AlphaStar
StarCraft II, **2019**



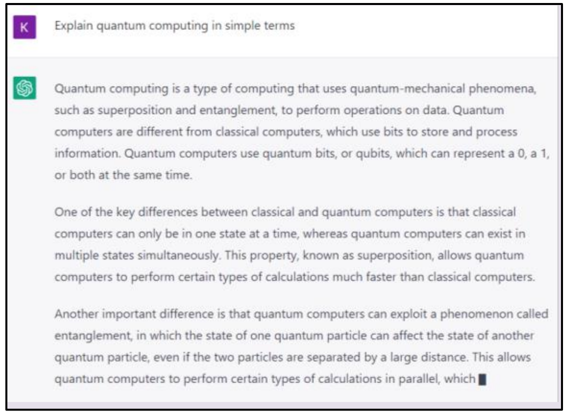
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction

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



K Explain quantum computing in simple terms

Quantum computing is a type of computing that uses quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data. Quantum computers are different from classical computers, which use bits to store and process information. Quantum computers use quantum bits, or qubits, which can represent a 0, a 1, or both at the same time.

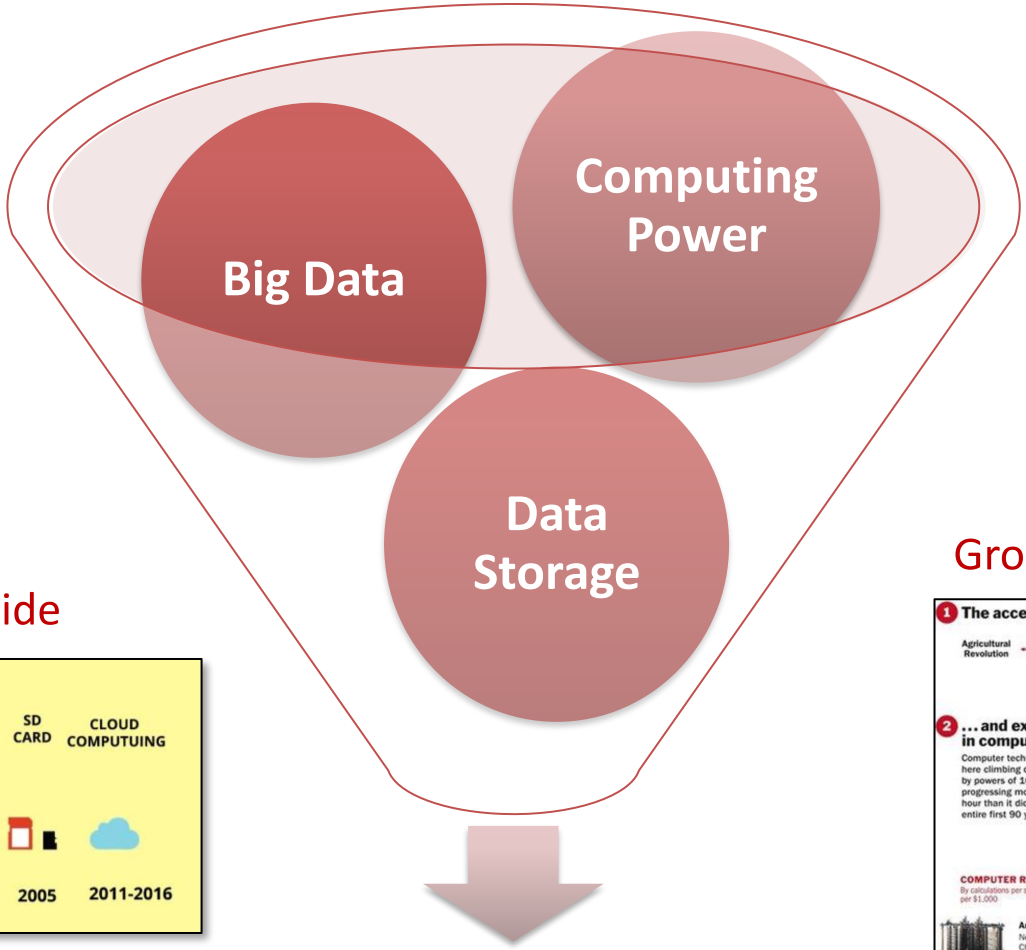
One of the key differences between classical and quantum computers is that classical computers can only be in one state at a time, whereas quantum computers can exist in multiple states simultaneously. This property, known as superposition, allows quantum computers to perform certain types of calculations much faster than classical computers.

Another important difference is that quantum computers can exploit a phenomenon called entanglement, in which the state of one quantum particle can affect the state of another quantum particle, even if the two particles are separated by a large distance. This allows quantum computers to perform certain types of calculations in parallel, which

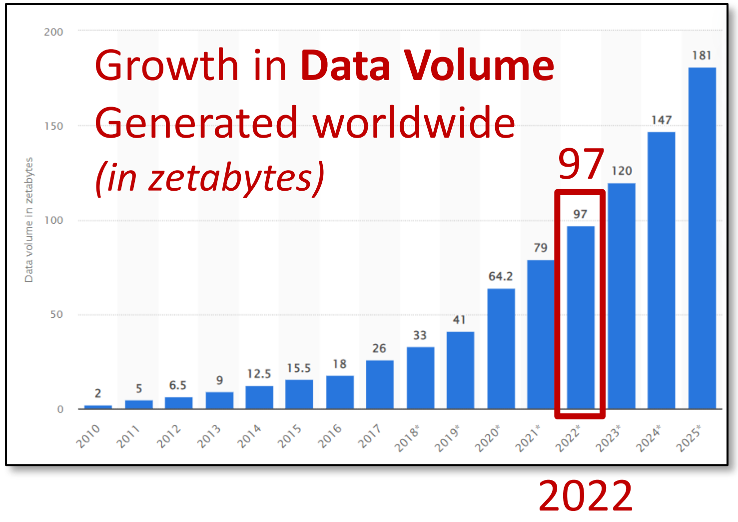
ChatGPT
2022

Machine Learning,
Data Science,
Data Analytics,

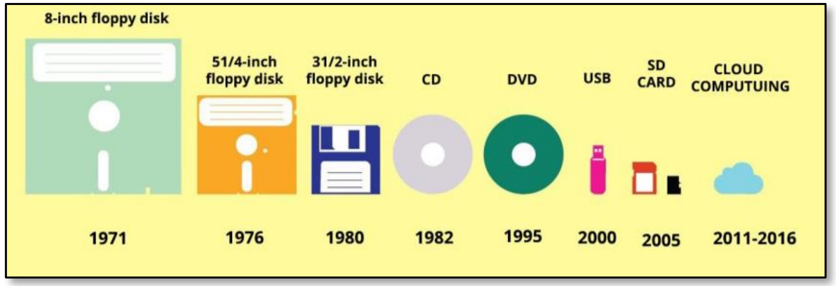
…why only now?



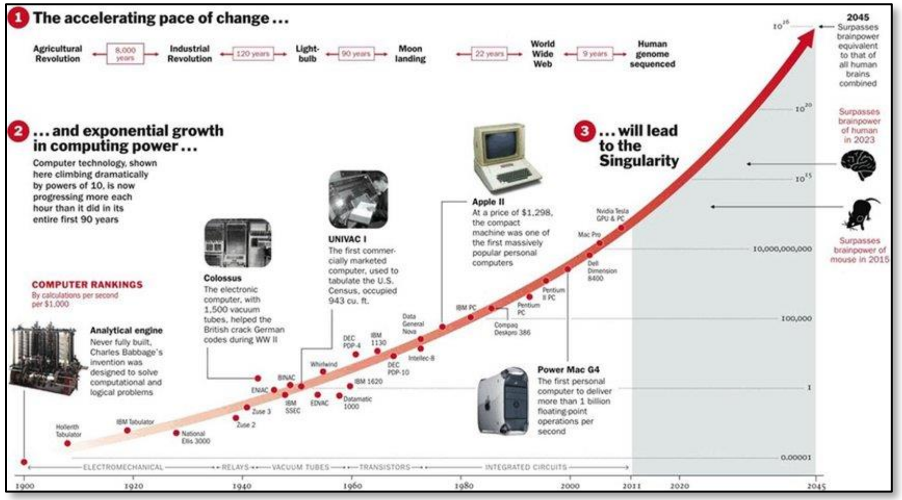
Machine Learning +
Practical Applications



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:

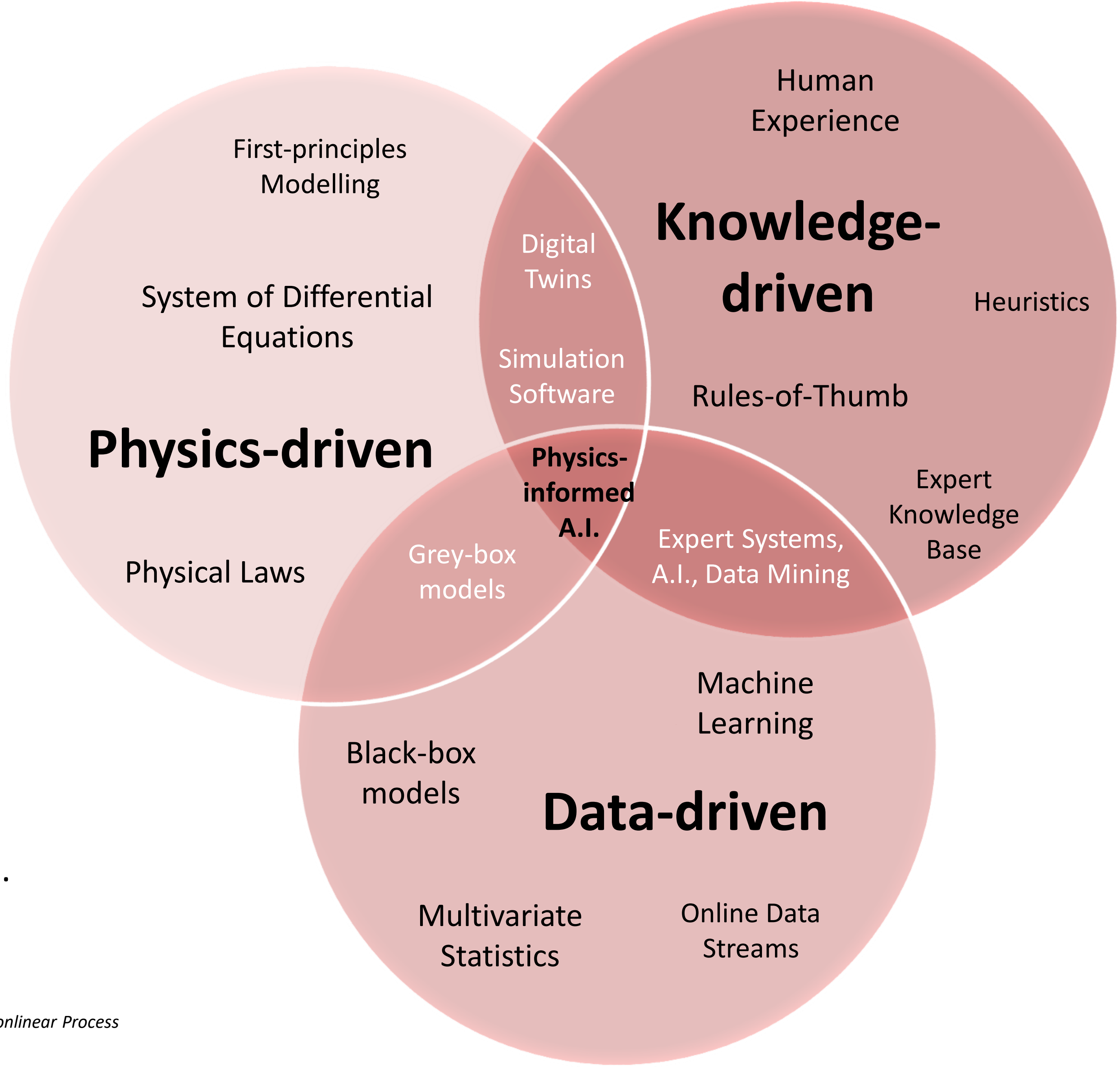
[1] Venkatasubramanian (2009). DROWNING IN DATA: Informatics and Modeling Challenges in a Data-Rich Networked World. *AIChE Journal.*
[2] Murphy (2012). Machine Learning: A Probabilistic Perspective. *MIT Press.*

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**.

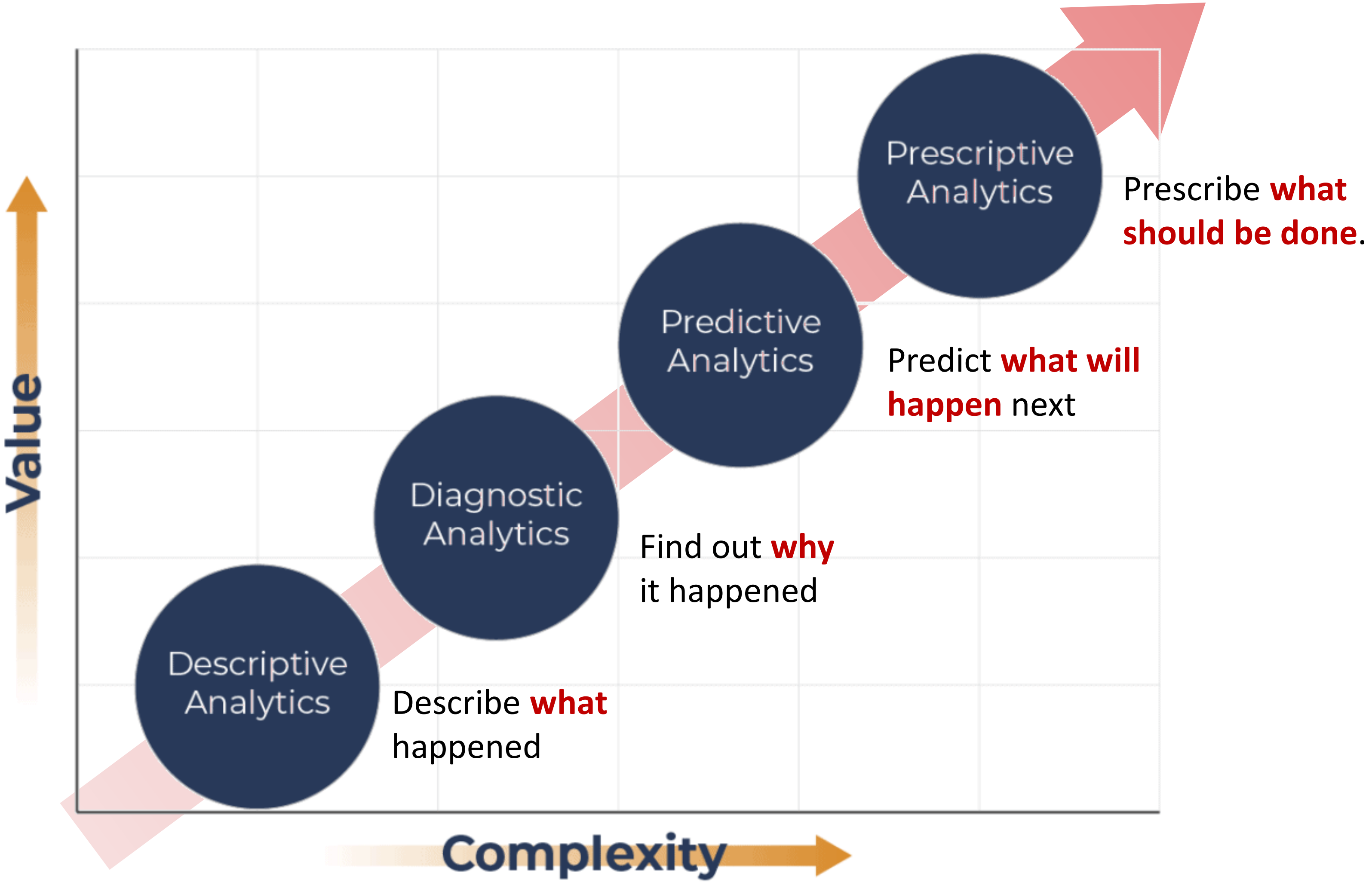


Reference: Pilario et al. (2020), *A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring.* MDPI: Processes, https://doi.org/10.3390/pr8010024

How to turn data into decisions?

Source: https://iterationinsights.com/article/where-to-start-with-the-4-types-of-analytics/

- Applying machine learning to your data is not enough.
- 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**.



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

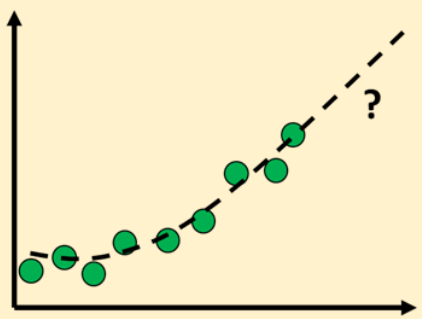
Supervised Learning

Learn a mapping or a function:

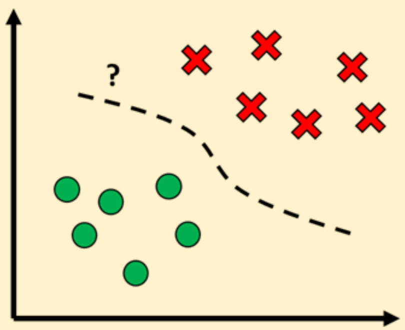
$$y = f(x)$$

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

Regression



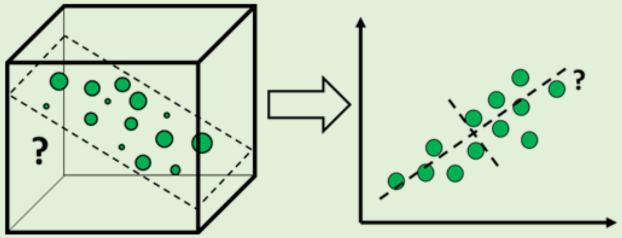
Classification



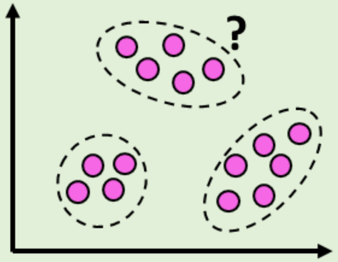
Unsupervised Learning

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

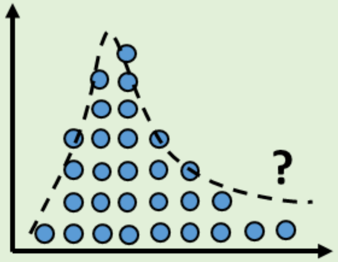
Dimensionality Reduction



Clustering



Density Estimation



Types of Learning Problems

A simple example…

Supervised Learning

These are images of dogs.



These are images of cars.

Now, what is this an image of?

Unsupervised Learning

Here are some images…




Is there an image that does not belong?

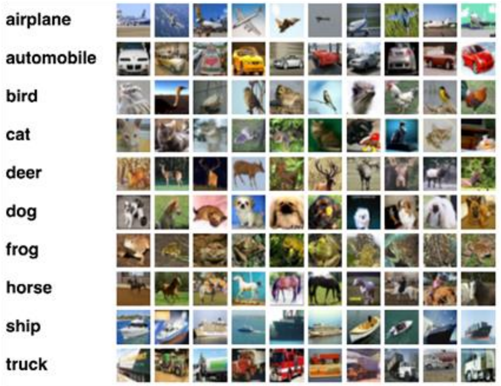
Are there images with similar patterns?

Types of Learning Problems

Semi-Supervised Learning

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

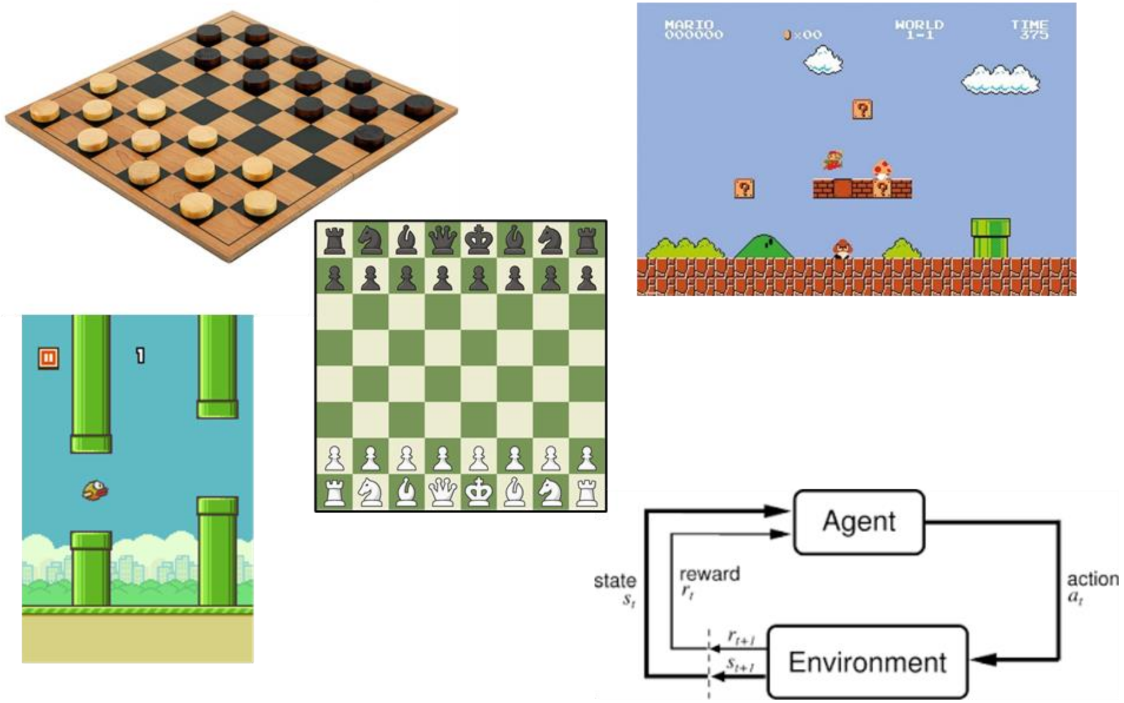
Labelled Data



Unlabelled Data

Reinforcement Learning

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



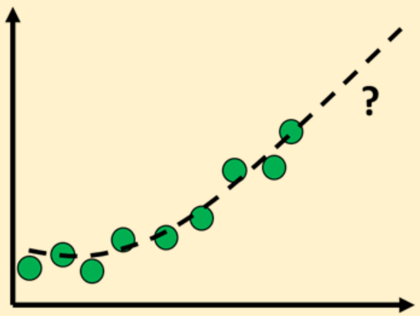
Supervised Learning

Learn a mapping or a function:

$$y = f(x)$$

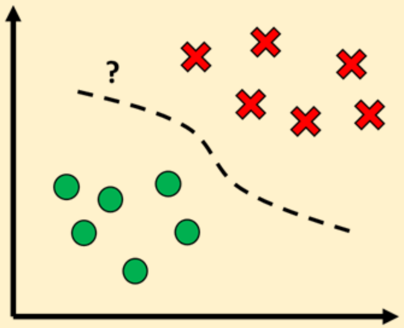
from inputs (x) to outputs (y) , *given a labelled set of input-output examples* (or).

Regression



- **Given:** Training Data $\{\boldsymbol{x}_i, y_i\}_{i=1,2\ldots,N}$
- 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

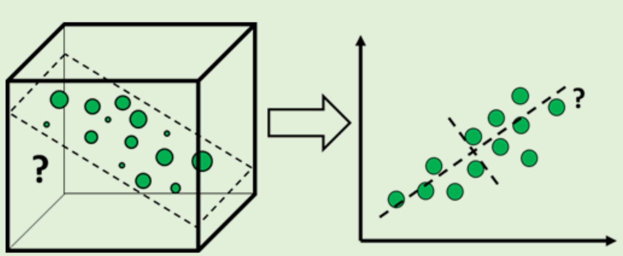


- **Given:** Training Data $\{\boldsymbol{x}_i, y_i\}_{i=1,2\ldots,N}$
- Target y_i is a **categorical** variable.
- Examples:
    - Classifying objects in images
    - Classifying chest X-ray images into COVID positive/negative
    - Handwritten digits recognition
    - Filter e-mails into spam/not spam
    - Classify critical equipment as to healthy or faulty
    - Activity recognition from wearable devices

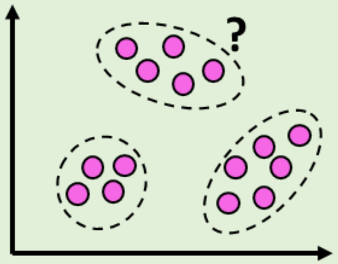
Unsupervised Learning

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

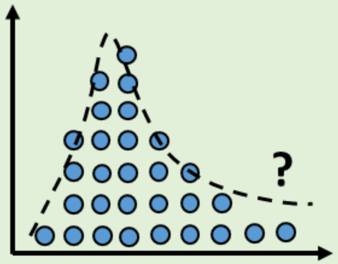
Dimensionality Reduction



Clustering



Density Estimation



Dimensionality Reduction

- **Given:** Data $\{\boldsymbol{x}_i\}_{i=1,2\ldots,N}$
- **Reduce features** but retain the most important information from the original data.
- <u>Examples:</u>
  - Feature Engineering
  - Image compression
  - Filtering noise from signals
  - Source separation in audio
  - Data visualization

Clustering

- **Given:** Data $\{\boldsymbol{x}_i\}_{i=1,2\ldots,N}$
- **Group** similar data points together.
- <u>Examples:</u>
  - Customer segmentation
  - Recommendation systems
  - Identifying fake news
  - Clustering documents, tweets, posts

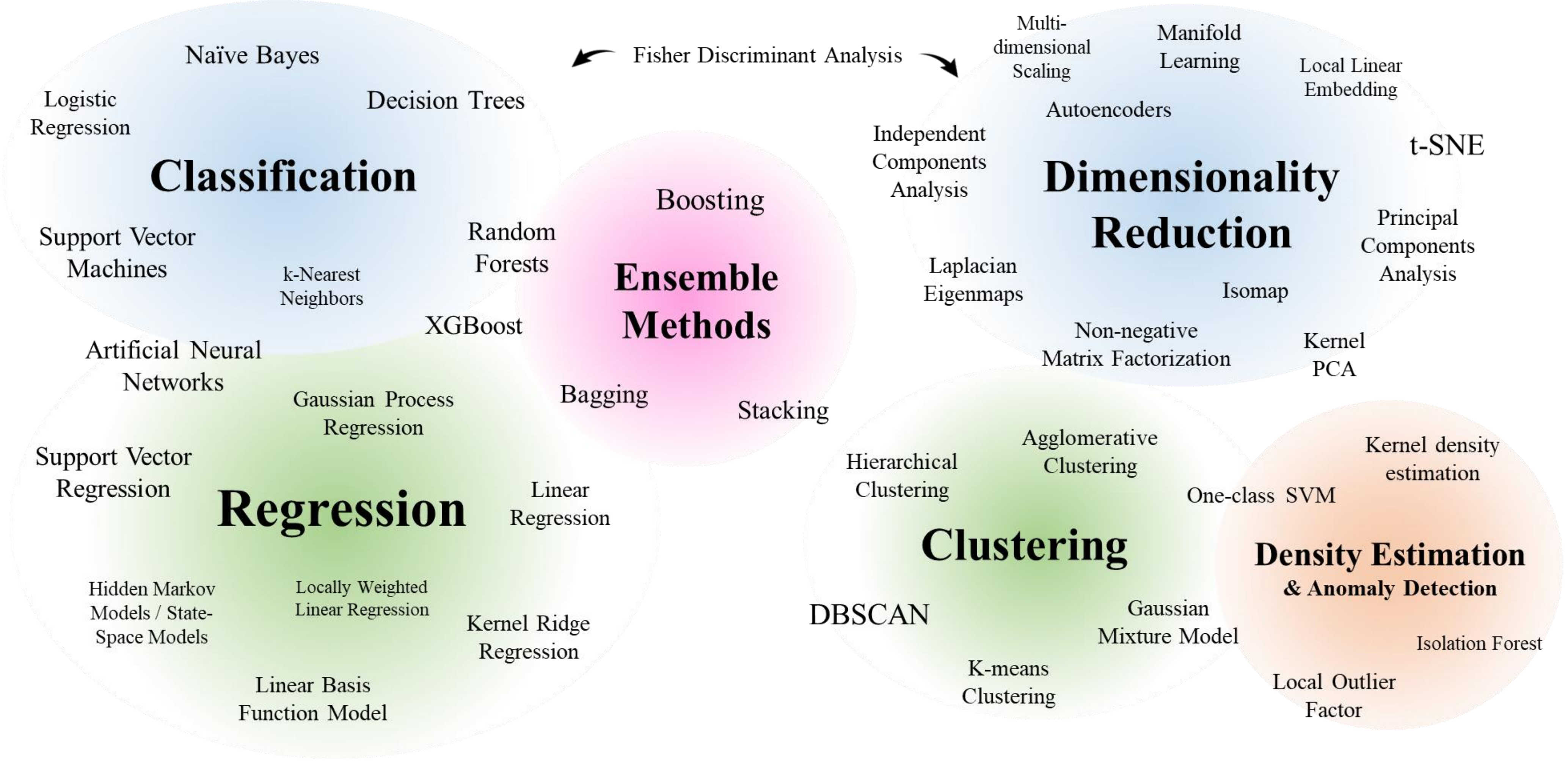
Density Estimation

- **Given:** Data $\{\boldsymbol{x}_i\}_{i=1,2\ldots,N}$
- **Estimate** the distribution of the data.
- <u>Examples:</u>
  - Anomaly Detection
  - Novelty Detection
  - Generative Models
  - Finding distribution modes
  - Spatio-temporal analytics

Machine Learning Methods

Supervised Learning

Unsupervised Learning



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

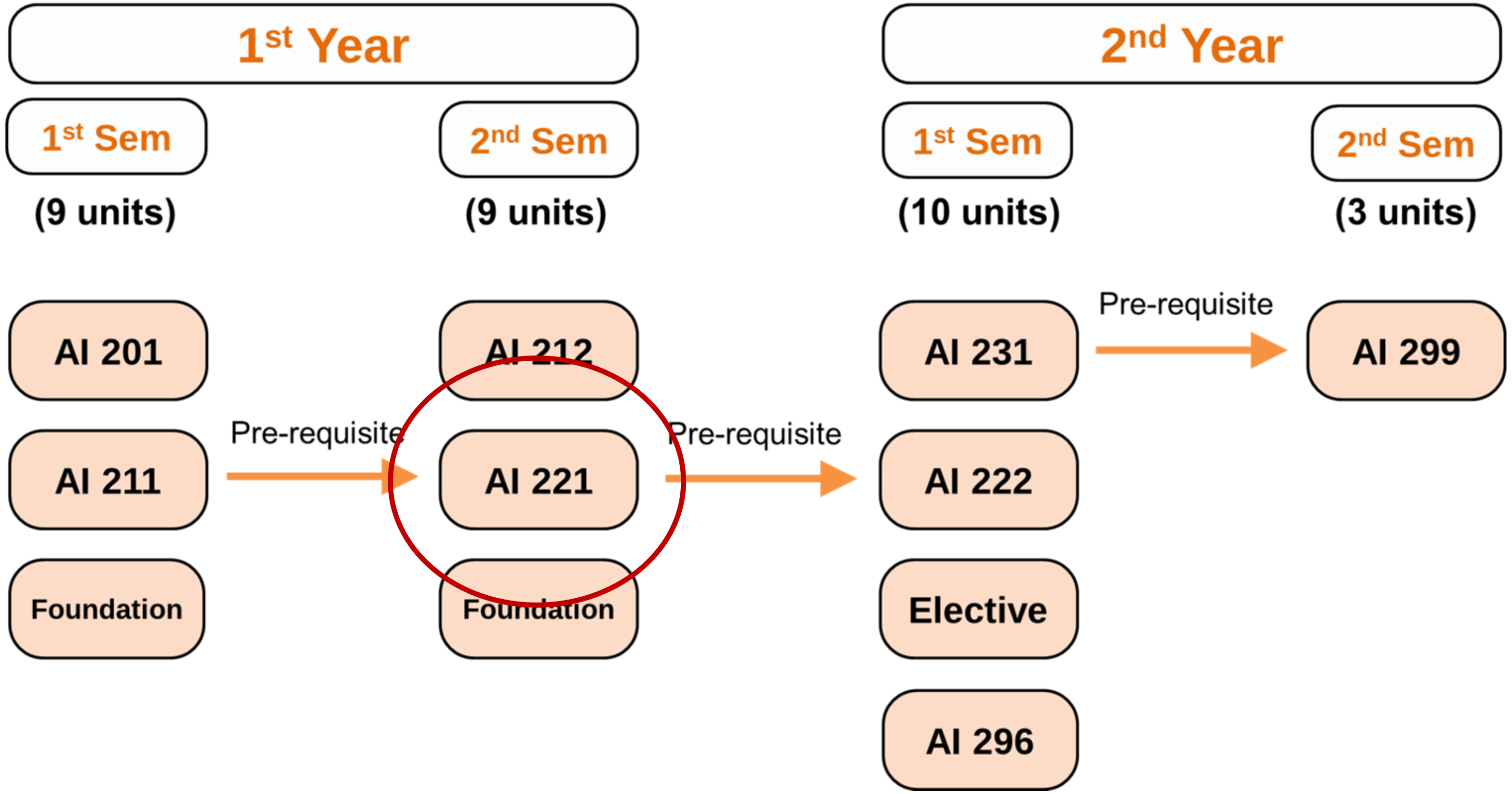
COURSE NUMBER:	AI 221
COURSE TITLE:	Classical Machine Learning
COURSE DESCRIPTION:	Linear Models. Kernel Methods. Neural Networks. Trees. Clustering. Dimensionality Reduction. Feature Engineering. Density Estimation. Ensemble Learning. Gaussian Processes. Bayesian Methods. Hyperparameter Search. AutoML. Explainability.
COURSE CREDIT:	3 units
	3.0 hours/week
COURSE LMS*:	**UVLE Course Page:** **AI 221 [TZZQ]**
Github:	https://github.com/kspilario/AI221

*LMS = Learning Management System

Introduction to the Course

MEngg in Artificial Intelligence

Total Units: 31 Units

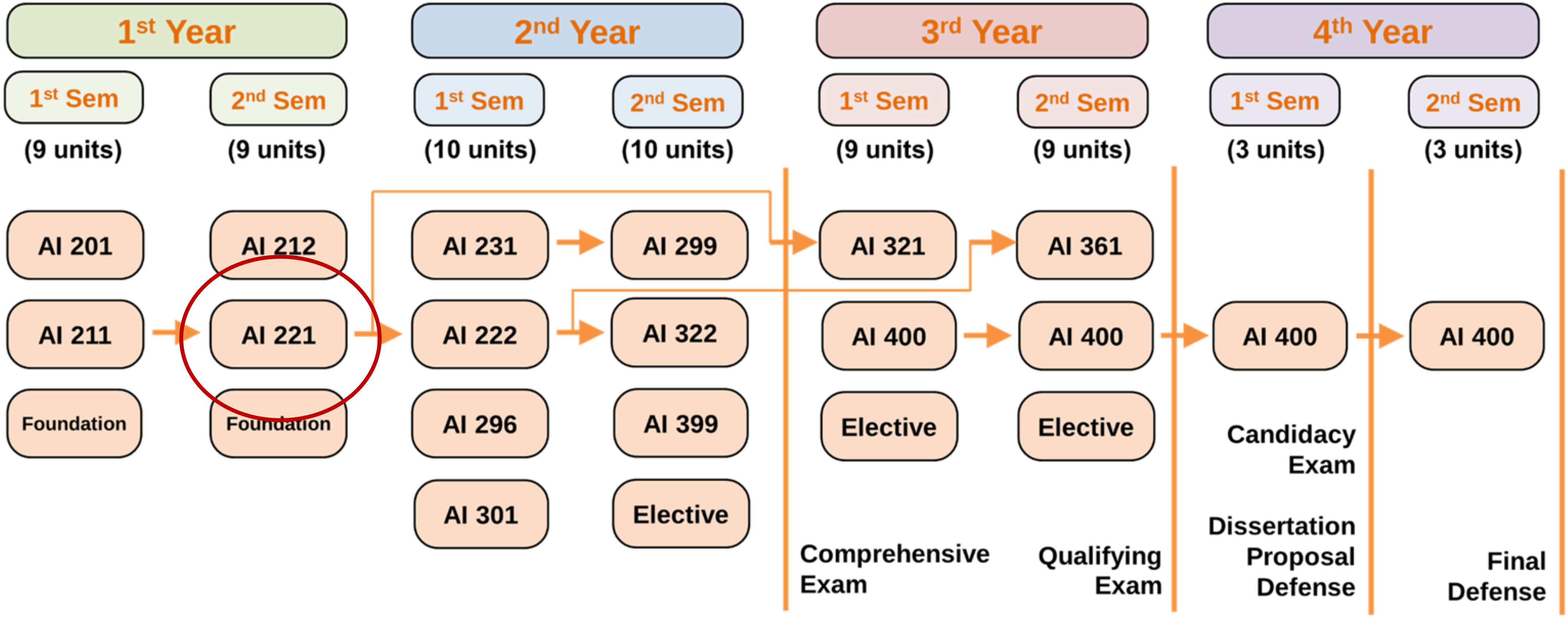


Introduction to the Course

PhD in Artificial Intelligence

Option A

Total Units: 62 Units
Requirement: 2 Publications



AI 221 Course Delivery

- **Meeting:** Every Tuesday, face-to-face, Room A301, Chemical Engineering Building.

Sunday	Monday	**Tuesday**	Wednesday	Thursday	Friday	Saturday
		6-9 PM				

- **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, Take-home
• Journal Critique	**20%**	Individual, Take-home

- **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

Sep 12 Week 1. Introduction to Machine Learning

Sep 19 Week 2. Exploratory Data Analysis

Sep 26 Week 3. Linear and Logistic Regression

Oct 3 Week 4. Support Vector Machines and Kernel Methods

Oct 10 Week 5. Cross-validation and Hyper-parameter Optimization

------------------- Reading Break -------------------

Nov 7 Week 6. Linear Dimensionality Reduction + Discriminant Analysis

Nov 14 Week 7. Nonlinear Dimensionality Reduction

Nov 21 Week 8. Clustering, Density Estimation, and Anomaly Detection

Nov 28 Week 9. Trees, Weak Learners, and Ensemble Learning (Boosting, Bagging, Stacking)

Dec 5 Week 10. Neural Networks for Classification, Regression, and Dim. Reduction

Dec 12 Week 11. Gaussian Processes and Bayesian Optimization

Dec 19 Week 12. AutoML and ML Explainability

Jan 9 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 to the class.
- NO two teams should have the same problem.
- Grading and deadlines:
  - Oral Presentation (40%) – Jan. 9, 2024
  - Written Report (60%) – Jan. 18, 2024

Machine Exercises (40%)

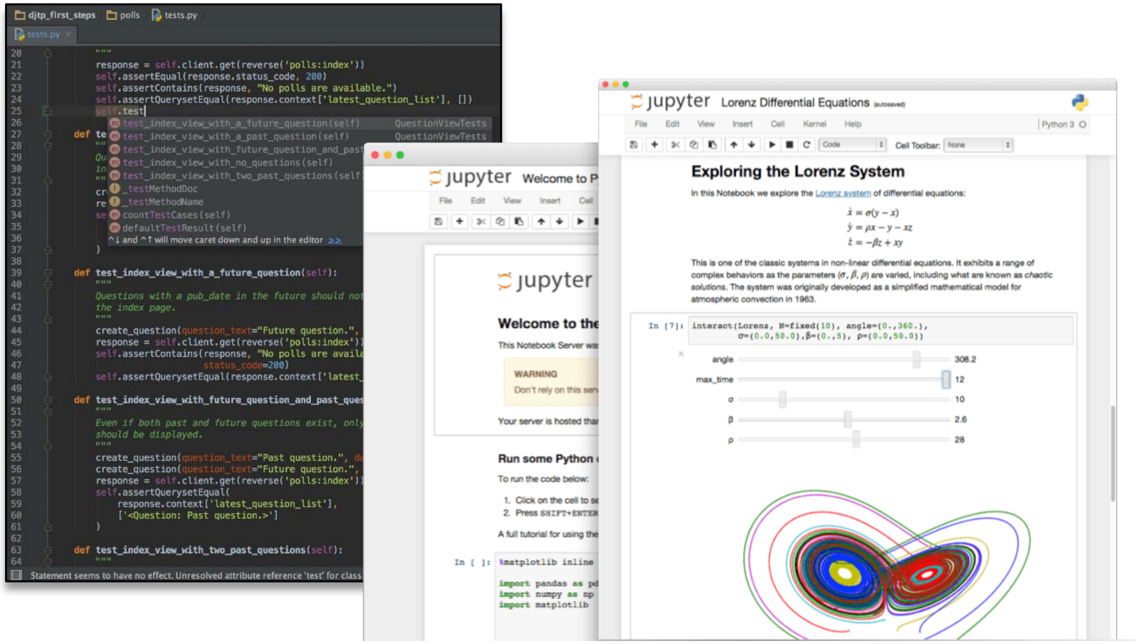
- Mode: Individual, take-home
- To be given every lecture (Weeks 3-12 only).
- Submission deadline is **2 weeks** after release.

Journal Critique (20%)

- Mode: Individual, take-home
- Find a paper from a reputable journal or conference proceedings related to your field.
  - Should at least have an impact factor.
  - Should be published in the last 5 years.
- Send me the paper for approval first, then I will send guide questions for you to answer.
- Deadline: January 6, 2024

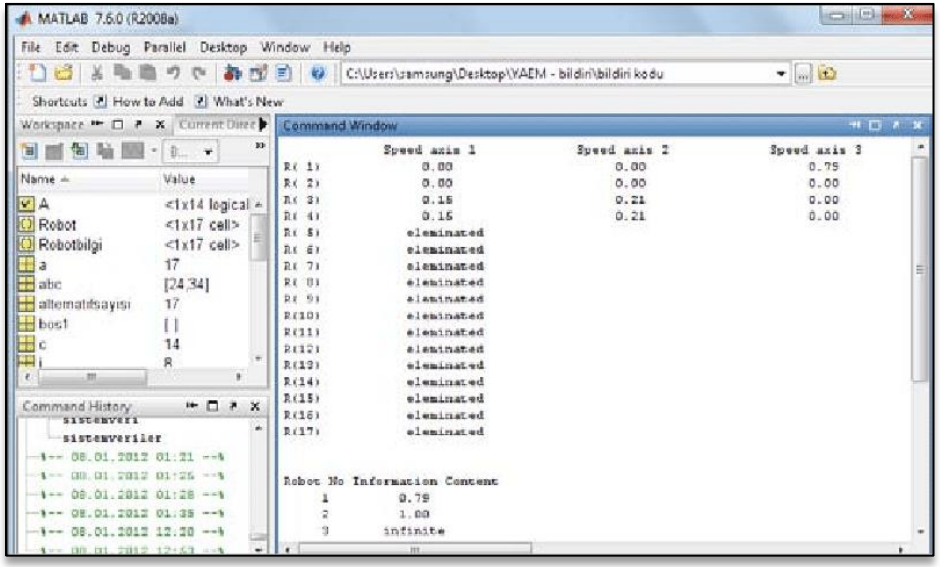
AI 221 Required Software

- Anaconda >> Spyder
- Anaconda >> JupyterLab
- Google Colab
- Jupyter Notebook
- PyCharm
- JupyterLite
- Microsoft Excel (New)



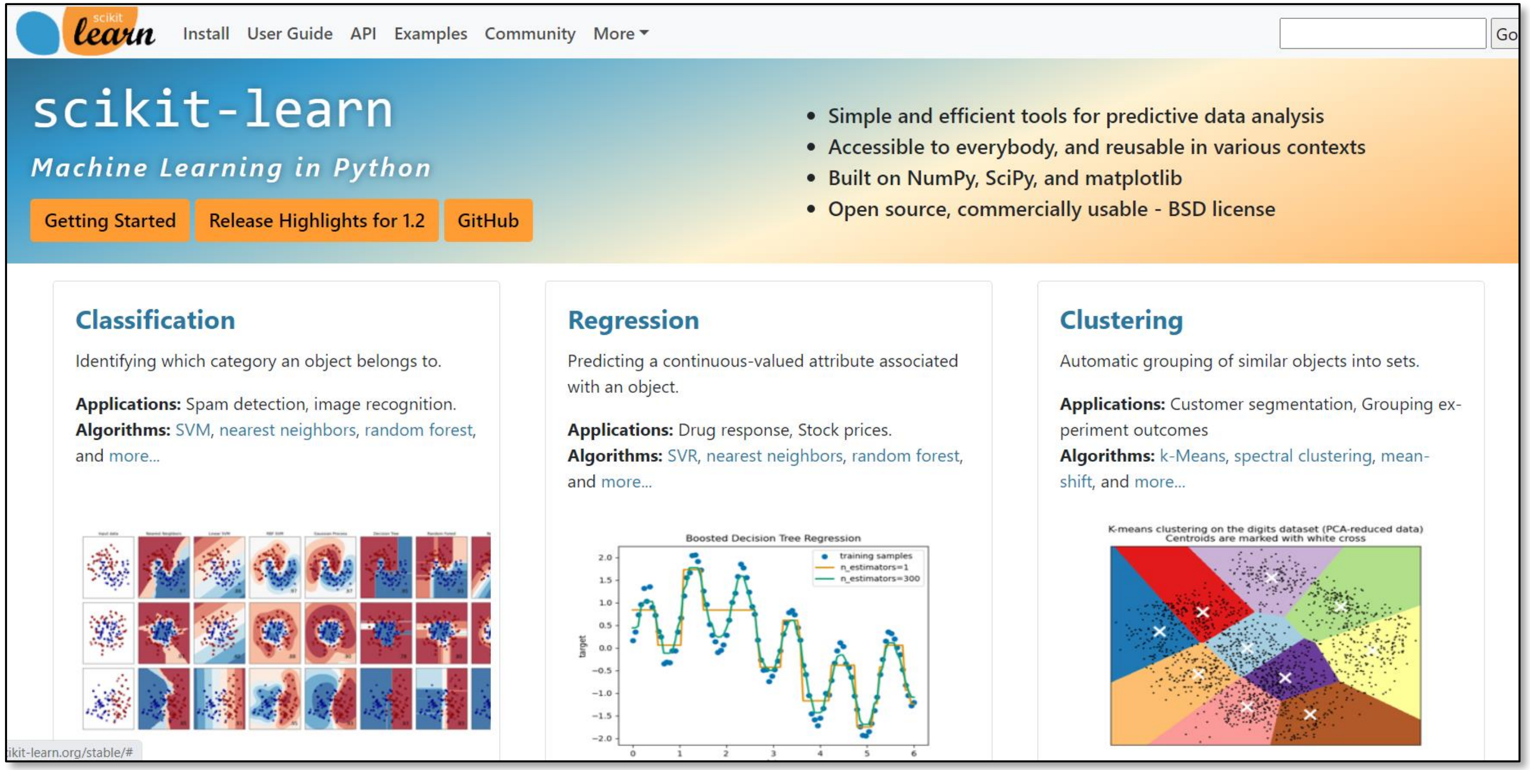
or

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



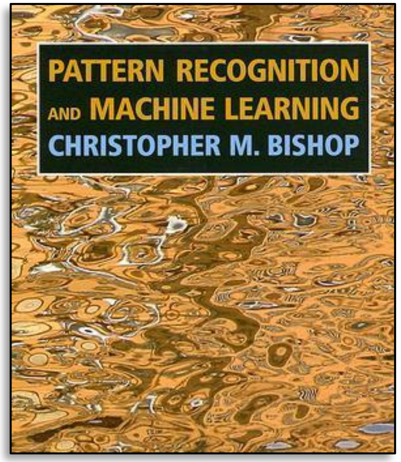
AI 221 Required Software

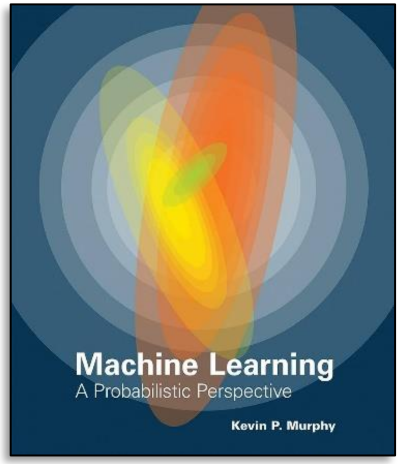
- Python 3
  - https://www.python.org
- Numpy
  - http://www.numpy.org/
- Scikit-Learn
  - https://scikit-learn.org/
- Juptyer Lab
  - https://jupyter.org/try-jupyter/lab/
  - https://nbviewer.org/
- MS Excel

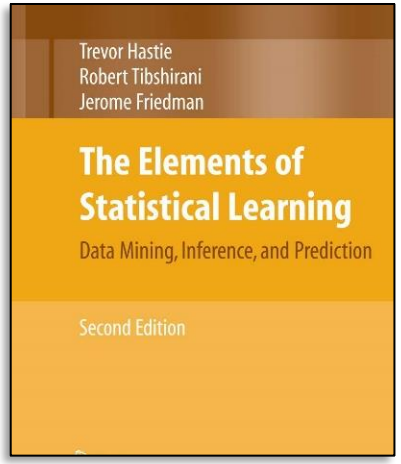


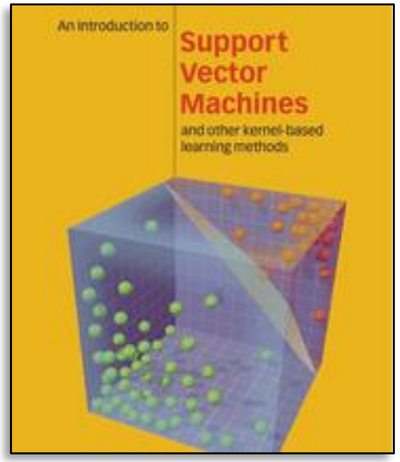
AI 221 References

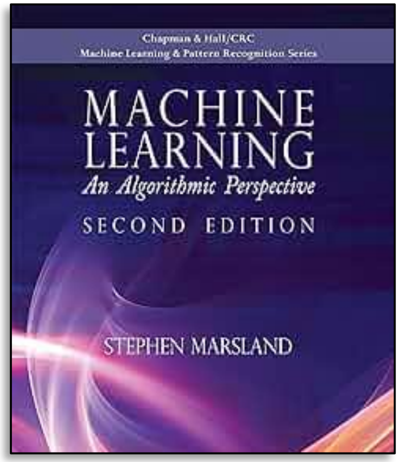
- Bishop (2006). *Pattern Recognition and Machine Learning.* Springer.
- Murphy, Kevin (2012). *Machine Learning: A Probabilistic Perspective.* MIT Press.
- Hastie et al. (2008). *The Elements of Statistical Learning.* 2nd Ed. Springer.
- Cristianini & Shawe-Taylor (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press.
- Marsland, Stephen (2014). *Machine Learning: An Algorithmic Perspective*. Chapman and Hall. 2nd Ed.
- Rasmussen and Williams (2006). *Gaussian Processes for Machine Learning.* MIT Press. https://gaussianprocess.org/gpml/
- Geron, Aurelien (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media.
- Journals and Conference Proceedings
- Python API, Sci-kit learn API: https://scikit-learn.org/stable/modules/classes.html
- Online Courses, Youtube Videos, etc.

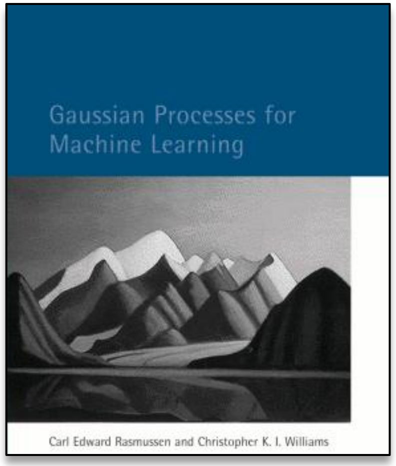


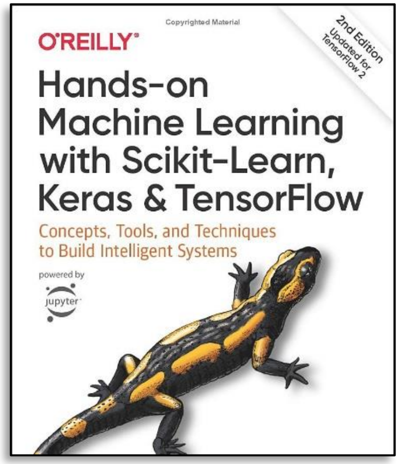












AI 221 Course Instructor

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

University of the Philippines, Diliman

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