Lecture 1

Introduction to Data Mining

Golnaz Taheri, PhD
Senior Lecturer, Stockholm University

Stockholms universitet

Course Logistics

- ○ https://ilearn.dsv.su.se/course/view.php?id=1678

- Course activities:
  - ○ 7 weeks: 35 - 41
  - ○ 12 Lectures: Aug 28 – Oct 13
  - ○ 6 Labs & Q&A sessions
  - ○ Written Exam: Oct 20

- Instructor and Responsible teachers:

○ Golnaz Taheri	golnaz.taheri@dsv.su.se
○ Ioanna Miliou	ioanna.miliou@dsv.su.se

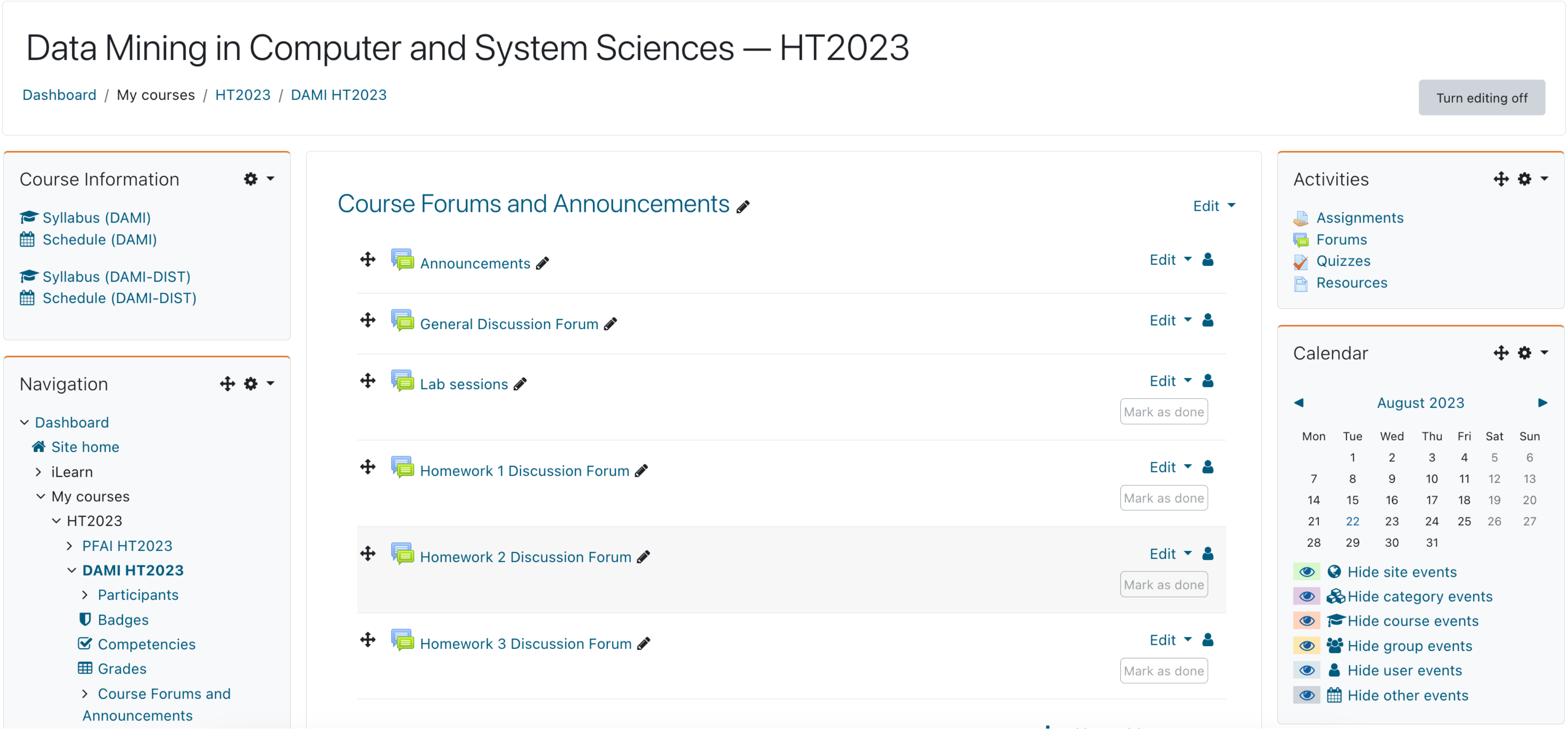
Course Logistics

- Course Assistants:

  - Luis Quintero luis-eduardo@dsv.su.se
  - Maria Bampa maria.bampa@dsv.su.se
  - Zed Lee zed.lee@dsv.su.se

Stockholms universitet

Course page on ilearn



Data Mining in Computer and System Sciences — HT2023

Dashboard / My courses / HT2023 / DAMI HT2023

Course Information

- Syllabus (DAMI)
- Schedule (DAMI)
- Syllabus (DAMI-DIST)
- Schedule (DAMI-DIST)

Navigation

- Dashboard
  - Site home
  - iLearn
  - My courses
    - HT2023
      - PFAI HT2023
      - DAMI HT2023
        - Participants
        - Badges
        - Competencies
        - Grades
        - Course Forums and Announcements

Course Forums and Announcements

- Announcements
- General Discussion Forum
- Lab sessions
- Homework 1 Discussion Forum
- Homework 2 Discussion Forum
- Homework 3 Discussion Forum

Activities

- Assignments
- Forums
- Quizzes
- Resources

Calendar

◀ August 2023 ▶

Mon	Tue	Wed	Thu	Fri	Sat	Sun
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

- Hide site events
- Hide category events
- Hide course events
- Hide group events
- Hide user events
- Hide other events

Stockholms universitet

Course Syllabus

Week 35	Week 36	Week 37	Week 39	Week 40	Week 41
Introduction to Data Mining 08/28	Dimensionality reduction 09/04	Clustering || 09/11	Classification || 09/25	Model evaluation 10/02	Advanced Topics II Graph Mining 10/09
Introduction to Python 08/30	Data Preparation using Python 09/06	Clustering using Python 09/13	Classification using Python 09/26	Advanced Topic | Neural Network 10/03	Deployment 10/10
Association Rules 08/31	Clustering | 09/07	Classification | 09/15	Classification ||| 09/29	Model Evaluation 10/05	Exam Review 10/13

Course workload

Assignments 3 hp

- Three programming assignments (Python)
- Online quizzes

Written Exam 4.5 hp

Homework Assignments

- To be done individually (strictly)
- Will involve programming in Python
- Each corresponding to a lab session

Plagiarism is not acceptable, such as:

- Borrowing code from the internet (chatGPT) and submitting it as is or with minor changes
- Borrowing code from each other and submitting it as is or with minor changes
- Borrowing code from previous years and submitting it as is or with minor changes

Homework Assignments

Submissions:

- Before a given deadline
- Late submissions: Not Allowed
- Second deadline: November 15th
  - OBS: penalty of 50% off the obtained grade

Quizzes

- 6 weekly online quizzes (lowest quiz grade to be dropped, and the best five will count)
- Questions on previous lectures (1 to 3 lectures)
- Only one attempt per quiz
- All quizzes will be timed!
- No make-up quizzes possible!

Homework Assignments

- HW1: 4 pts
- HW2: 5 pts
- HW3: 6 pts
- Quizzes: 5 pts
- Total points: 20 pts
- To pass you need 12 pts
- Grading scheme: A – F

Exam

- Two versions:
  - **DAMI**: on-campus
  - **DAMI-DIST**: online
- Two parts:

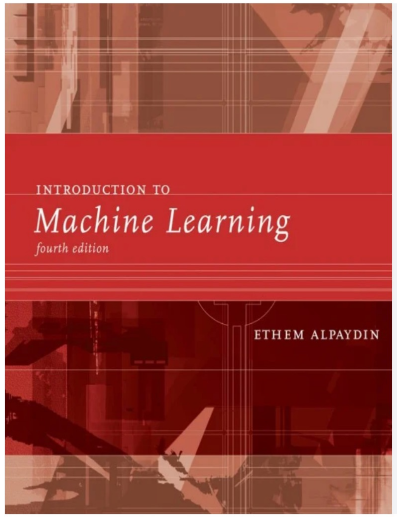
  Part A: multiple-choice questions

  Part B: free text questions
- This will examine your ability on what you have learned
- To pass you need at least 60% of the points
- Grade scheme: A – F

Course textbook

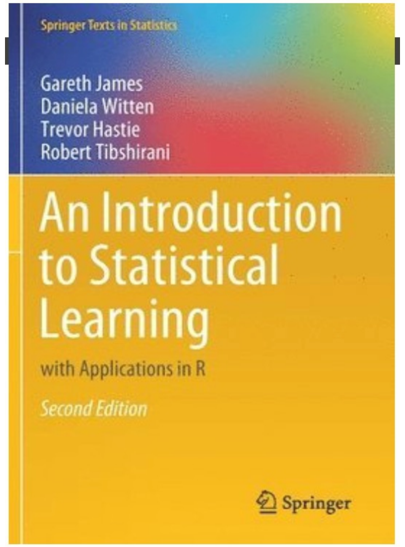
Main:

- Introduction to Machine Learning (fourth edition)
  Publisher: MIT Press
  Year: 2022
  ISBN: 978-0-2620-4379-3



Additional:

- An Introduction to Statistical Learning with applications in R
  Publisher: Springer
  Year: 2013
  ISBN: 978-1-4614-7138-7
  URL: http://www-bcf.usc.edu/~gareth/ISL/



Learning Objectives

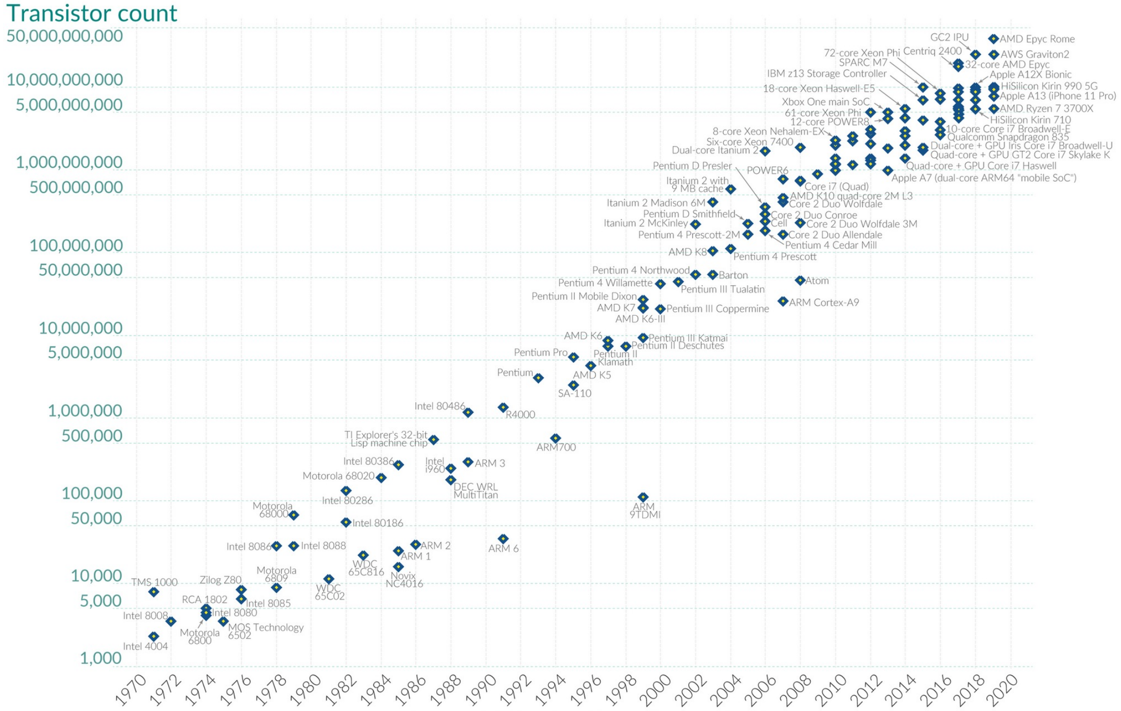
- Become familiar with data science and its algorithms
- Be able to identify a correct algorithmic solution to a given problem
- Be able to apply these algorithmic solutions to solve practical problems
- Be able to perform basic data analysis on real data using Python

Introduction

- Why we need Data Analysis?
- What is Data Science (DS) ?
- What is Data Mining (DM) ?
- What is Artificial Intelligence (AI)?
- What is Machine Learning (ML) ?

Why we need Data Analysis?

- Computational power
  - More efficient processors, larger memories



Moore’s Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World in Data

Stockholms universitet

Why we need Data Analysis?

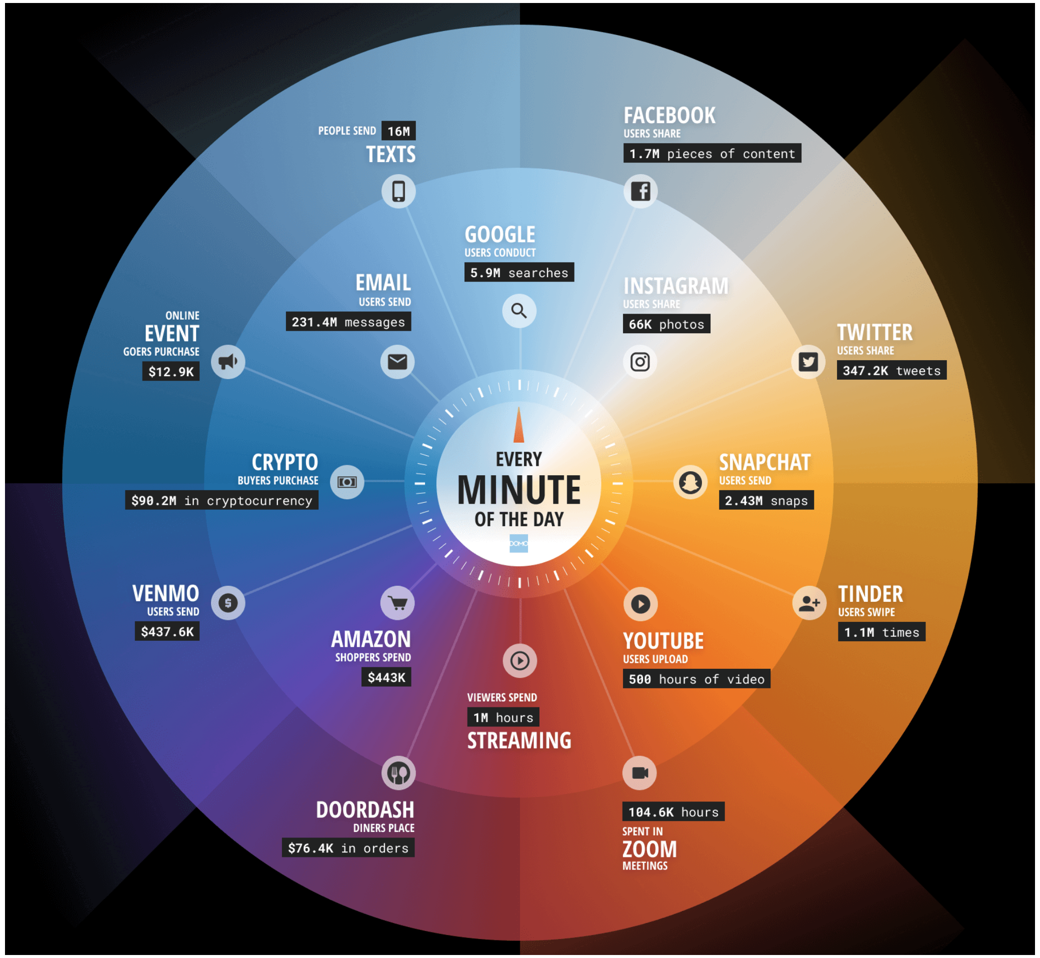
- Data collection and transfer
  - Communication and measurement technologies have improved
- Data storage
  - Huge hard disks
  - Data on the cloud

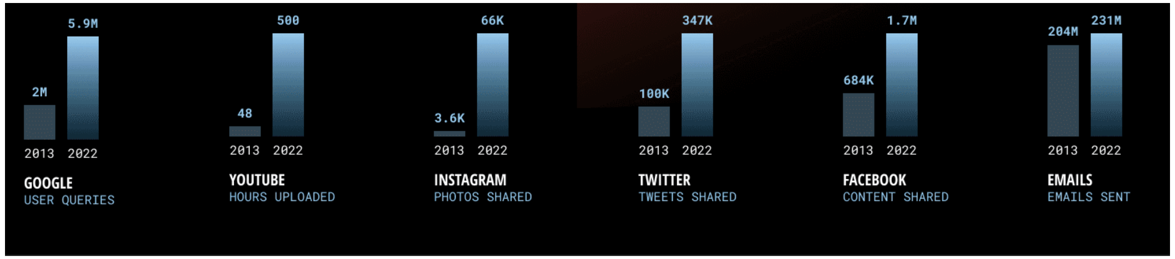
Stockholms universitet

Why we need Data Analysis?

- It is possible to collect and store lots of raw data
- But…data analysis methods are lagging behind
- Need to analyze the raw data to extract knowledge

Data Never Sleeps!





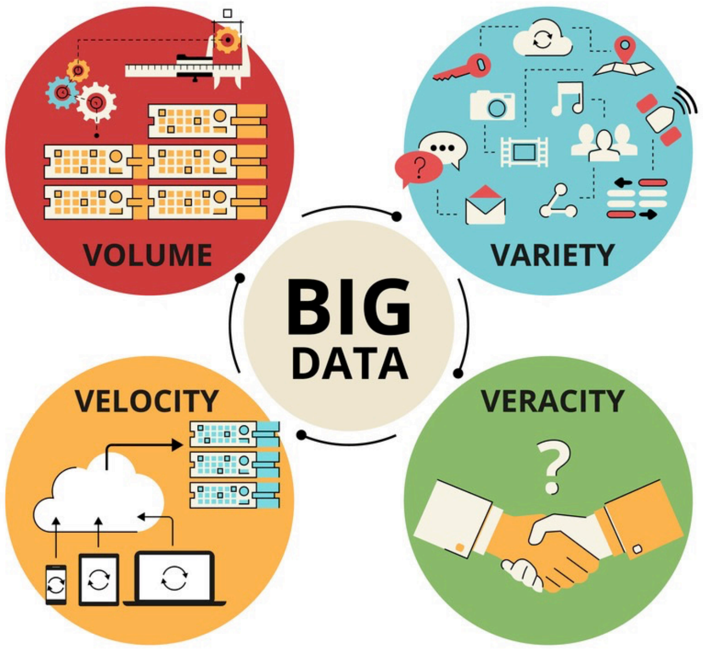
The Four V's of Big Data

Volume: The first V of big data is all about the amount of data.

Velocity: The second V of big data, is all about the speed new data is generated and moves around.

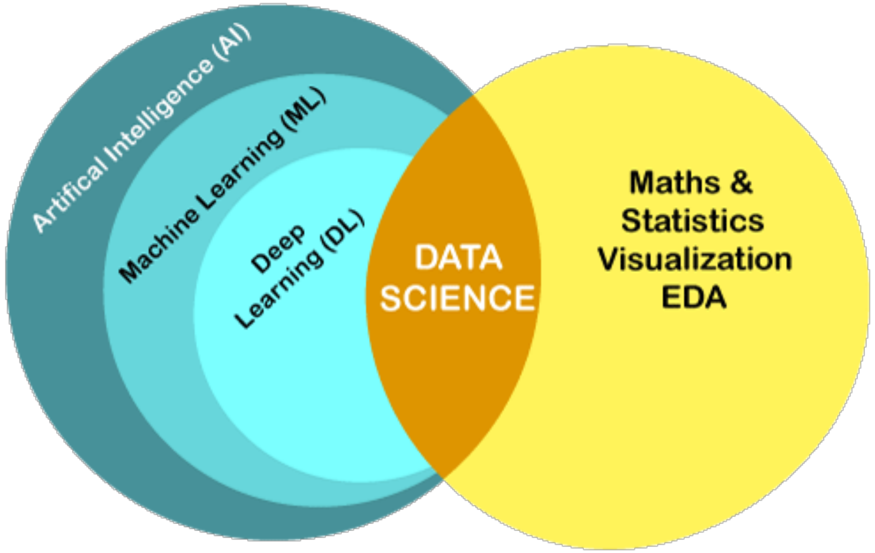
Variety: Data is generally one of three types: unstructured, semi-structured and structured and algorithms required to process the variety of data generated varies based on the type of data to be processed.

Veracity: Denotes the trustworthiness of the data. Is the data accurate and high-quality?



What is Data Science

- Data science is an interdisciplinary field that uses statistics, scientific computing, scientific methods, algorithms and systems to extract knowledge and insights from structured, and unstructured data.

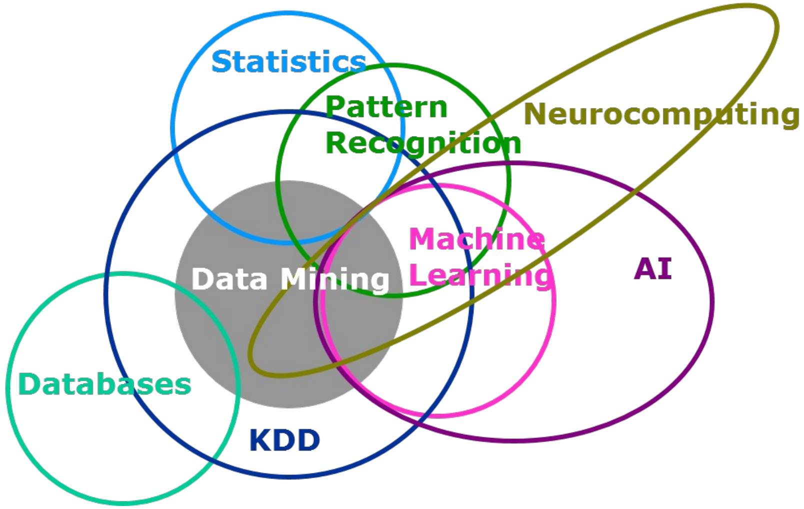

Artifical Intelligence (AI)
Machine Learning (ML)
Deep Learning (DL)
DATA SCIENCE
Maths & Statistics Visualization EDA

Why Data Science is important?

- Data science is revolutionizing the way companies operate. Many businesses, regardless of size, need a robust data science strategy to drive growth and maintain a competitive edge.
  - DS allows businesses to uncover new patterns and relationships that have the potential to transform the organization
  - DS can reveal unnoticed gaps and problems. Greater insight about purchase decisions, customer feedback, and business processes can drive innovation in internal operations and external solutions.
  - DS can help companies predict change and react optimally to different circumstances.

What is Data mining

- Data mining is the process of extracting and discovering patterns in large data sets.
- The overall goal of data mining is extracting information (with intelligent methods) from a data set and transforming the information into an understandable structure.
- Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

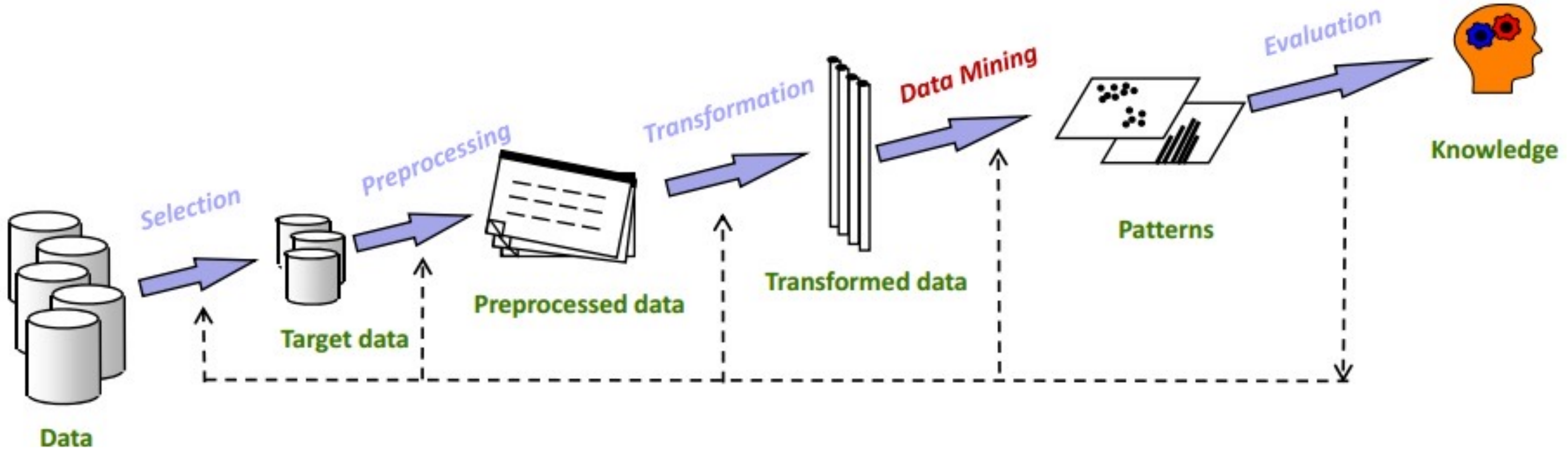


Stockholms universitet

What is KDD

- Knowledge Discovery in Databases (KDD) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets.
- KDD is a multi-step process that encourages the conversion of data to useful information. Data mining is one of the steps of KDD which is the pattern extraction phase of KDD.

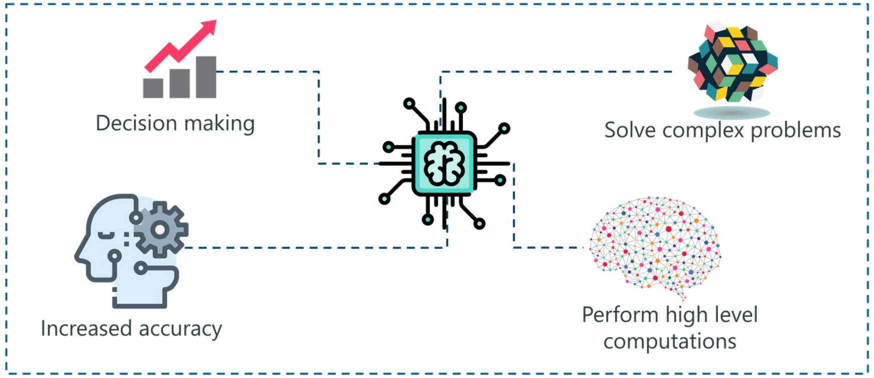
[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



What is Artificial Intelligence

- In 1956, the term Artificial Intelligence (AI) was defined by John McCarthy as:

  ‘The science and engineering of making intelligent machines.’

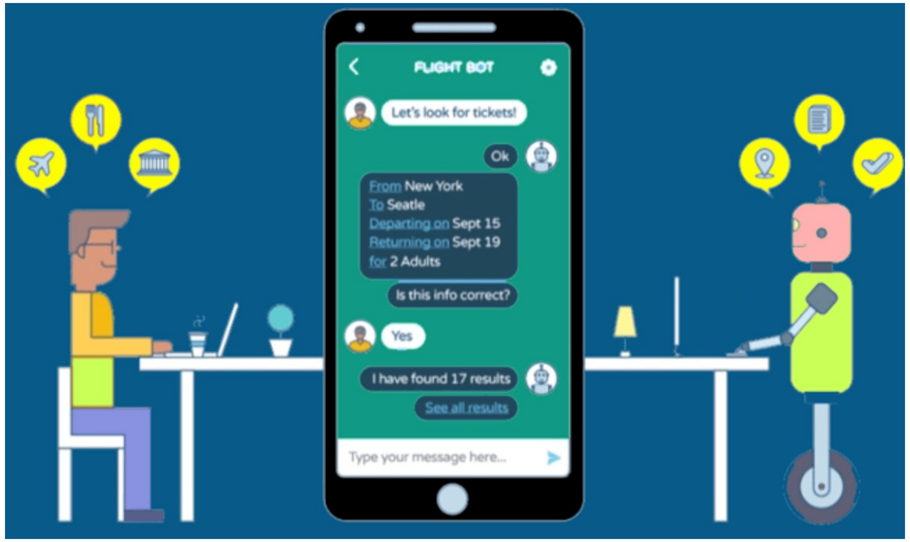


- AI is a machine's ability to perform the cognitive functions we associate with human minds, such as reasoning, learning, interacting with an environment, problem solving, and even exercising creativity.

AI example

- Chatbots

- Answering a customer's inquiries can take a long time.
- The use of algorithms to train machines to meet customer needs through chatbots is an AI solution.
- This allows machines to answer as well as take and track orders.



Let's look for tickets!

Ok

From New York
To Seatle
Departing on Sept 15
Returning on Sept 19
for 2 Adults

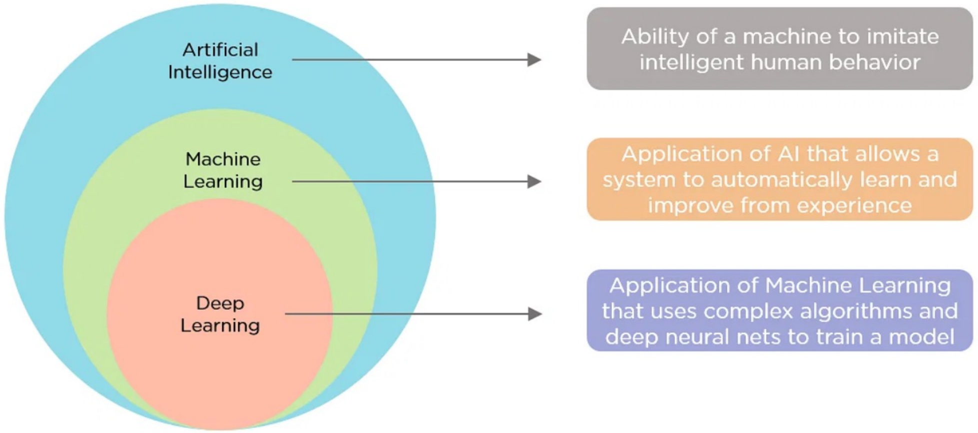
Is this info correct?

Yes

I have found 17 results

What is Machine Learning

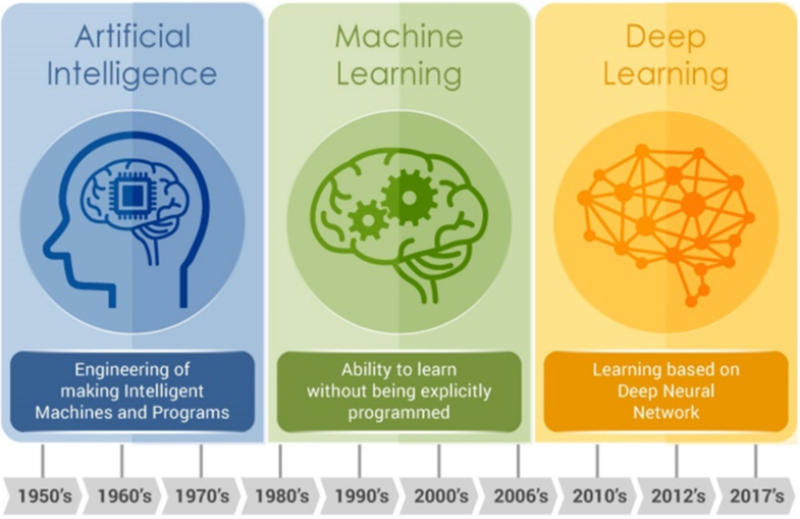
- Machine learning is a branch of AI and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.



Ability of a machine to imitate intelligent human behavior

Application of AI that allows a system to automatically learn and improve from experience

Application of Machine Learning that uses complex algorithms and deep neural nets to train a model

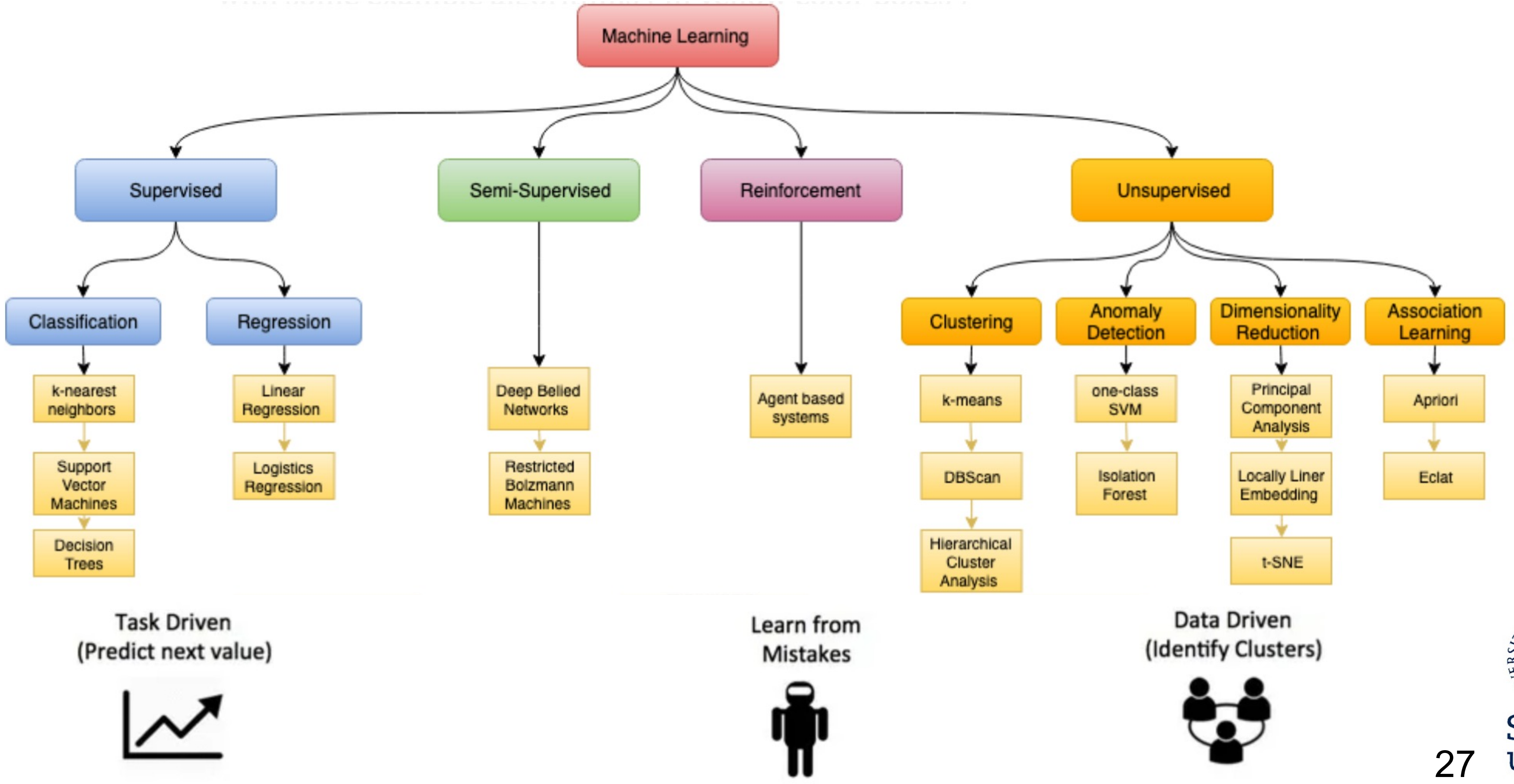


Engineering of making Intelligent Machines and Programs

Ability to learn without being explicitly programmed

Learning based on Deep Neural Network

Types of Machine Learning



Machine Learning

- Supervised
  - Classification
    - k-nearest neighbors
    - Support Vector Machines
    - Decision Trees
  - Regression
    - Linear Regression
    - Logistics Regression
- Semi-Supervised
  - Deep Belied Networks
  - Restricted Bolzmann Machines
- Reinforcement
  - Agent based systems
- Unsupervised
  - Clustering
    - k-means
    - DBScan
    - Hierarchical Cluster Analysis
  - Anomaly Detection
    - one-class SVM
    - Isolation Forest
  - Dimensionality Reduction
    - Principal Component Analysis
    - Locally Liner Embedding
    - t-SNE
  - Association Learning
    - Apriori
    - Eclat

Task Driven
(Predict next value)

Learn from
Mistakes

Data Driven
(Identify Clusters)

Thanks!

golnaz.taheri@dsv.su.se