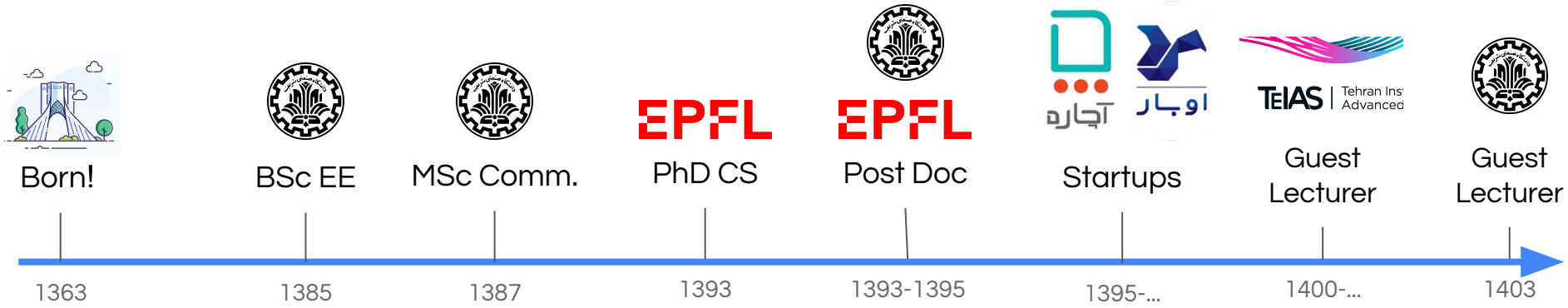
Applied Data Science

1403-02

Hamed Shah-Mansouri
hamedsh@sharif.edu

Amir Hesam Salavati
saloot@gmail.com

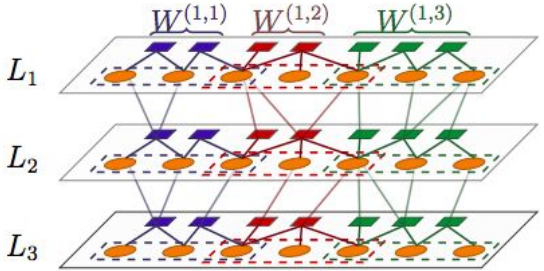
About Me


Born!
1363
BSc EE
1385
MSc Comm.
1387
EPFL
PhD CS
1393
EPFL
Post Doc
1393-1395
آچاره
اوبار
Startups
1395-...
TEIAS
Guest Lecturer
1400-...
Guest Lecturer
1403

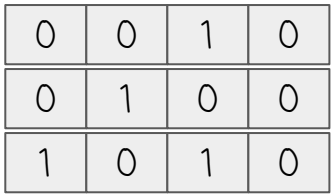
How I Got Here: Research

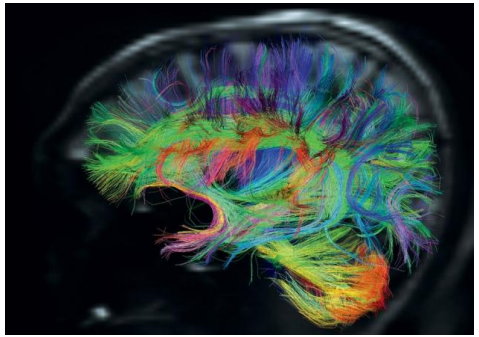
MSc:
Error Correcting Codes

PhD: *Enhanced* Neural Associative Memories

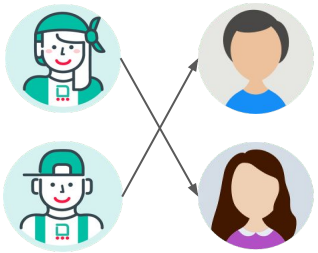


PostDoc:
Connectome Mapping
from Neurons ***Data***



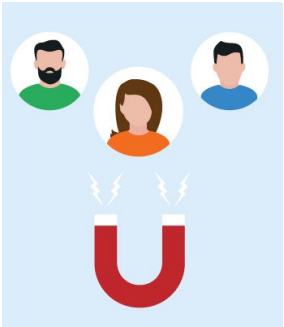


How I Got Here: Work



Matching

Personalization/Recommendation



Credit: clickdimensions.com

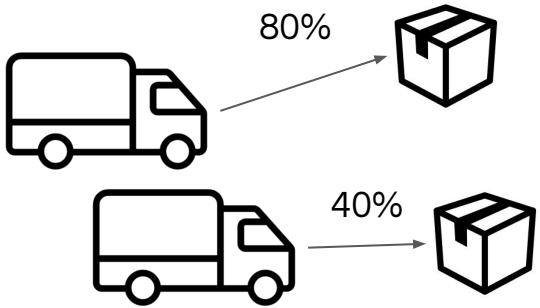
Customer Retention

Credit: friss.com

Fraud Detection

How I Got Here: Work

Price Prediction



Matching

& Probability Estimation

About this Course

Credit: https://www.inklyo.com/how-to-succeed-online-course/

Syllabus

- Introduction to data science
- Data preparation
  - Exploratory data analysis
  - Cleaning and preprocessing
- Data visualization and reporting
  - Web crawling and data scraping
- Different type of problems in data science
  - Supervised vs. unsupervised
  - Classification, regression, association
- Various metrics to measure the accuracy of algorithms
- Regression problems and famous algorithms

Syllabus (Continued)

- Classification problems and famous algorithms
  - SVMs
  - Decision trees and their extensions
  - Multiclass problem
- Overfitting and techniques to mitigate that
- Neural networks
  - Intro
  - Deep learning (in image classification)
- Generative AI
- Model explainability
- Deployment and its challenges in the “live” environment

Prerequisites

- Math: Statistics and Basics of Data Science
- Programming: Python + Pandas (and NumPy + Scikit-learn)
- Development Environment: Google Colab
    - If you are comfortable, you can do your homeworks on your local machine/laptop as well. But again, we cannot help you set it up, Google is a good friend though ;)

NOT EVERYTHING will be explained!
Searching for a solution is actually a part of the course :)

Not EVERYTHING will be Explained...

- Data science is a rapidly changing field
- Searching for a solution is a actually part of the course ;)

APPLIED Data Science

- Main aim of the course is to get you STARTED on using data science in PRACTICE

- We go over all steps of a data problem *workflow*
  - Gathering → Cleaning → Preprocessing → Analysis → Visualization

- The course covers a lot of topics, but we won't go deep (breadth vs. depth)
  - We will learn how to find further details we need online

- We work on several different sample problems to help you be ready for various problems you'll face later in your career

Other Courses and Resources Which Might Help

- Applied Data Science at EPFL
  https://dlab.epfl.ch/teaching/fall2020/cs401/
- Applied Data Science Specialization at Coursera
  https://www.coursera.org/specializations/applied-data-science
- Python for Data Science and Machine Learning Bootcamp at Udemy
  https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp
- AI Python for Beginners (by Andrew Ng)
  https://learn.deeplearning.ai/courses/ai-python-for-beginners

Homework and Final Project

- Homeworks *should* be done individually
- Final projects must be done individually
- You have to submit a pre-compiled Jupyter notebook
- Use of ChatGPT/Bard/etc. are allowed and encouraged!
    - But please first do the assignments yourself, and then seek help from generative AI!

Homeworks are be graded on

- Notebook runs without a problem: 20%
- It solves/addresses the problem: 45%
- It is clear and well-commented: 35%
- **Note:** Homeworks usually come with bonuses. **These bonuses only apply to homeworks** !

Logistics

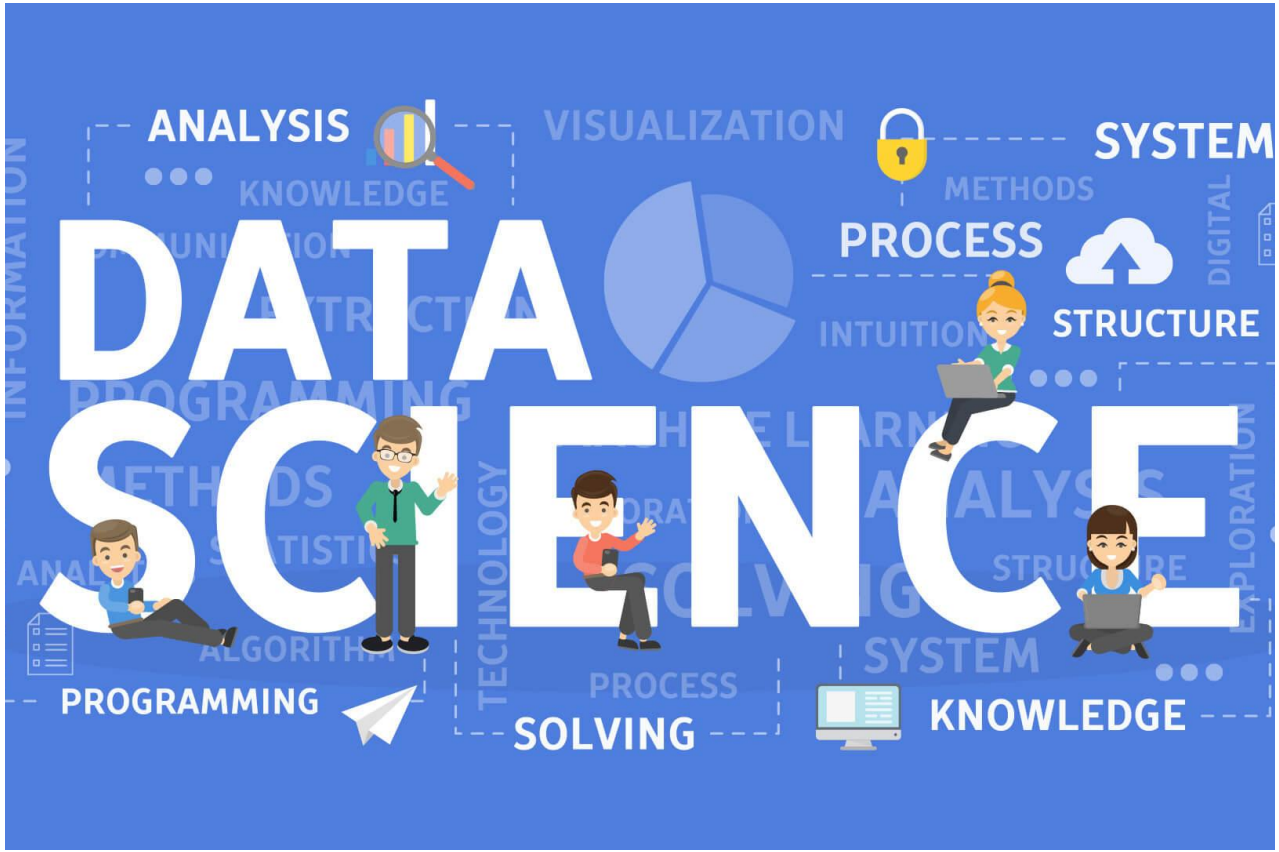
- Lectures and Lab sessions: 80% online, 20% in person
- Course Website: https://saloot.github.io/ADS2025/
- Communications: *MS Teams*
- **TAs: TBA** **Mohmmad Ali Yousefzadeh**

Actively Participate Please!

- On communication channel, please help each other
- During lectures, ask questions if you have any
- At the end of each session, please give us feedback via the link provided.

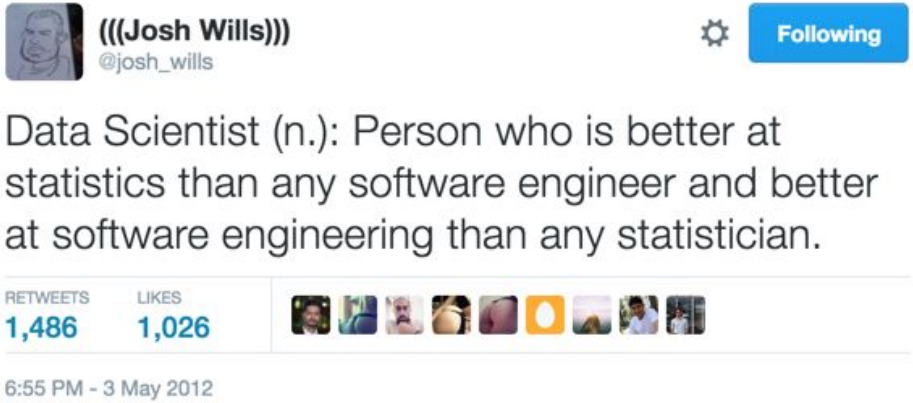
Questions So Far?

Data Science Intro



What is Data Science?

"... while many people think of data science as a profession, it's better to think of data science as a way of thinking, a way to extract insights using the scientific method."
— Bob E. Hayes
https://onalytica.com/blog/posts/big-data-2016-top-100-influencers-and-brands/



"For me, data science is a mix of three things: **quantitative analysis** (for the rigor necessary to understand your data), **programming** (so that you can process your data and act on your insights), and **storytelling** (to help others understand what the data means)."
— Edwin Chen (data scientist at Twitter)
https://www.technologyreview.com/2012/03/22/187074/twitter-data-scientist-takes-on-mcdonalds-entire-menu-survives/

What is Data Science?

From Wikipedia:

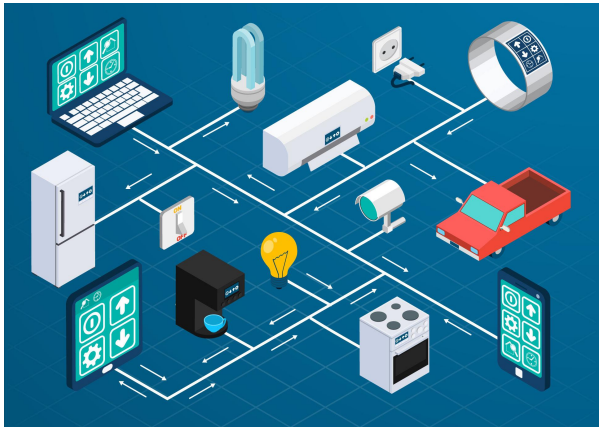
Data science is an **interdisciplinary field** that **uses tools** including scientific methods, processes, algorithms and systems **to extract knowledge and insights** from structured and unstructured **data sources** ...

Why has Data Science Become Important?

Decision-making needs data

Credit: kdnuggets.com

We have LOTs of data



Credit: verdict.co.uk

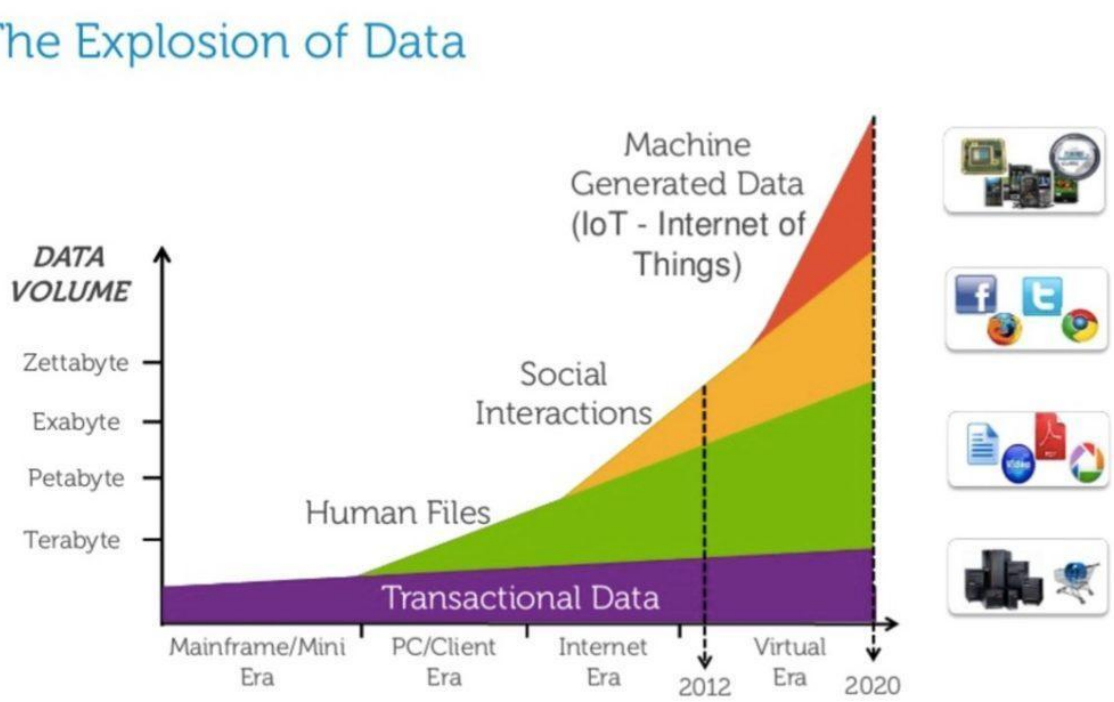
and Computational resources

Credit: verdict.co.uk

Explosion of Data Volume

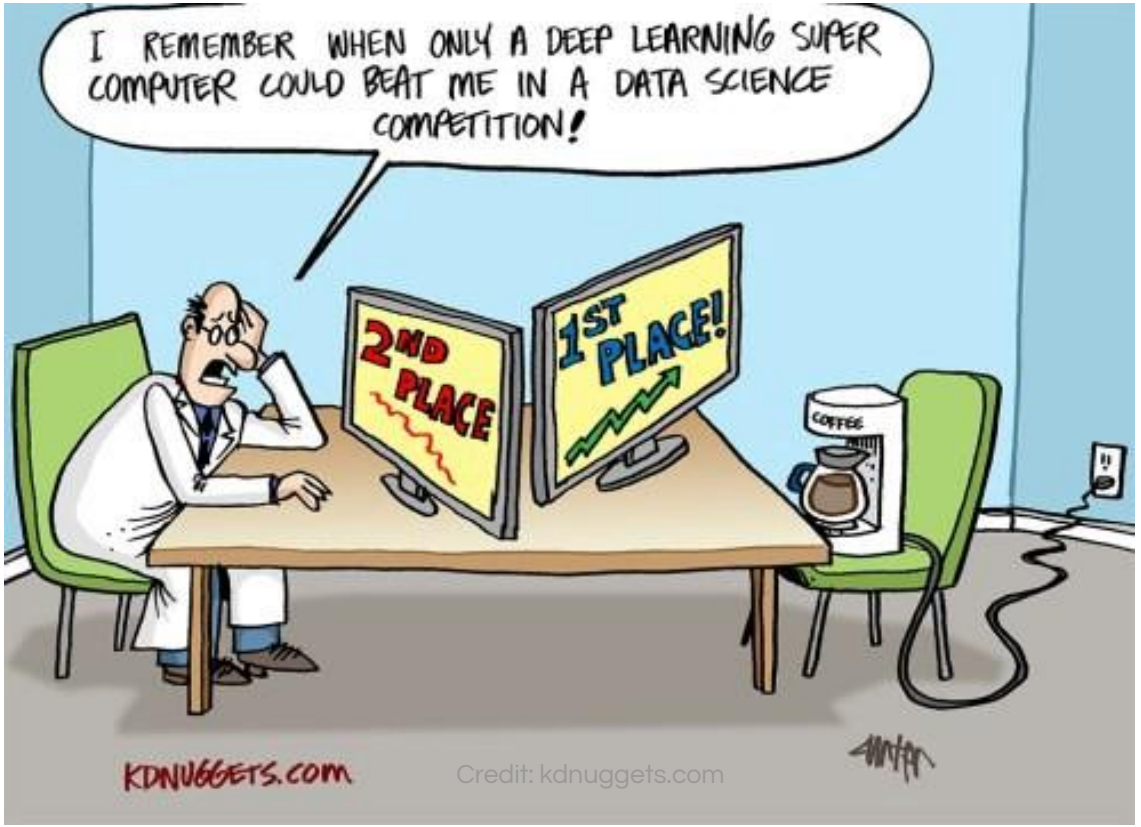
"There was 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing."

Eric Schmidt, Google (2010)



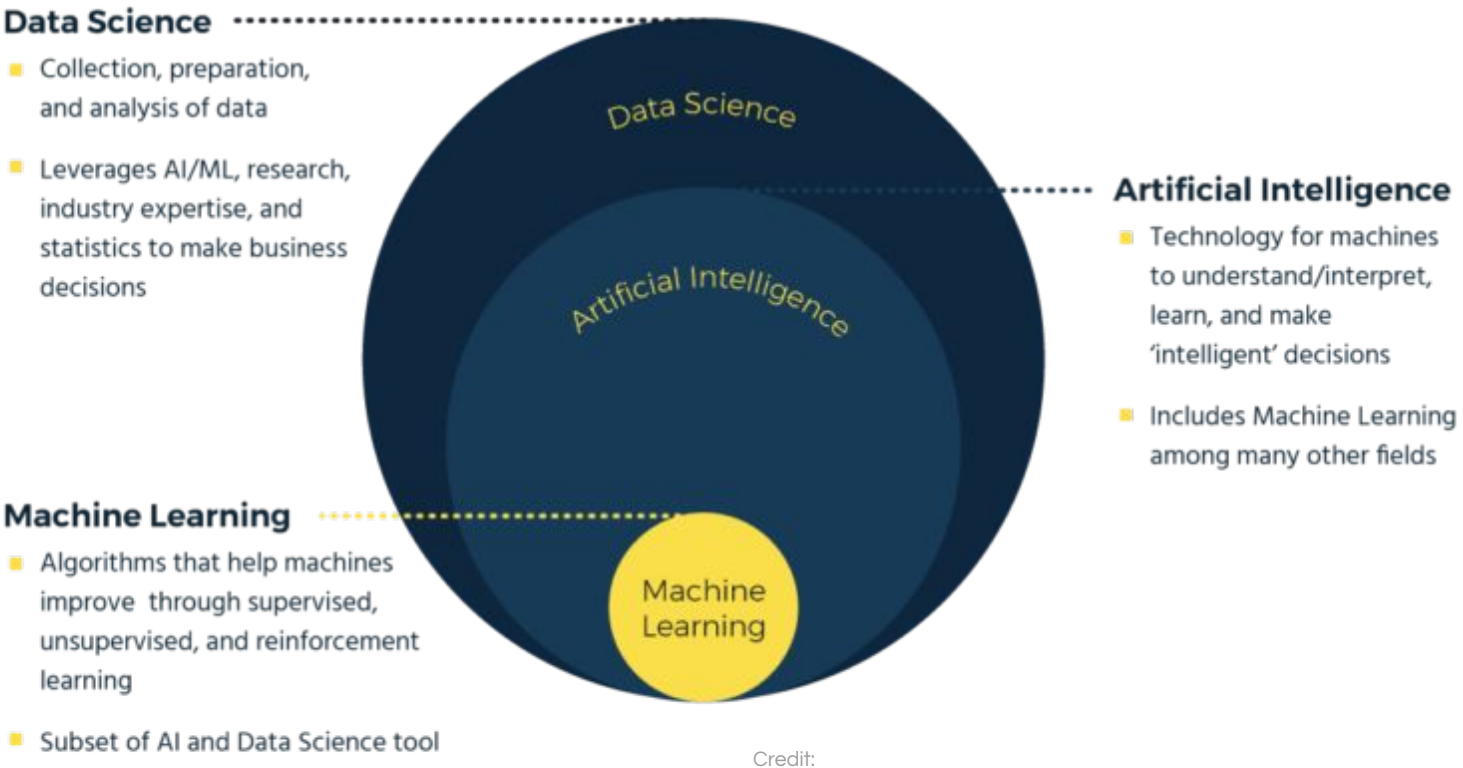
Credit: www.alation.com

Smarter Devices



Credit: kdnuggets.com

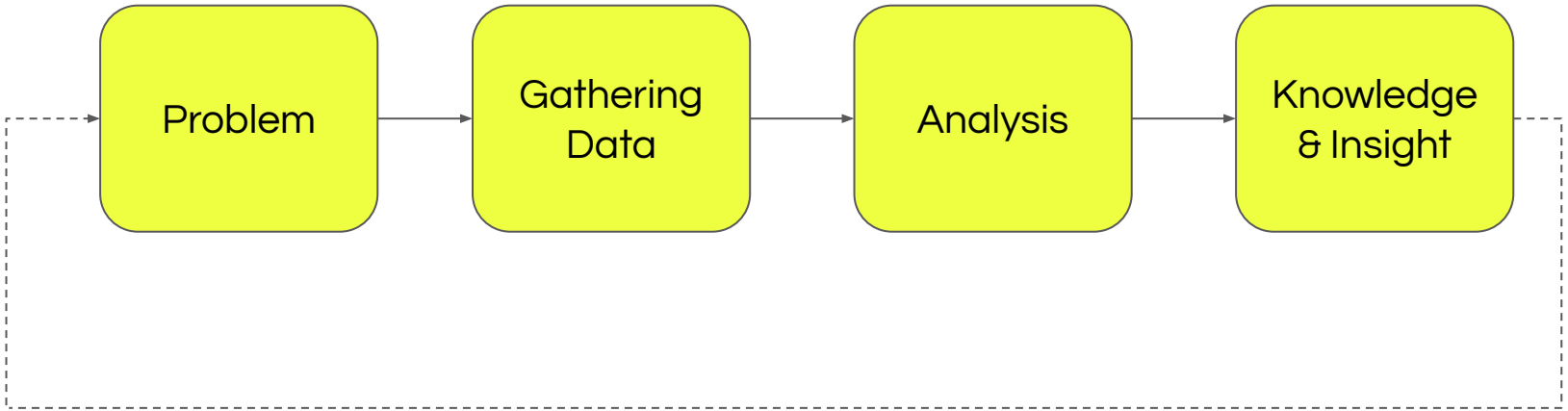
Difference Between AI, Data Science and Machine Learning



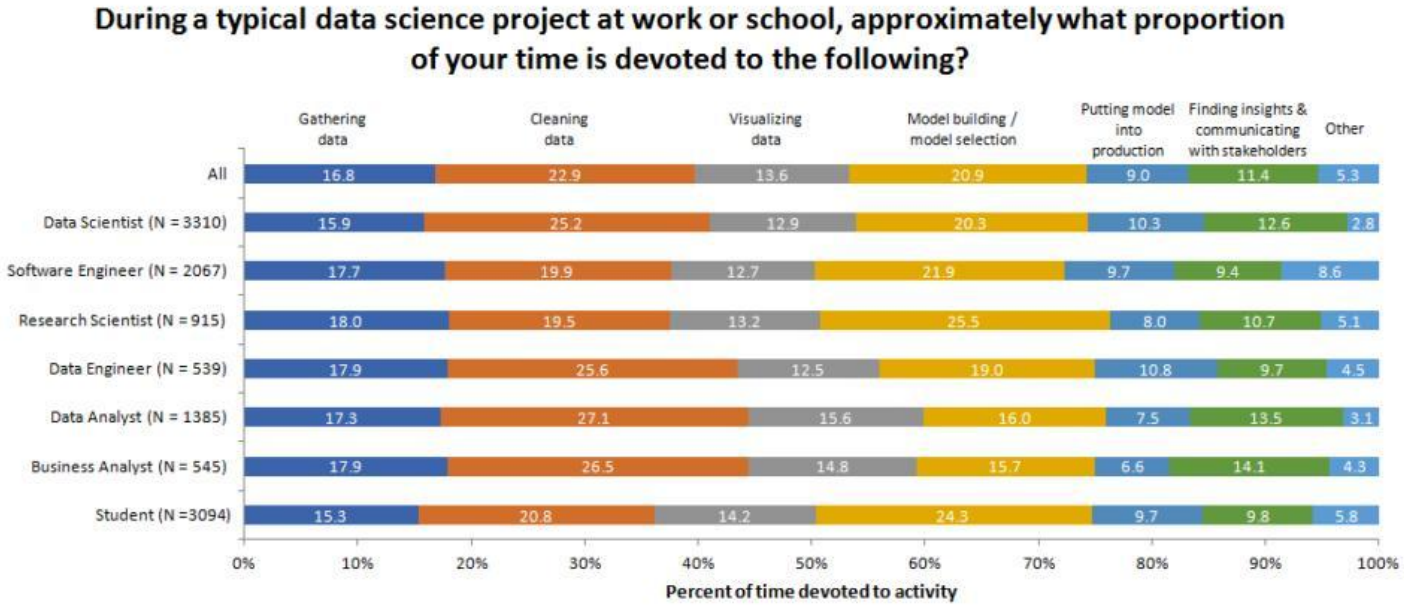
Credit:
https://braincube.com/resource/manufacturing-ai-vs-machine-learning-vs-data-science/

Data Science as an Interdisciplinary Field

Data science is an **interdisciplinary field** that, using structured and unstructured **data sources**, applies **analysis tools** including scientific methods, processes, algorithms and systems **to extract knowledge and insights** ...



Life of a Data Scientist

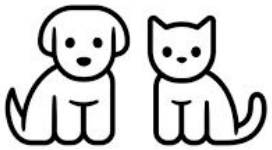


Note: Data are from the 2018 Kaggle ML and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018.
A total of 23859 respondents completed the survey; the percentages in the graph are based on a total of 15937 respondents who provided an answer to this question. Only selected job titles are presented.

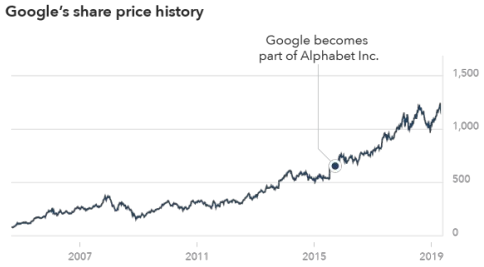
BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

Different Problem Types

Supervised



Classification



Regression

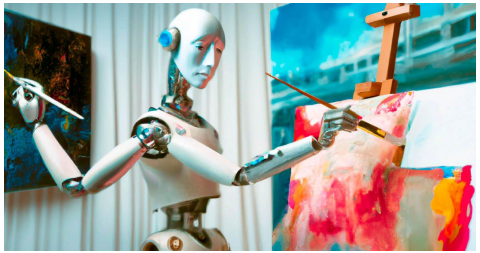
Generative AI

Unsupervised

Clustering

Apple

Association



Example of Data Science Problems

Personalization

Identify users' specific taste and offer them personalized products/services/treatments

Image credit: emotivebrand.com

Example of Data Science Problems

Tumor Detection

Using ML algorithms, identify suspicious tumors from CT scans, MRI results, etc.

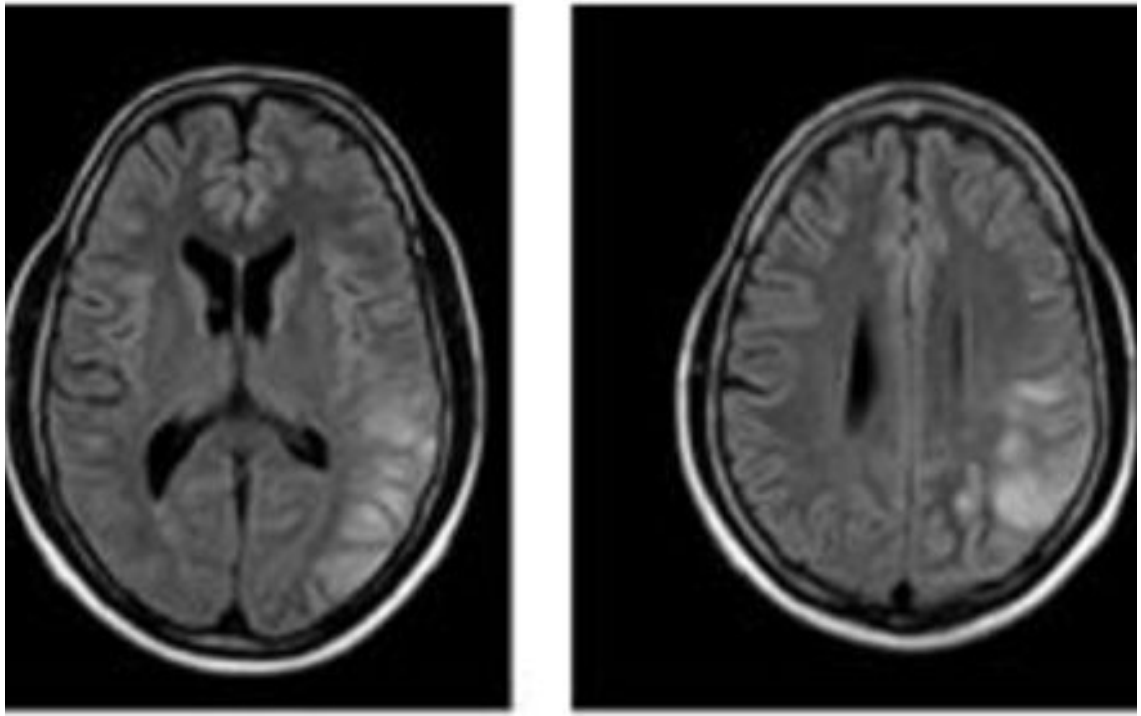
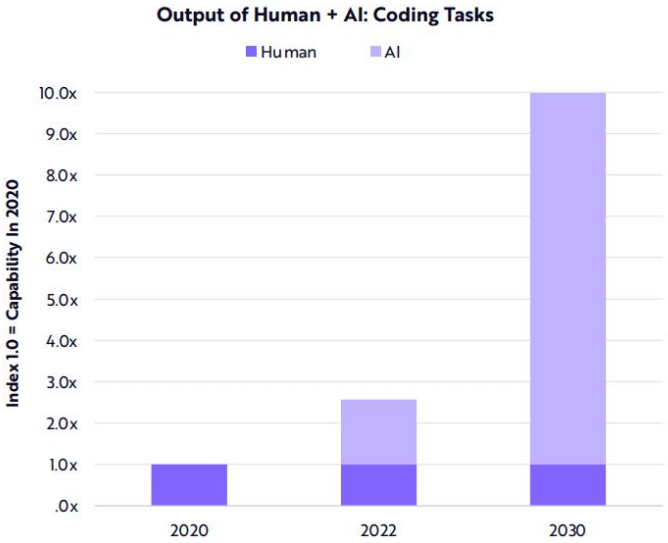


Image credit: www.deepai.com

Example of Data Science Problems

Generative AI

Using generative AI as a co-developer, copywriter, assistant, business consultant, etc.



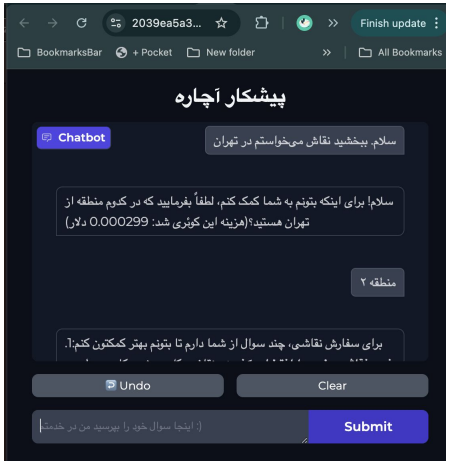


Image credit /thebusinessdive.com/ai-assistant-statistics

Other Examples of Data Science Applications?

ToDo List for Next Session

- Checkout the Google Colab notebook before our lab session:

  https://saloot.github.io/ADS2025/schedule

- **And please don't forget to give us feedback at:**
  https://survey.porsline.ir/s/lavoOaO