LING 506 - TOPICS IN COMPUTATIONAL LINGUISTICS

Introductory Machine Learning

Yan Tang

Department of Linguistics, UIUC

Week 1

Instructor and office hours

- Yan Tang. Assistant Professor
  - yty@Illinois.edu
- Office hours:
  - Walk-in: 3-4 pm, Friday
  - By appointment: 10-11am, Thursday
  - FLB 4023

Course website

- https://uiuc-ling-cl.github.io/ling506_ml/

DRES

If a student has a disability or condition that requires special consideration, the student is expected to present the requisite letter from the University Division of Disability Resources and Educational Services (DRES) no later than the beginning of the second week of class.

Academic integrity

This course follows the University of Illinois Student Code regarding Academic Integrity. The College of Liberal Arts and Sciences also has an excellent web page on the topic. You are expected to read these resources prior to the second day of class, and to understand your responsibilities regarding Academic Integrity.

All work submitted for this class must be solely your own. Violations of Academic Integrity include, but are not limited to, copying, cheating, and unapproved collaboration.

Asking questions and discussion

- Course Piazza site linked off course web page
- Verify that you are enrolled in the course Piazza site
- Ask questions via Piazza
- Do not ask questions via email

Sign up: https://piazza.com/illinois/spring2020/ling506

Home: https://piazza.com/illinois/spring2020/ling506/home

Other business

- https://uiuc-ling-cl.github.io/ling506_ml/syllabus.html
- Student responsibilities
- Absences and late work policy
- etc...

Course Overview

- Concepts of contemporary machine learning (ML)
- Unsupervised ML techniques for clustering
- Supervised ML for classification and regression
- Doing ML on Python
- Focus: practical implementations

More details: https://uiuc-ling-cl.github.io/ling506_ml/schedule.html

Objectives

- To have a broad understanding on ML and its techniques
- To be able to interpret output of ML
- To be able to implement ML on Python to solve simple practical problems
- To be able to perform basic diagnosis on model performance
- To develop self-learning and problem-solving abilities

Course schedule

- Mondays: lecture
- Wednesdays: lecture, discussion and demos
- Fridays: lab sessions

Evaluation

- Lecture attendance, lab exercises and assessments:
  - Attendance: **10%**
  - Mid- and final-term assessments: **20%**
  - Lab exercises: **20%**
- After-class homework: **50%**

- A total of **100%**

Essentials: students

- Python programming skills
  - Not a dedicated course for learning programming!
- Basic git commands
  - *clone*, *add*, *commit*, *pull* and *push*
- Some sense of mathematics and probability
  - No panic! We focus on what and how, less why…

Essentials: hard- and software

- Functional computer
- Python 3
  - Anaconda distribution
- Core Python modules for ML: **scikit-learn**
  - Can be installed via anaconda: https://www.anaconda.com/distribution/
- Git
  - Mac users: do nothing
  - Window users: https://git-scm.com/download/win

https://uiuc-ling-cl.github.io/ling506_ml/syllabus.html

Other useful Python modules

- Matrix and mathematical operations
  - *numpy*
  - *scipy*
- Data structure and manipulation
  - *panda*
- Visualisation:
  - *matplotlib*
  - *yellowbrick*
  - *seaborn*

Applications of ML

- Speech/audio processing:
  - Speech recognition and synthesis
  - Speech enhancement
  - Audio object identification
- Image processing
  - Image object identification
  - Edge detection and optimisation
- Natural language processing:
  - Understanding , answering and reasoning
  - Automatic document summarizing
  - Email spam filtering
- Marketing: customised advertising and automatic recommendation

Applications of ML

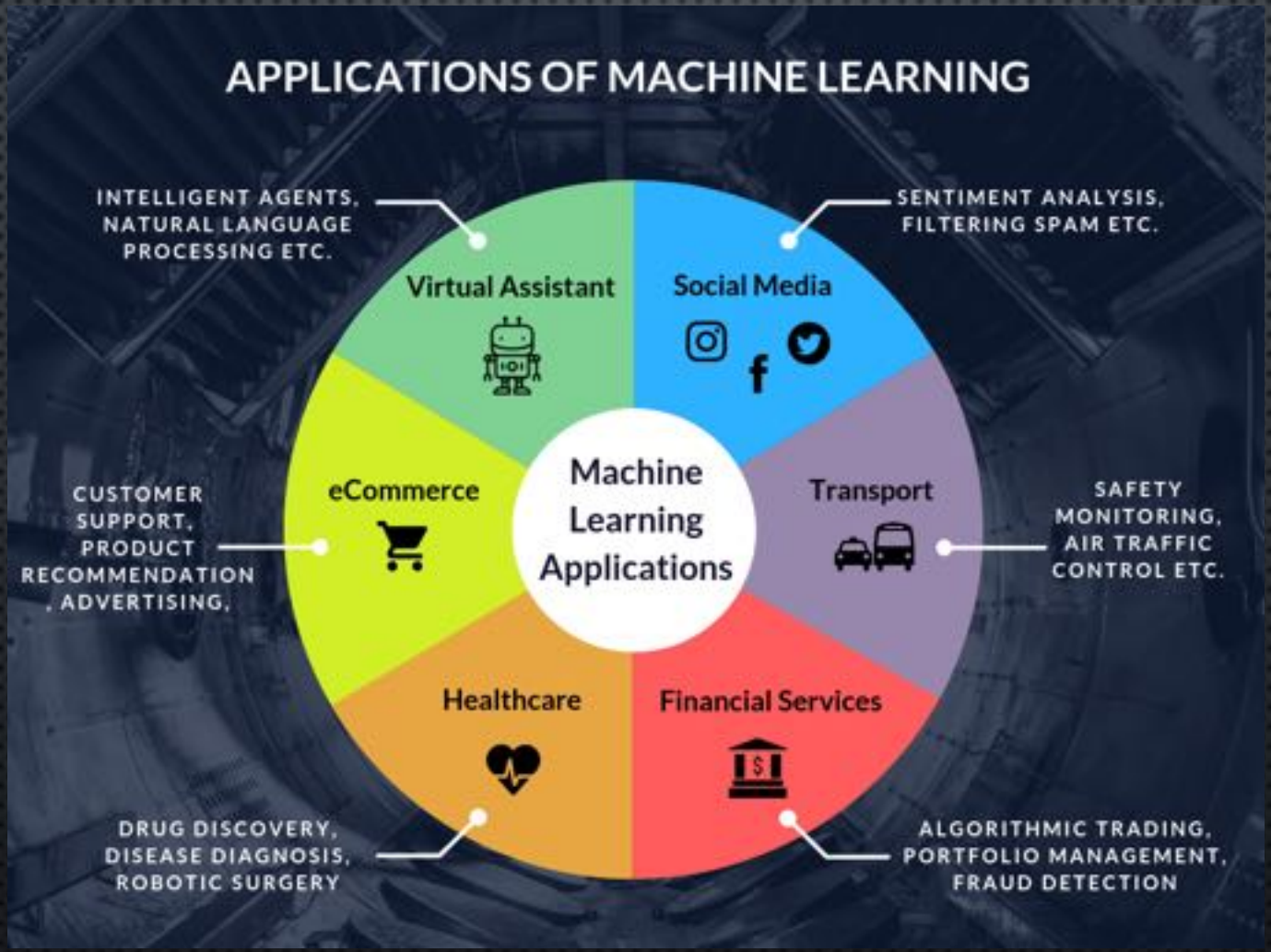


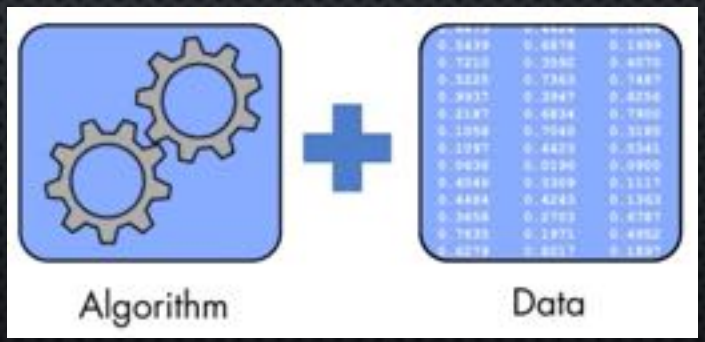
Image Source: https://blog.quantinsti.com/machine-learning-basics/

What is machine learning?

- Using machines to gain meaning from data, and giving machines the ability to learn from data.
- Given some data, how to make the consequence of similar data foreseeable?
  - Algorithmic prediction
- ML – A process during which the machine identifies and analysing the patterns from the given data set, and then makes autonomous decisions based on the knowledge acquired from its previous observations.

How does machine learn?

- Imagine how a human baby learns
  - To be exposed to new things (data)
  - To extract features and patterns of things received
  - To build generalised experiences (models) in mind (using algorithms)
  - To take decisions using on-hand experiences (models) for new things (data)



data – models - data

All is about data!!!

History of ML

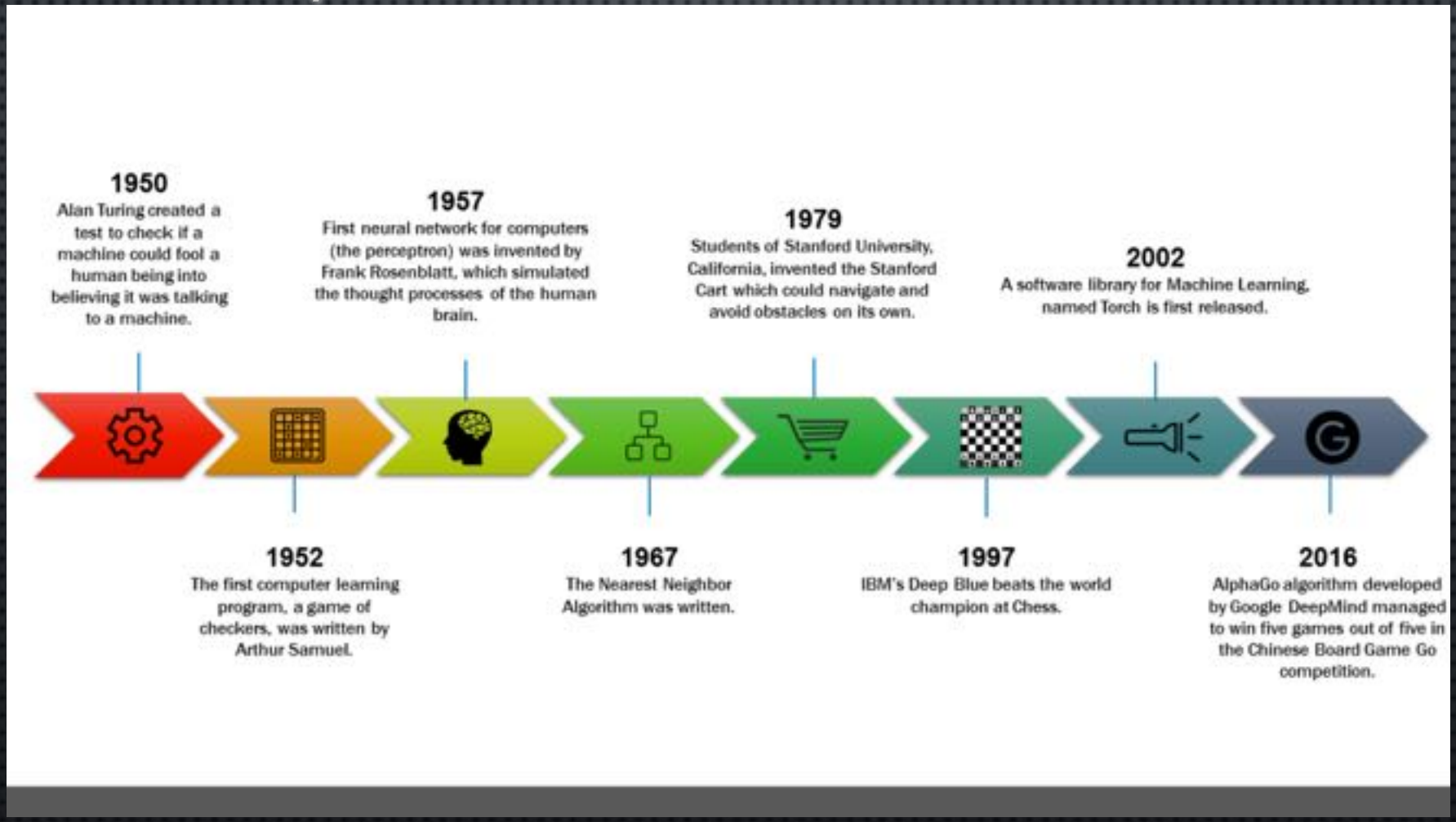


Image Source: https://blog.quantinsti.com/machine-learning-basics/

Relationships between ML, deep learning and artificial intelligence

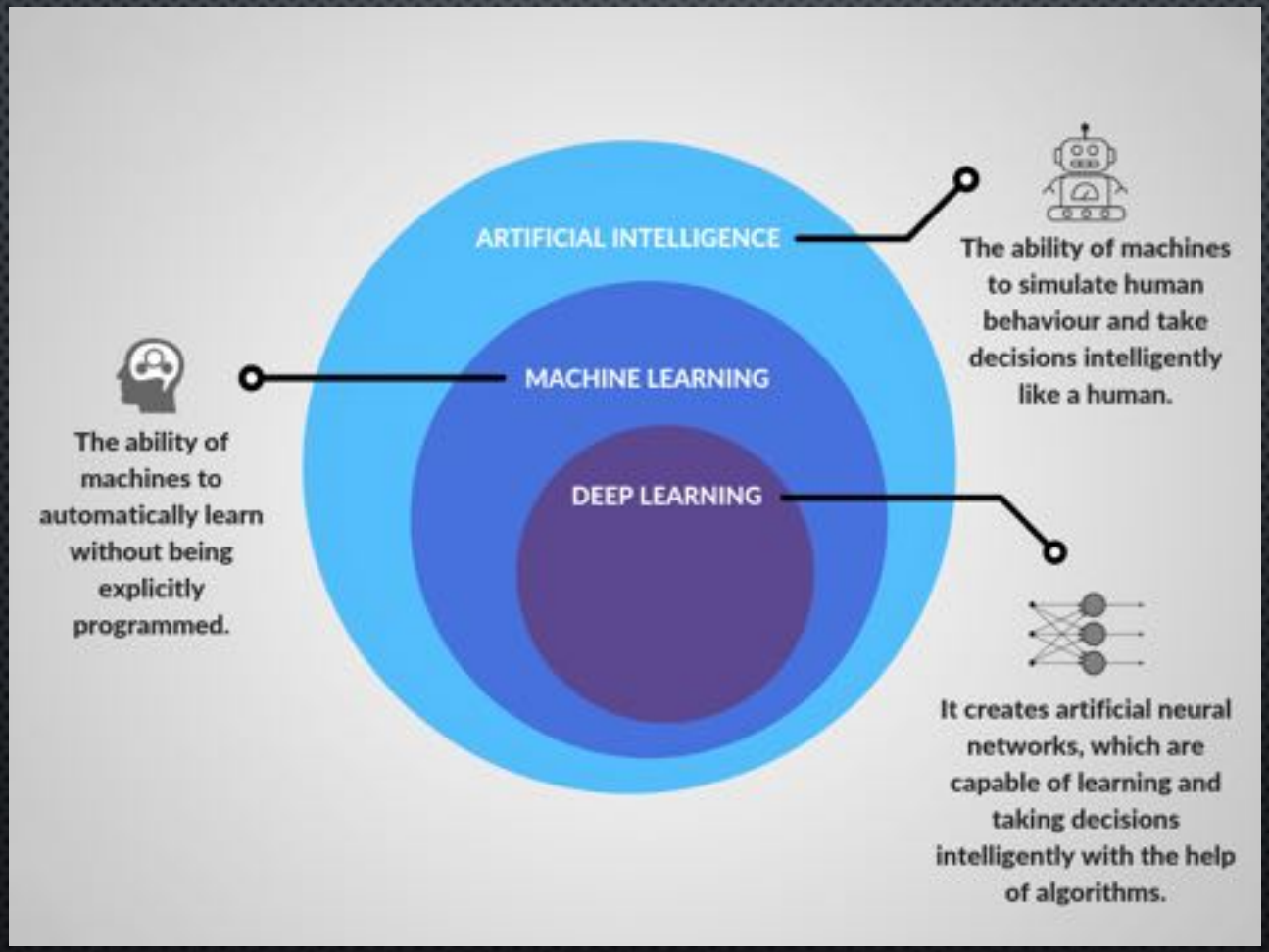


Image Source: https://blog.quantinsti.com/machine-learning-basics/

Types of ML algorithms

- Unsupervised learning
- Supervised learning
- Reinforcement learning

Unsupervised learning

- Resemblance: self-learning student
  - Materials ***without*** correct answers - unlabelled training data
  - Student, the algorithm, discovers, reveals and summaries the underlying structure/distribution of given materials on her own – clustering the data
- Algorithms:
  - K-means clustering
  - Gaussian Mixture model (GMM)

Supervised learning

- Resemblance: teacher and student
  - Teach provides the correct keys to questions - labelled training data
  - Student, the algorithm, learns to answer questions with correct keys.
  - Goal: to continuously progress until a target performance is achieved - trained model
- Algorithms:
  - K-nearest neighbour (KNN) and Naïve Bayes classification
  - Decision trees and random forest
  - Support Vector Machines (SVM)
  - Linear regression

Reinforcement learning

- Resemblance: bounty hunter
  - Hunter, the algorithms, take a decision which will be either rewarded or punished as per prior rules - data
  - Goal: maximising reward in long run - an ideal model
  - Balance between taking new risks and making empirical decision.
- Applications:
  - Gaming, e.g. virtual character
  - Automation control, e.g. lift scheduling

In this course…

- Unsupervised learning
  - Reducing data dimension
  - Clustering: k-means and Gaussian Mixture models
  - Interpretation of natural patterns in data

- Supervised learning
  - Building classification and regression models
  - Algorithms: k-NN, Naïve Bayes, SVM, Trees
  - Creating simple feed-forward NN

Challenges to ML

- Insufficiency of training data
  - Data is more dominant than algorithms
- Nonrepresentative training data
  - Sampling bias
- Data with poor quality
  - Outliners or wrong labels
  - Missing data for features
- Irrelevant data
  - Increases model working load

Challenges to ML

- Overfitting the training data
  - Model performs well only with training data, but significantly less well with unseen data, i.e. cannot generalise.
- Underfitting the training data
  - Model is too simple to capture the underlying structure of the data

Friday: lab session