Introduction to Machine Learning

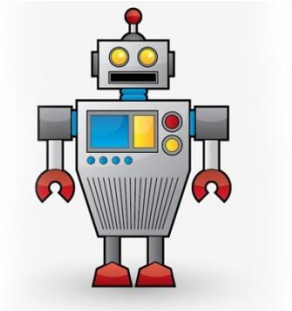
Ruxandra Stoean

rstoean@inf.ucv.ro

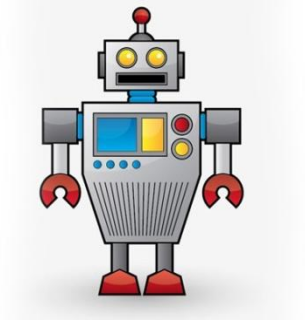
ruxandra.stoean@e-uvt.ro

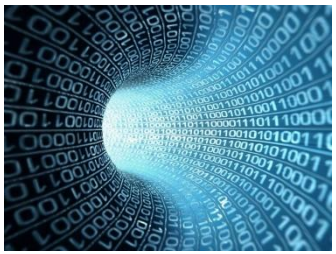
Intelligent machine

- A machine that can be programmed to act like a human



- A machine that can be taught to learn like a human





What is Machine Learning?

Definitions

- "The field of study that gives computers the ability to learn without being explicitly programmed."
  – **Arthur Samuel (1959)**

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."
  – **Tom M. Mitchell (1999)**

- “A subset of artificial intelligence focused on developing algorithms and statistical models that enable computers to learn from data and improve their performance on specific tasks without explicit programming.

- It involves collecting and preparing data, designing mathematical models, training the models on labeled data, evaluating their performance on new data, and deploying them for predictions or decision-making.

- It encompasses supervised learning (learning from labeled data), unsupervised learning (discovering patterns in unlabeled data), and reinforcement learning (learning through interaction with an environment).

- Its applications span a wide range of fields, revolutionizing technology and problem-solving across domains like image recognition, natural language processing, recommendation systems, and autonomous vehicles.”

– **ChatGPT**

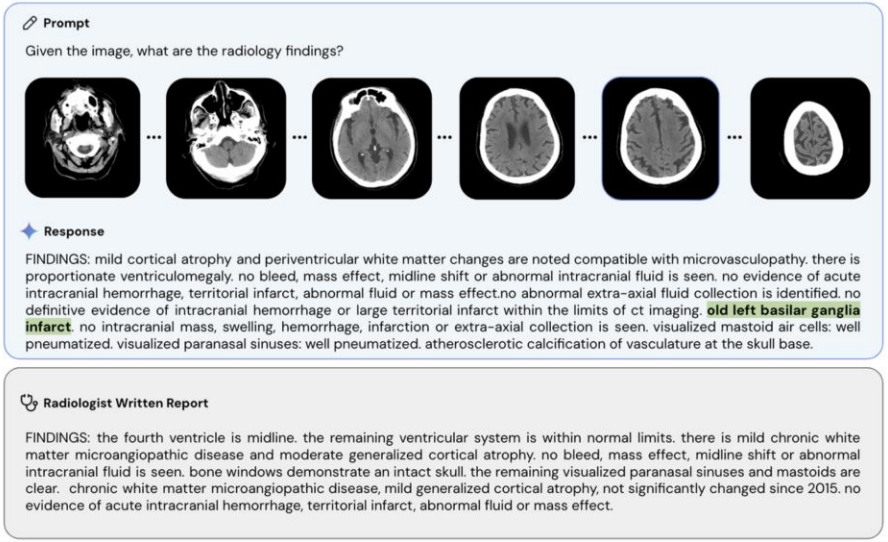
https://stablediffusionweb.com/

Applications. Autonomous vehicles

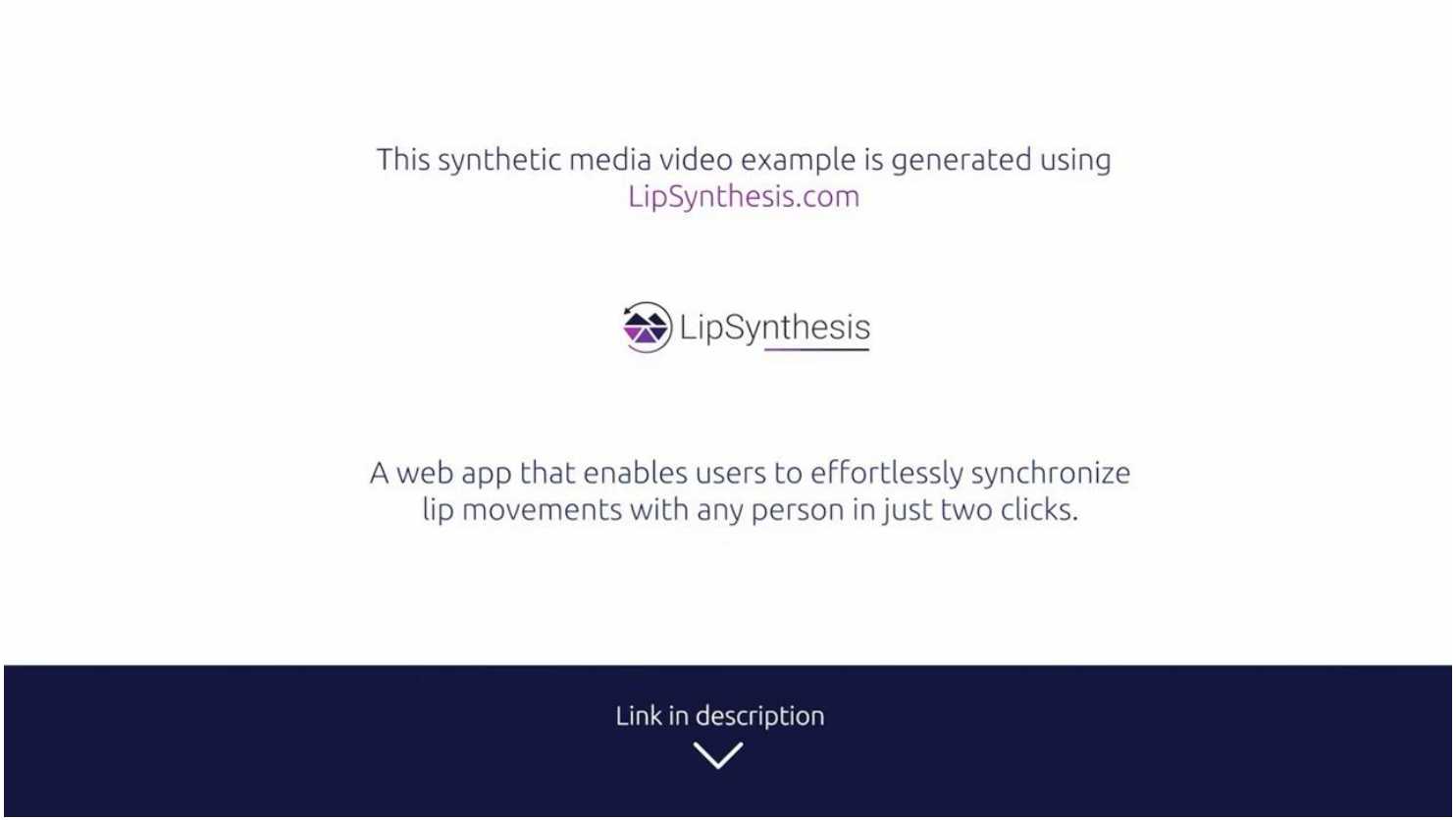
- Waymo from Google
  - https://www.youtube.com/watch?v=0kJPDg207oc
- Microsoft
  - https://www.youtube.com/watch?v=zkfrkgtrLdo

Medical diagnosis support

- MedGemma - Hugging Face
  - Image classification, text analysis, report generation, patient triage
  - https://huggingface.co/spaces/google/appoint-ready
- Microsoft AI Diagnostic Orchestrator (MAI-DxO)
  - 304 initial real cases from the New England Journal of Medicine
  - Foundation models: GPT, Llama, Claude, Gemini, Grok, DeepSeek.
  - https://youtu.be/JkIjmXEK0Yg?si=XDXPrtj2XamZQyLv
- Med-Gemini - Google
  - Imaging, genomic data



Deep fake



https://www.youtube.com/watch?v=HjwJklIuYFo


• FAKE

OpenAI

- ChatGPT - GPT5
- OpenAI Jukebox (2020)
  - Alternative, Suno AI text-to-speech audio model (2023)
    - https://suno.com/
- Dall-E: text-to-image
  - https://openai.com/index/dall-e-3/?video=866470243
  - Included in Microsoft Copilot
  - Replaced by GPT Image 1 within ChatGPT

Still OpenAI

- Movies: OpenAI's image-making model DALL-E 2 "The Frost"
  - https://www.youtube.com/watch?v=IgPvoPBrlTE
- Sora, text-to-video model
  - https://www.youtube.com/watch?v=HK6y8DAPN_0
- OpenAI voice chat with ChatGPT
  - Alternative, Gemini from Google

Google Deep Mind

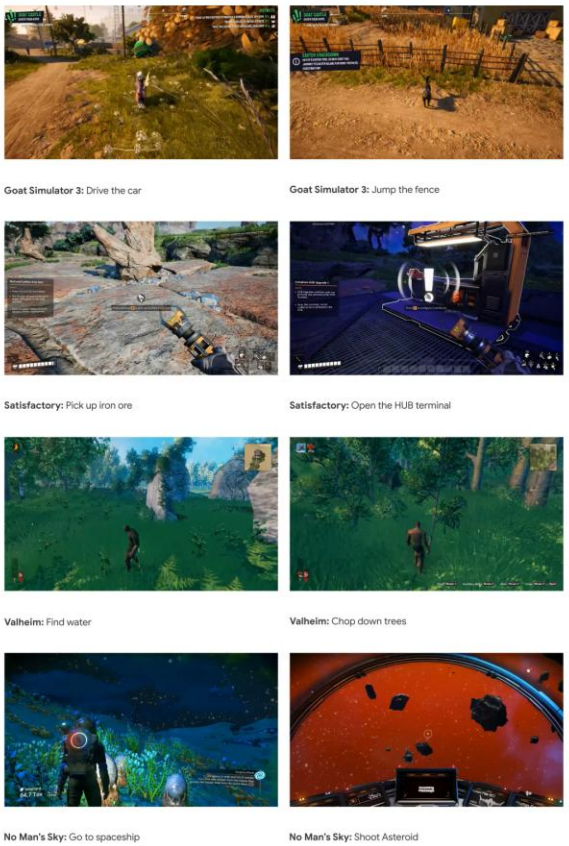
- Veo
  - Competitor to Sora
- Veo2 – "The Bridge"
  - https://youtu.be/YDlME4qvER8?si=lVwfUA4MSrI6Bsr7
- Flow – filmmaking tool with
  - Veo 3 for generative video
  - Gemini for prompts
  - Imagen for asset creation
  - https://youtu.be/A0VttaLy4sU?si=LVHobw2AJ7NVwlkt

Detection of AI generated content

- SynthID, Google
  - https://youtu.be/9btDaOcfIMY?si=C0DK2Jrv-gIRDmP0
- DeepFake Detector, Intel & McAfee
  - https://www.intel.com/content/www/us/en/content-details/832352/mcafee-demo-deepfake-detector-empowering-users-to-spot-ai-scams-with-intel.html

AI in games

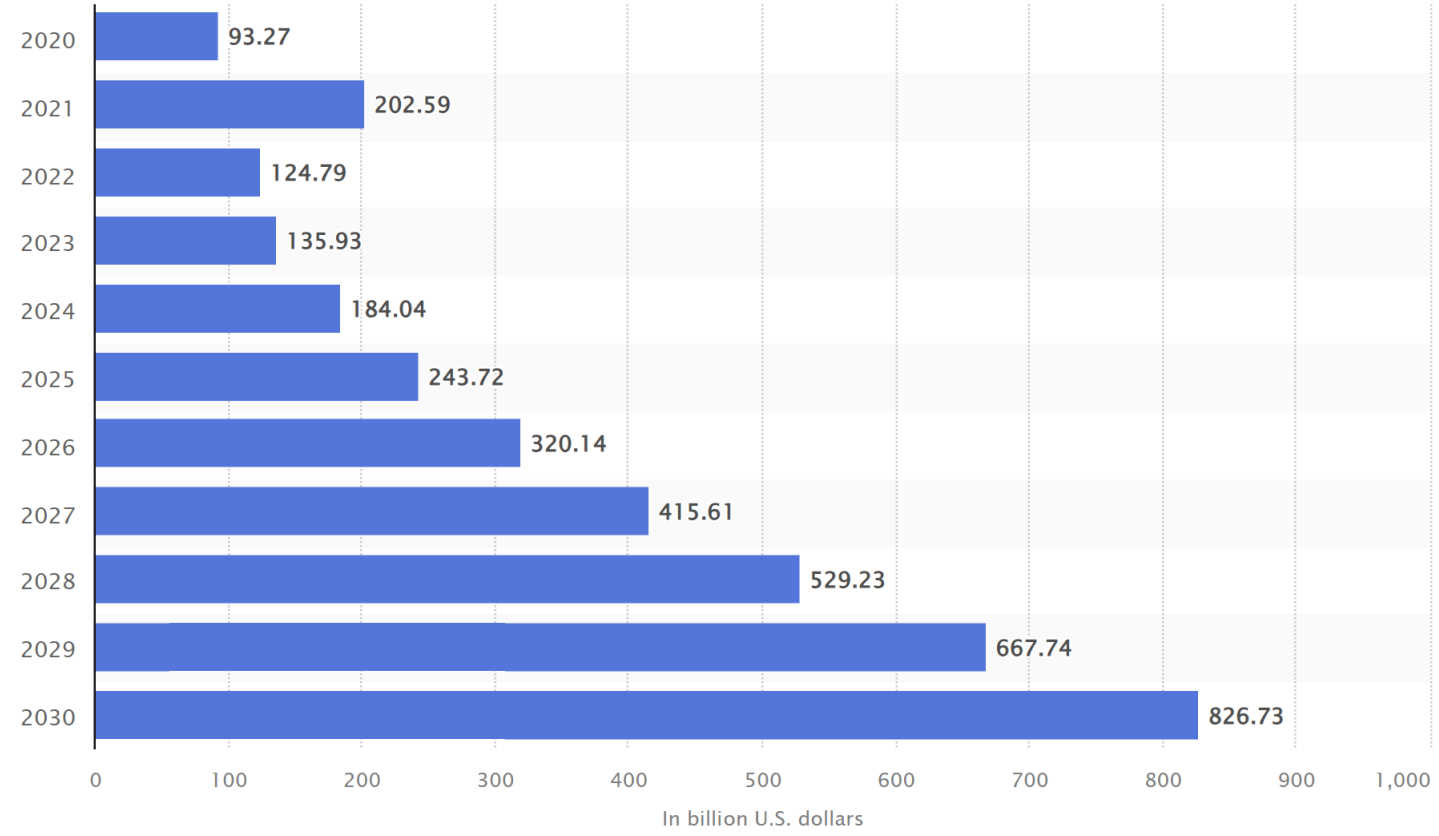
- Sima, Google DeepMind
  - Trained on many video games from several game studios for visual information
  - Follows natural language instructions
  - Zero-shot generalization



What is your favorite AI breakthrough?

Artificial intelligence (AI) market size worldwide from 2020 to 2030

(in billion U.S. dollars)

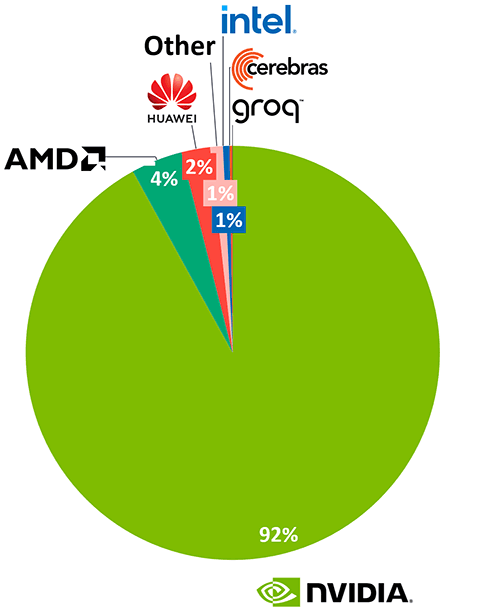


IOT ANALYTICS

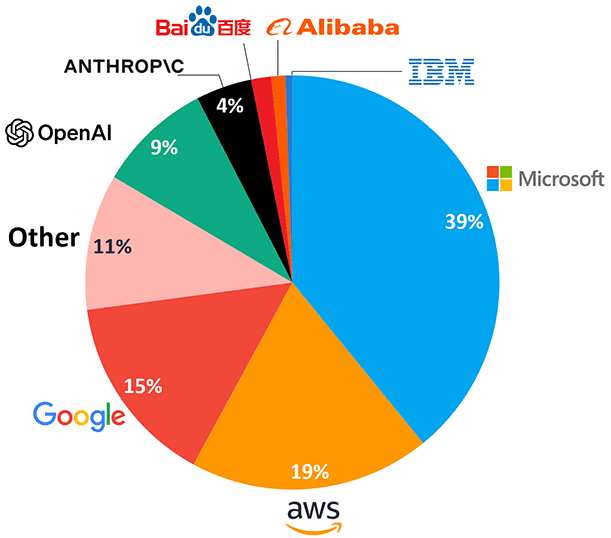
March 2025

Generative AI: Market share of leading vendors

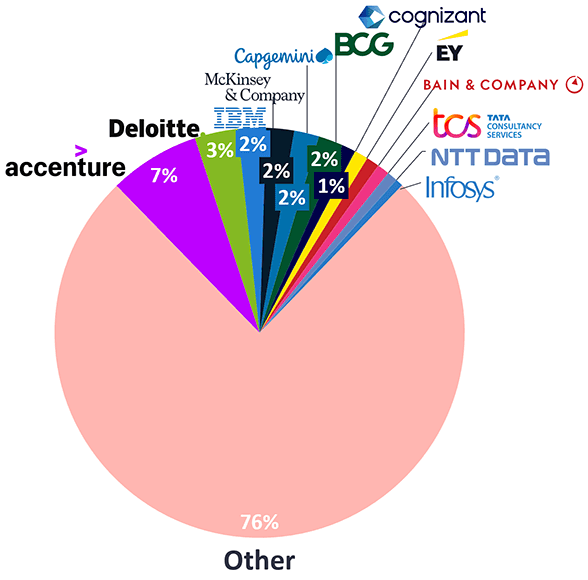
1 Datacenter GPUs



2 Foundation models & model managment platforms



3 GenAI services



Note: Numbers are rounded and might not add up to 100%; Market is estimated based on first three quarters of 2024.

AI Trends 2025

- Ethics and regulations for responsible AI
- Post-truth
- AI for cybersecurity
  - Breaches, anomaly identification, system automation
- Sustainable AI
  - Less consumption in data centers, more concentrated on renewables
  - Support for environmental protection

Recommended reading - theory

- C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 2006
- T. Hastie,R. Tibshirani,J. Friedman, *The Elements of Statistical Learning*, Springer, 2009
- M. P. Deisenroth, *Mathematics for Machine Learning, Cambridge*, 2020.
- C. Huyen, *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*, O'Reilly, 2022
- I. Goodfellow,Y. Bengio, A. Courville, *Deep learning*, MIT, 2016
- J. Krohn, G. Beyleveld, A. Bassens, *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence* (Addison-Wesley Data & Analytics Series), 2019

Recommended reading - lab

- A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* 3rd Edition, O'Reilly, 2022
- S.Raschka, *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*, Packt, 2022
- B. Lantz, *Machine Learning with R: Learn techniques for building and improving machine learning models, from data preparation to model tuning, evaluation, and working with big data*, 4th Edition 4th Edition, Packt, 2023
- F. Chollet, *Deep Learning with Python*, Second Edition 2nd Edition, Manning, 2021

About this lecture

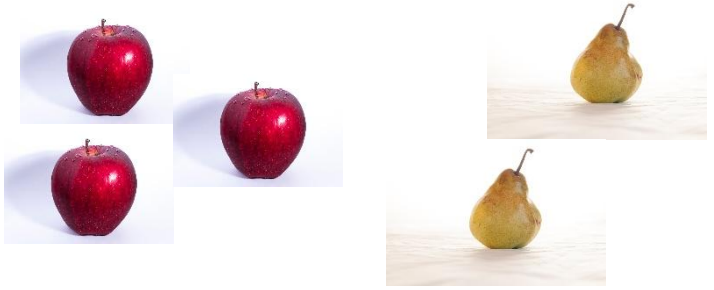
- Lecture: theoretical concepts, models, explanations, discussions
- Lab: model implementations in R & Python
- Final mark:
  - Written exam (30%)+
    - Lab activity (30%)
    - Project (with PPT presentation) (40%)
    - Uploaded to Classroom and presented to class

Machine Learning types

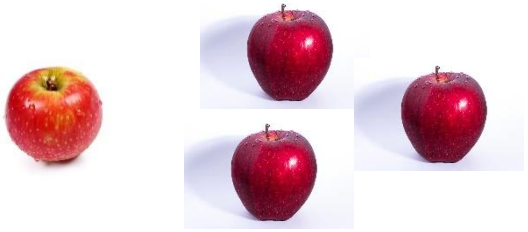
- Supervised learning:
  - Input data with labelled output
  - Learns the association between input and output
  - Prediction on the output of new data records
- Unsupervised learning:
  - Input data without labels
  - Learns/discovers the structure, pattern of the data
- Reinforcement learning
  - Learns through response of interaction with the environment

Supervised, unsupervised, RL

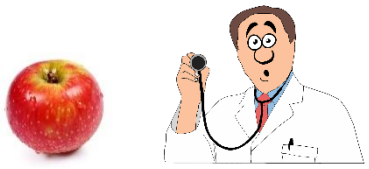
- Supervised learning
  - These are apples and these pears.



- Unsupervised learning
  - This apple is like the other apples.

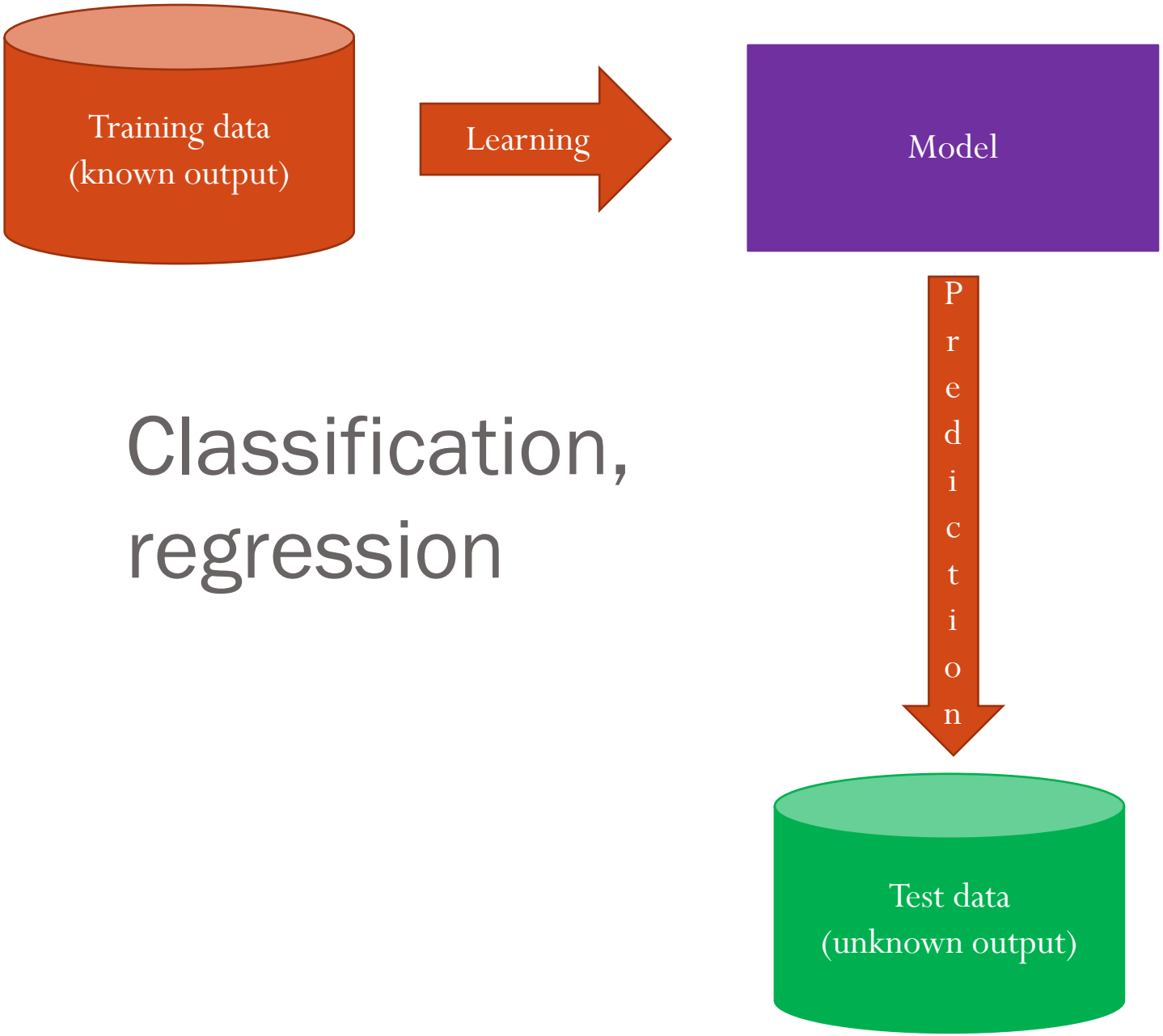


- Reinforcement learning
  - Eat this apple to keep the doctor away!

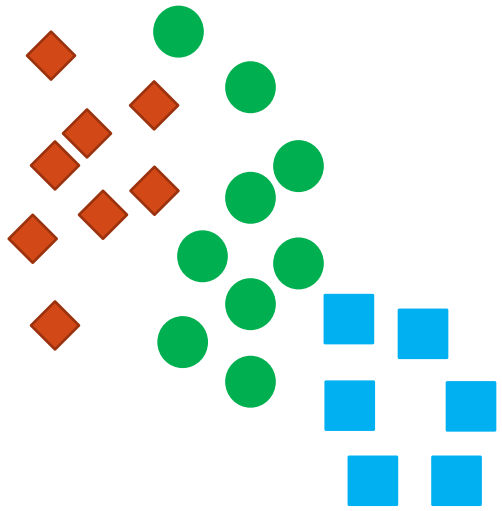


Machine Learning models

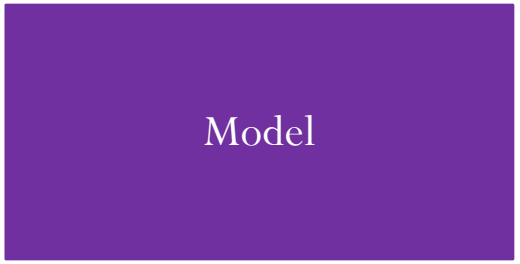
- Supervised learning:
  - Classification: data output made of several qualitative classes
    - Predict the class of new data
  - Regression: data output is quantitative
    - Predict the output value of new data
- Unsupervised learning:
  - Clustering: group data
  - Dimensionality reduction: data transformation- high to low-dimension
- Reinforcement learning
  - Trial and error to choose the best action, get reward
- Collateral:
  - Model evaluation, selection, parametrization

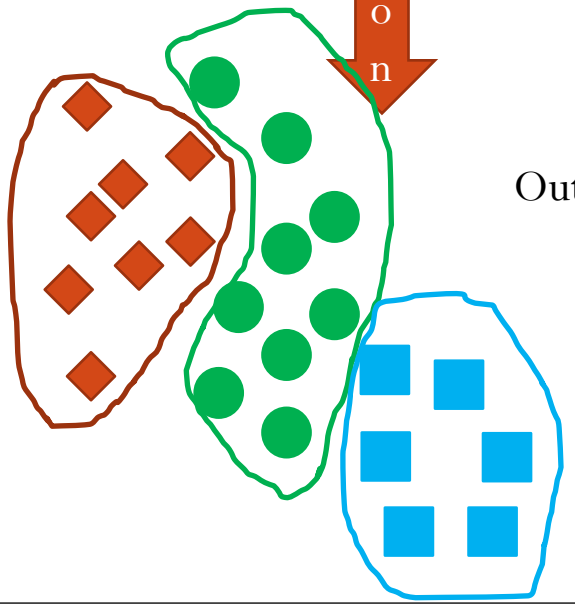


Classification,
regression



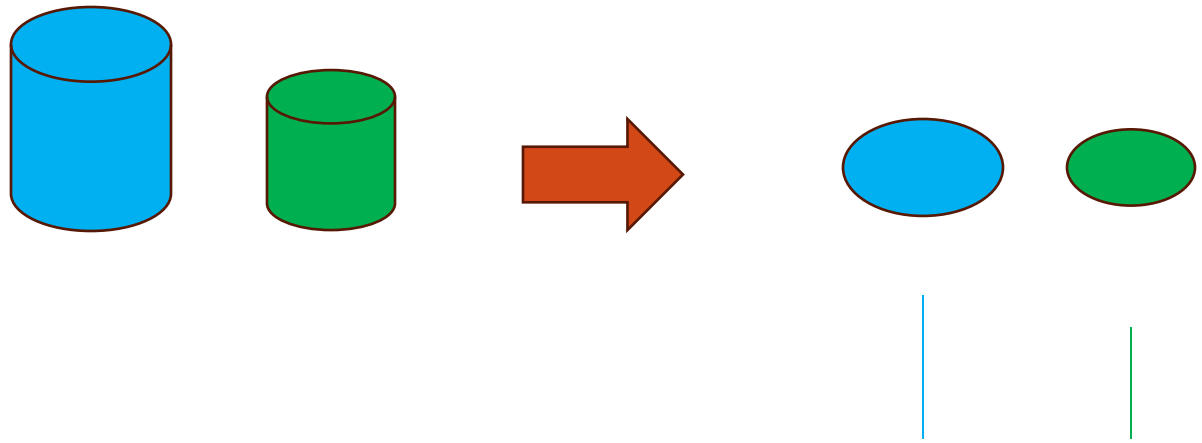
Input data





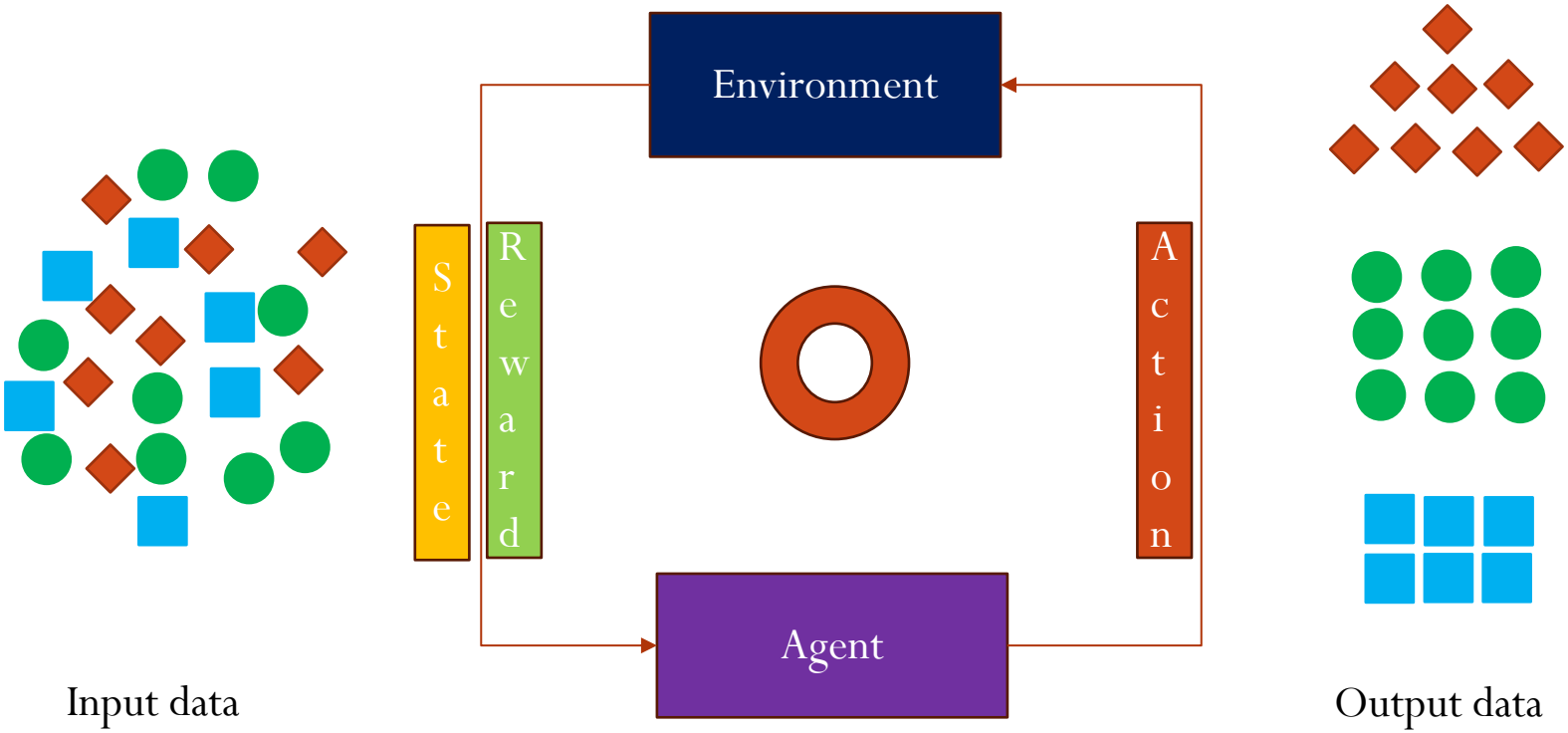
Output data

Clustering

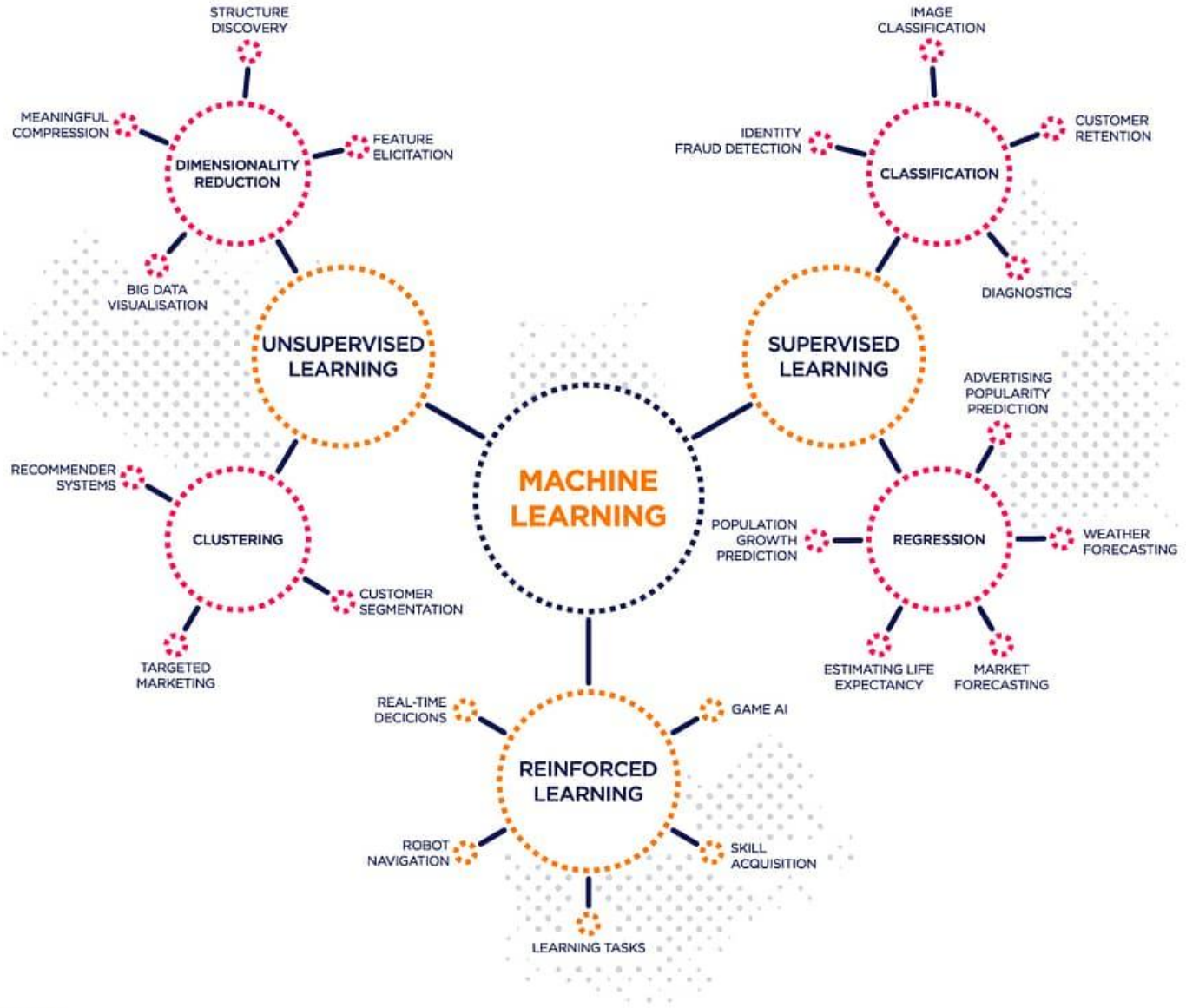


Dimensionality reduction

Reinforcement learning



Machine Learning applications



ITERATORS

Source: www.iteratorshq.com

Ex. 1 – Classification Iris

Fisher's *Iris* Data

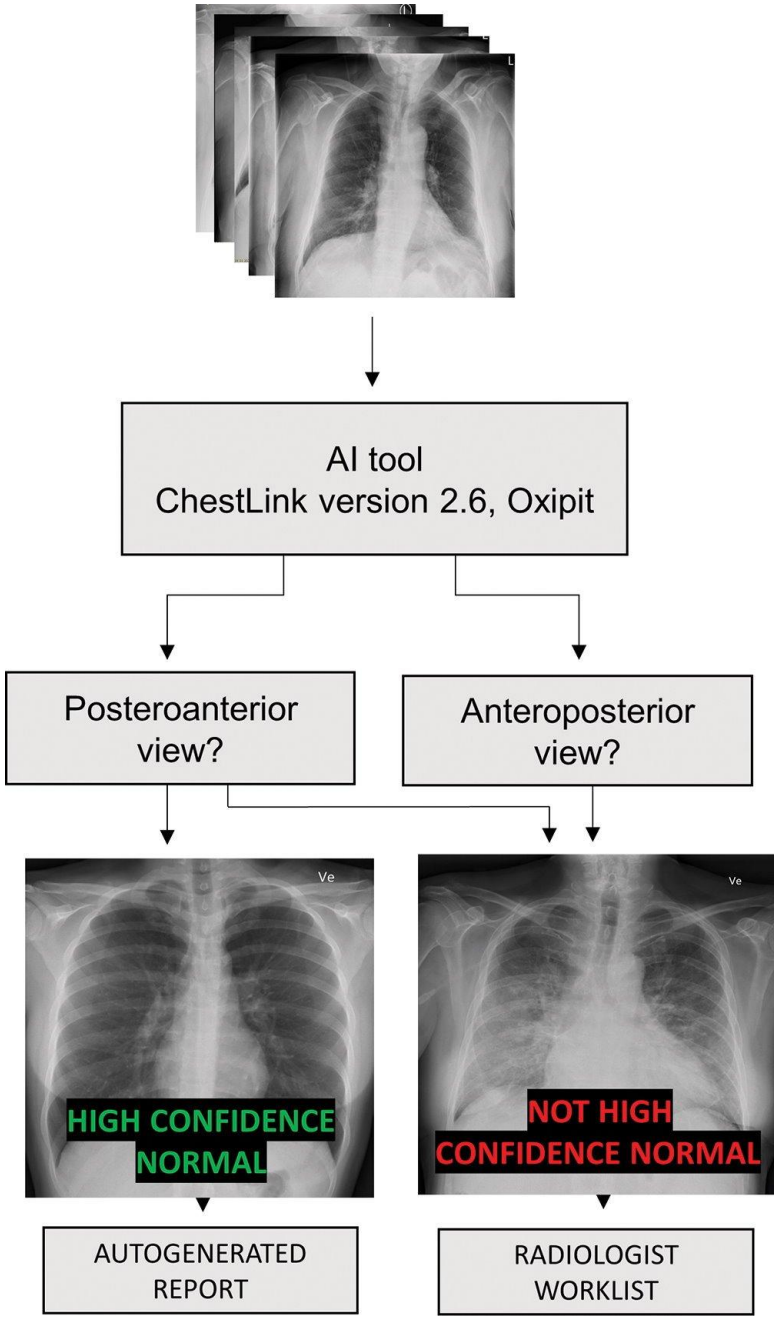
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.1	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
5.4	3.7	1.5	0.2	*I. setosa*
4.8	3.4	1.6	0.2	*I. setosa*
4.8	3.0	1.4	0.1	*I. setosa*
4.3	3.0	1.1	0.1	*I. setosa*
5.8	4.0	1.2	0.2	*I. setosa*
7.0	3.2	4.7	1.4	*I. versicolor*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.5	2.8	4.6	1.5	*I. versicolor*
5.7	2.8	4.5	1.3	*I. versicolor*
6.3	3.3	4.7	1.6	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.6	2.9	4.6	1.3	*I. versicolor*
5.2	2.7	3.9	1.4	*I. versicolor*
5.0	2.0	3.5	1.0	*I. versicolor*
5.9	3.0	4.2	1.5	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.1	2.9	4.7	1.4	*I. versicolor*
5.6	2.9	3.6	1.3	*I. versicolor*
6.7	3.1	4.4	1.4	*I. versicolor*
5.6	3.0	4.1	1.3	*I. versicolor*
5.5	2.5	4.0	1.3	*I. versicolor*
5.5	2.6	4.4	1.2	*I. versicolor*
6.1	3.0	4.6	1.4	*I. versicolor*
5.8	2.6	4.0	1.2	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
5.6	2.7	4.2	1.3	*I. versicolor*
5.7	3.0	4.2	1.2	*I. versicolor*
5.7	2.9	4.2	1.3	*I. versicolor*
6.2	2.9	4.3	1.3	*I. versicolor*
5.1	2.5	3.0	1.1	*I. versicolor*
5.7	2.8	4.1	1.3	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
5.8	2.7	5.1	1.9	*I. virginica*

Iris setosa

Iris versicolor

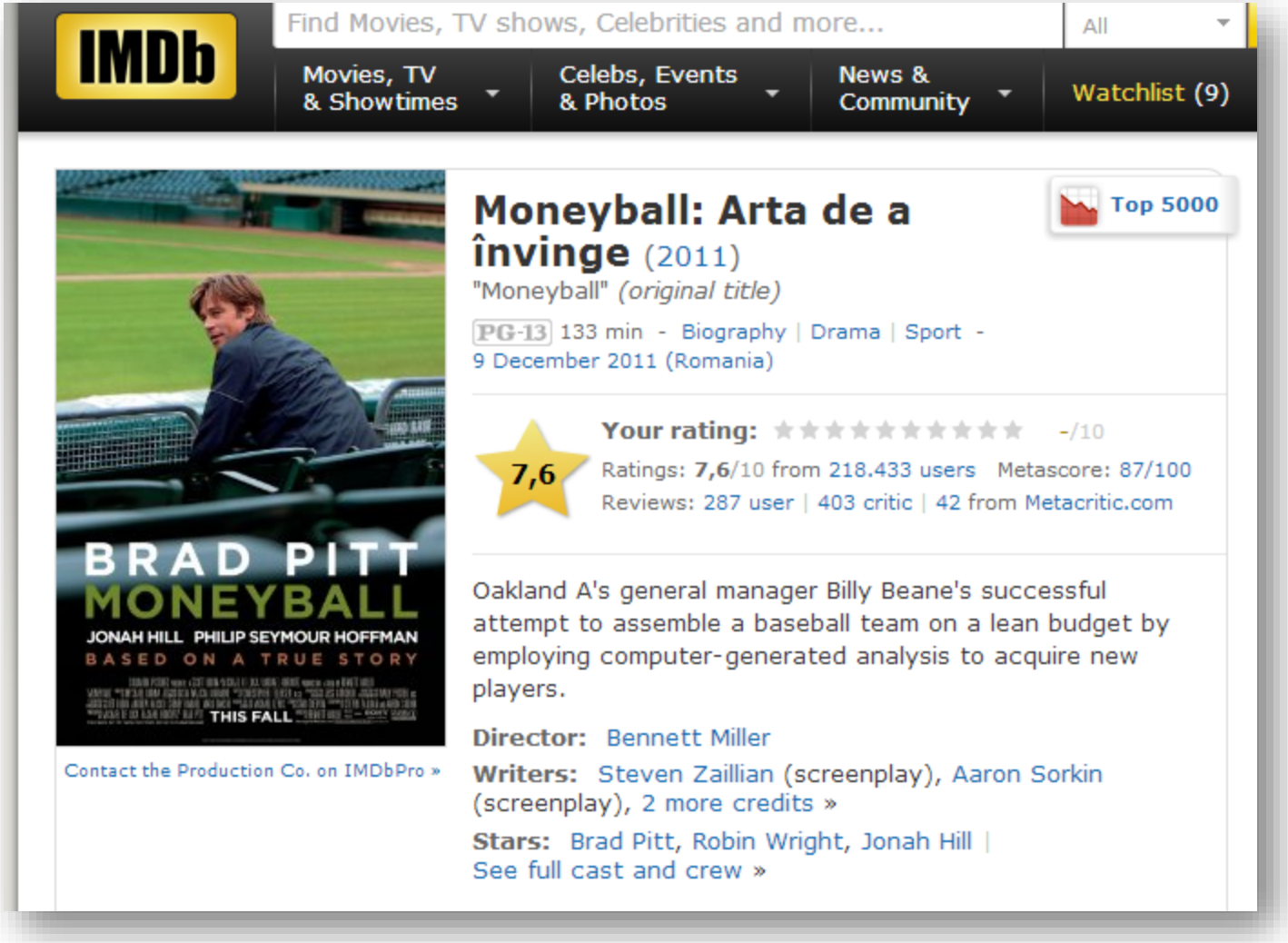
Iris virginica

Ex. 2 Medical diagnosis support



healthcare-in-europe.com
pubs.rsna.org

Ex. 3 – Regression in sport/movies

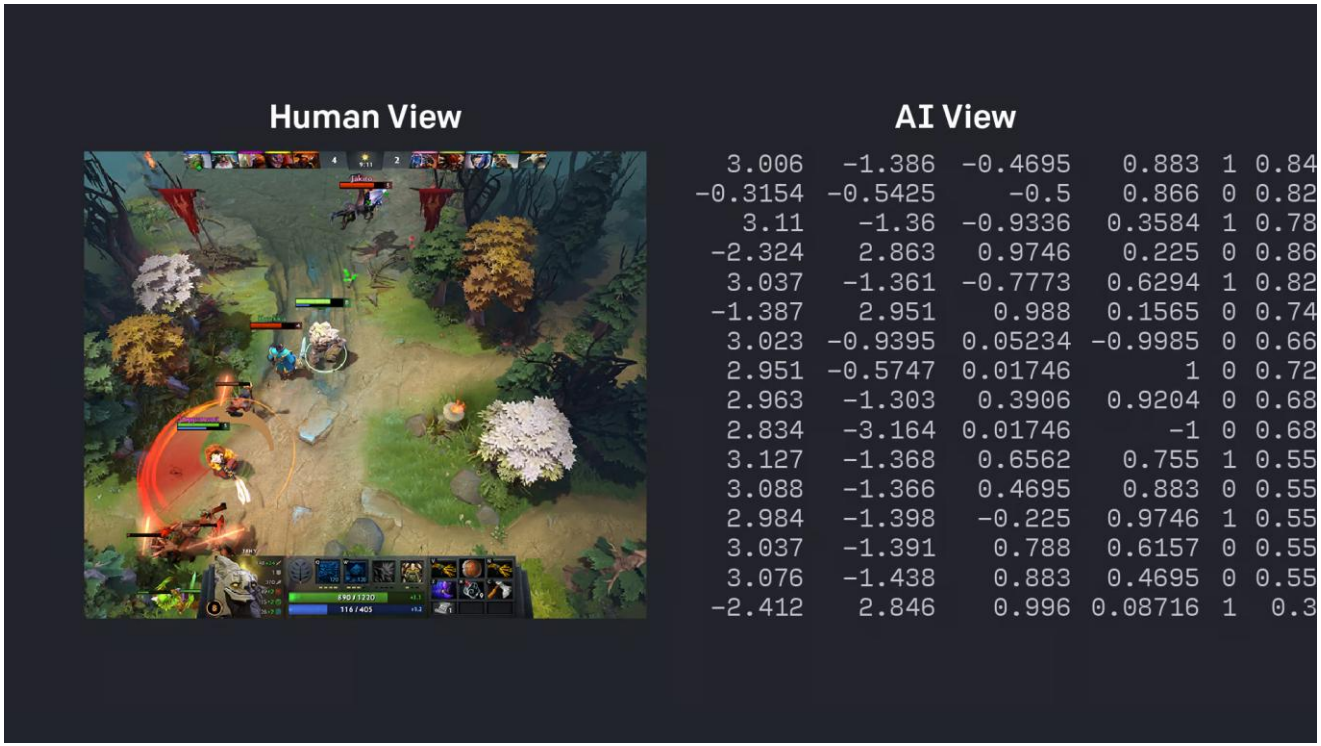


Ex. 4 – Recommendation Clusters

recoai.net

Ex. 5 RL in gaming

- OpenAI Five used deep reinforcement learning to beat the world champions in Dota 2.
- https://www.twitch.tv/videos/410533063?t=1h14m30s



openai.com

Three main types of Machine Learning Algorithms

Reddit

Lecture contents

- Linear models (LM)
- Support vector machines (SVM)
- Neural networks (NN)
- Decision trees (DT) and ensemble methods
  - Bagging
  - Boosting
  - Random Forests (RF)

Contents (cont.)

- Deep learning
  - Convolutional neural networks (CNN)
  - Semantic segmentation (U-Net)
  - Recurrent neural networks (LSTM, GRU)
  - Transformers
  - Generative models
  - Explainability
- Deep reinforcement learning (RL)

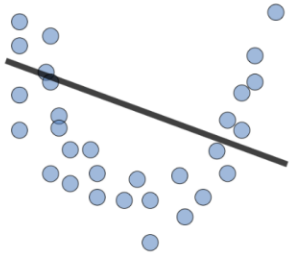
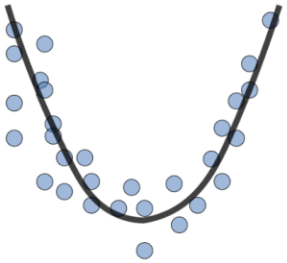
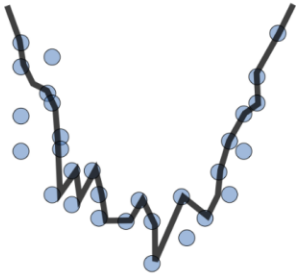
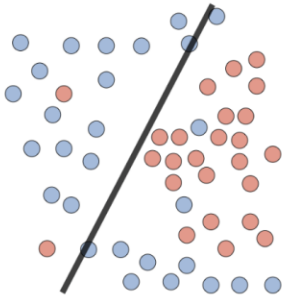
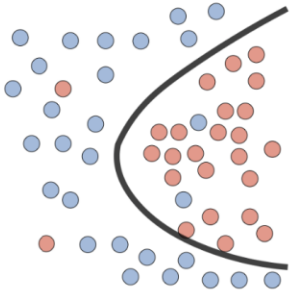
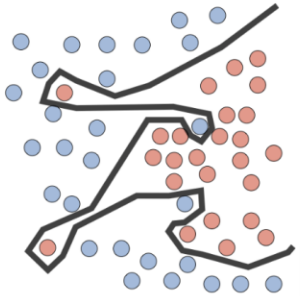
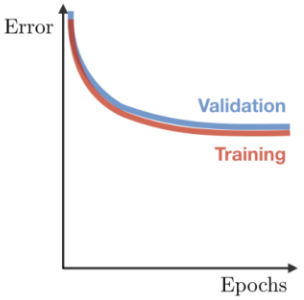
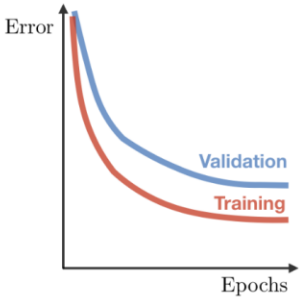
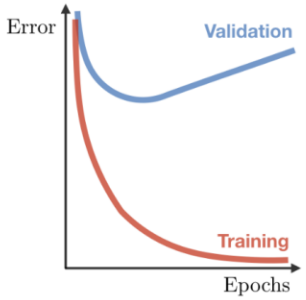
Further points

- Applications to real-world tasks
- Selection and evaluation of constructed models:
  - Model agreement with data / Model diagnosis
  - Performance measures
  - Accuracy/error estimation
  - Comparison between models – Testing the statistical significance
  - Benchmarking

Model selection

- From several possible models, choose the best for the problem under consideration.
- The problem is represented by the data

- Special attention to the trade-off between bias and variance
  - Generalization – simple model – underfitting (high bias)
    - Bias – the difference between the model prediction and the real value over the training data
  - Specialization – complex model – overfitting (high variance)
    - Variance – the inconsistency of the model prediction over different training data sets

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

https://stanford.edu/~shervine/teaching/cs-229

Performance improvement

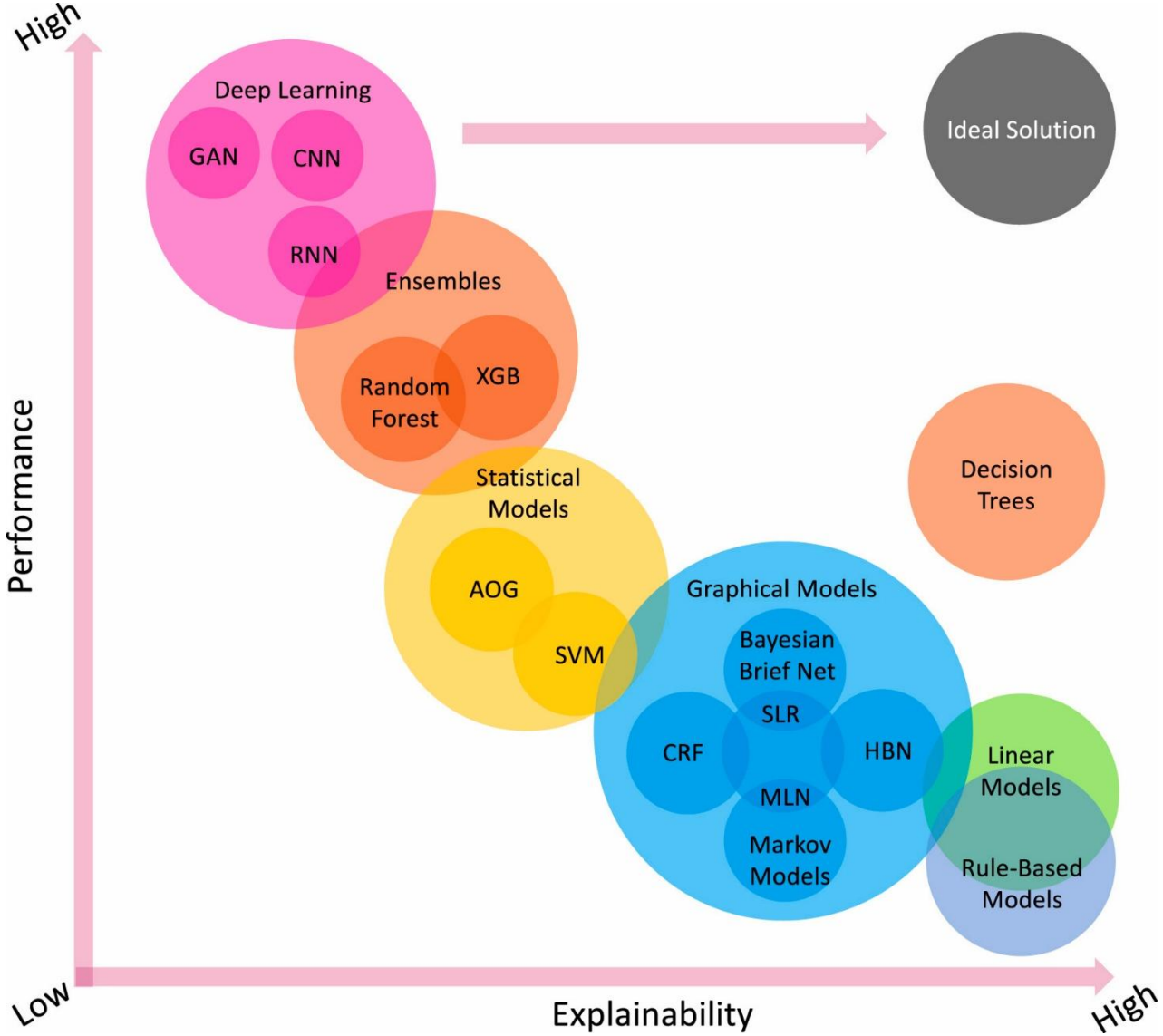
- Parameter tuning
  - Manual
  - Automatic
- Feature inspection
  - Feature selection
    - Reduction of the attributes to the most important
  - Feature extraction
    - Combination of the original attributes to a smaller set of new features

Interpretability / Explainability

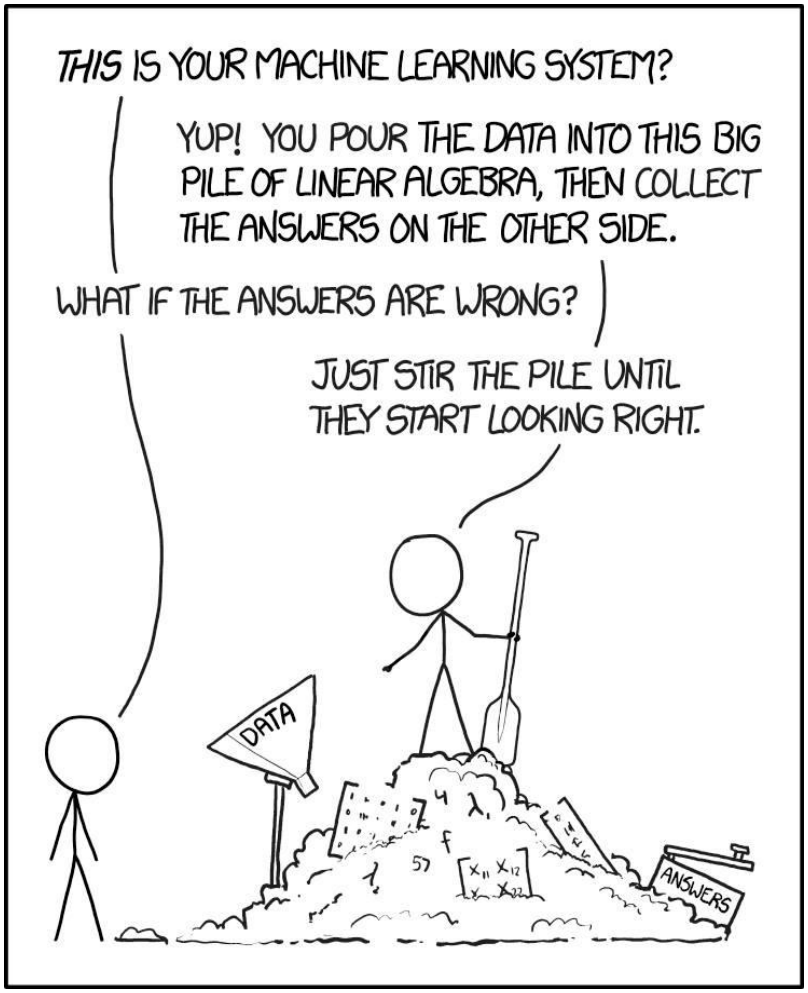
- Interpretable model
  - Internal mechanisms are understood
  - How does the model work to produce its outputs?

- Explainable model
  - The reasons behind a specific prediction of a model
  - Why did it produce this output? Justification needed

XAI

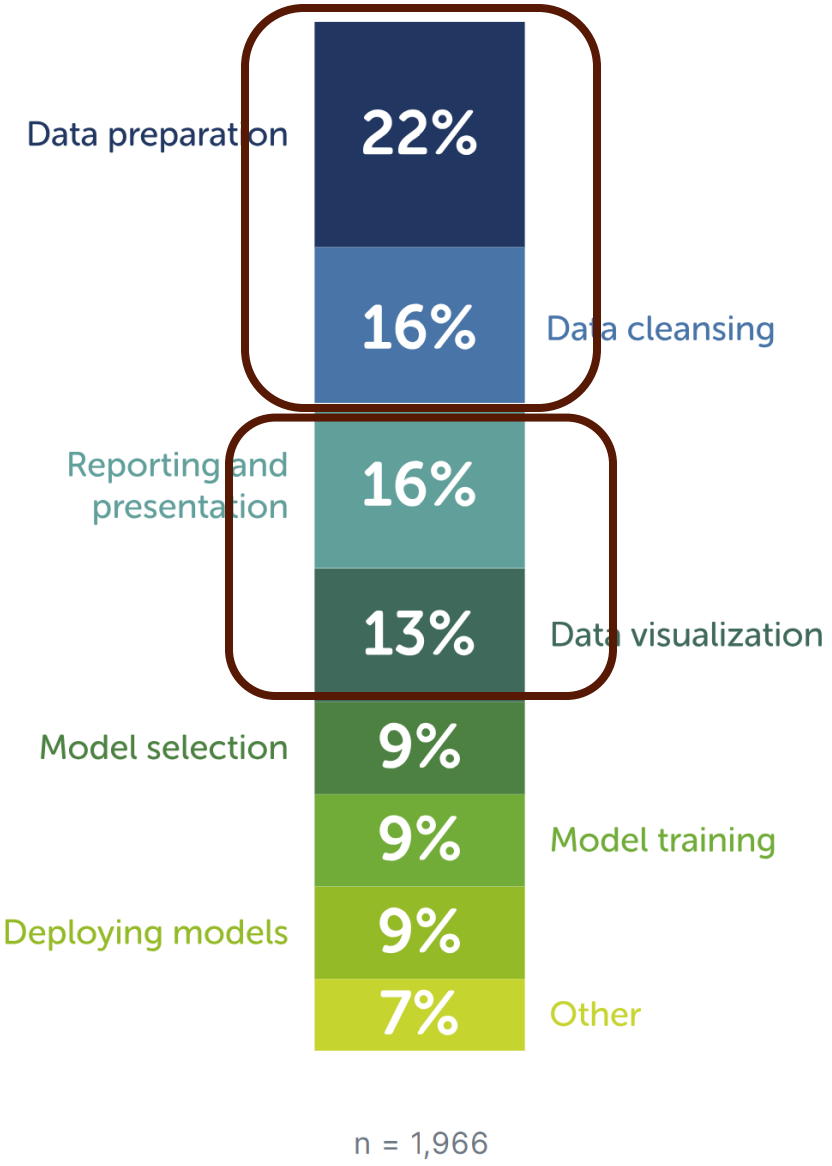


https://doi.org/10.1016/j.inffus.2021.07.016



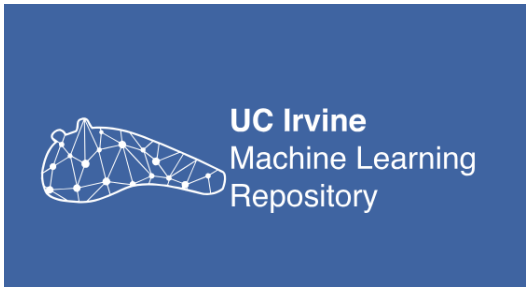
Machine Learning Memes for Convolutional Teens

Task importance by time



Resources

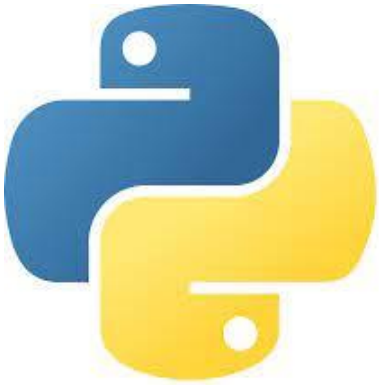
- https://archive.ics.uci.edu/
  - Buttons for download and for import to Python
- https://www.kaggle.com/
  - Data sets
  - Codes



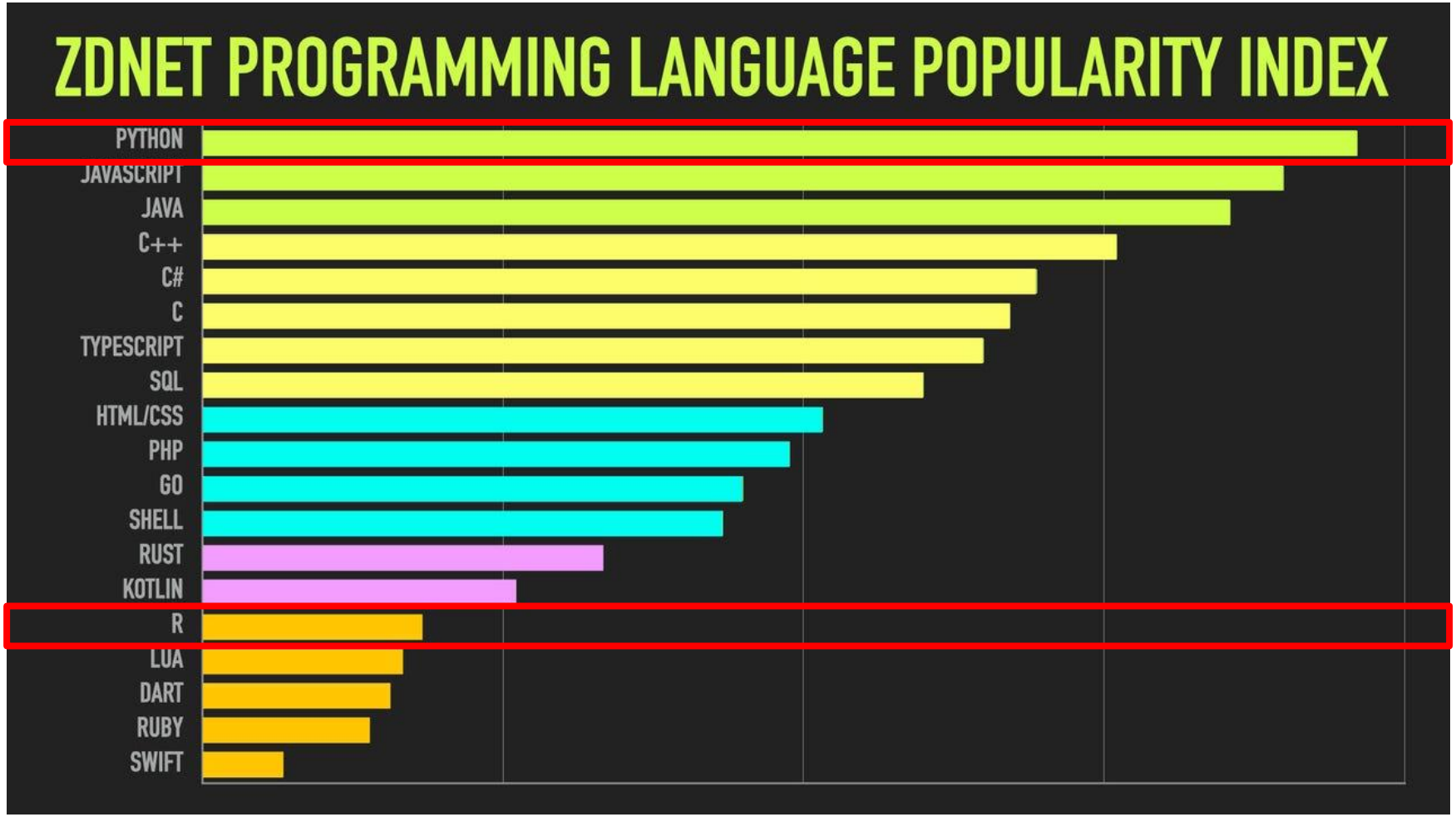
kaggle

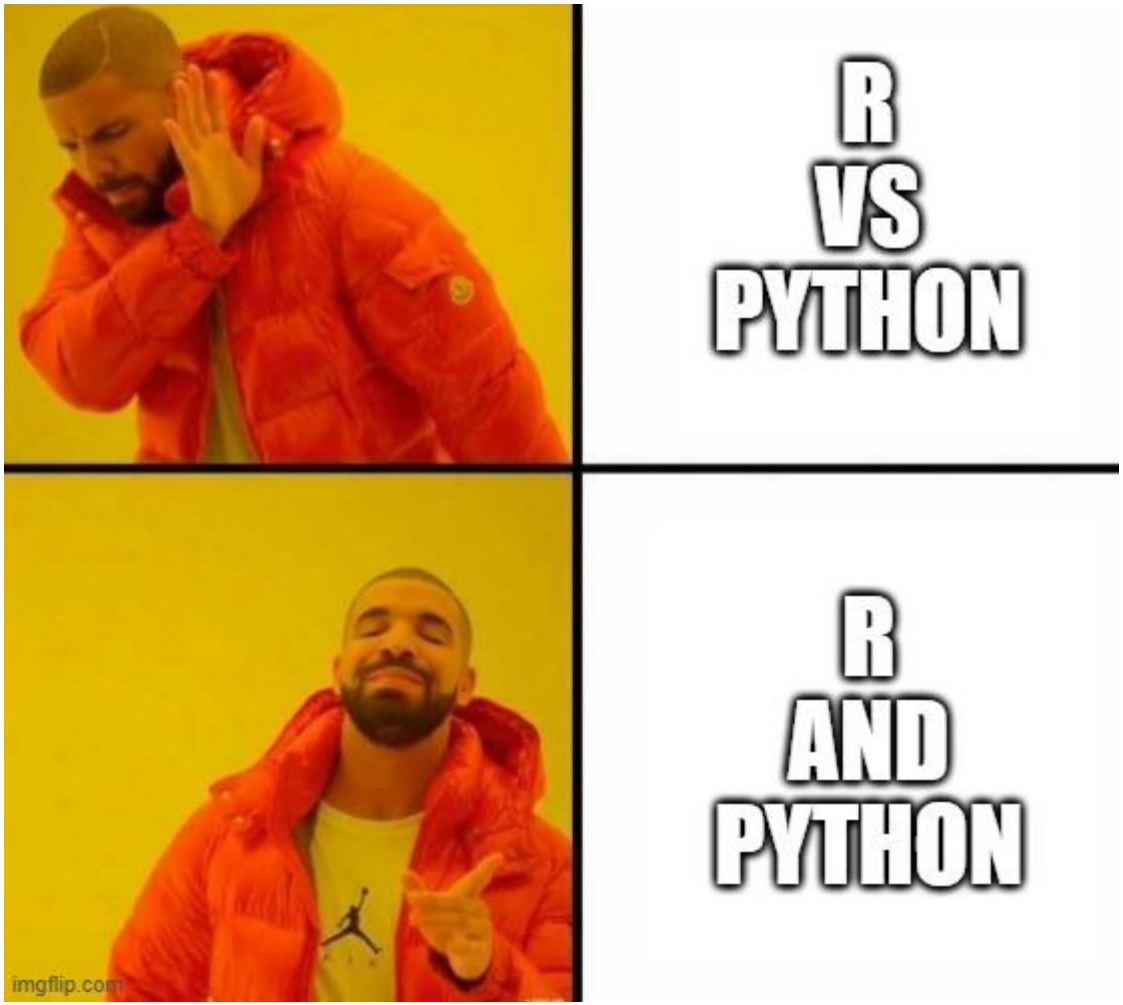
Model implementation

- Experimentation with the machine learning models can be done in R or Python.
- R – the language and environment for statistics, data mining, machine learning, graphical models
- Python – the simple language developed for general programming, but currently Lingua Franca for machine learning



Popularity 2025




R
VS
PYTHON
R
AND
PYTHON
imgflip.com

Python

- Anaconda (the distribution of Python)
  - https://www.anaconda.com/download
- conda install ipykernel (IPython Kernel for Jupyter)
- Install Visual Studio Code
  - https://code.visualstudio.com/
- Extensions -> Python (Install/Reload) -> Select Kernels -> base
- New File -> Jupyter Notebook

- Or use Google Colab

R

- Install R (https://cloud.r-project.org/)

- Install Visual Studio Code: https://code.visualstudio.com/
- Extensions -> R
- New File -> R file
- From terminal: install.packages("languageserver")
- For every new package referred in code:
  - Terminal: install.packages("nume")

- or Install RStudio


IT'S NOT THE PLANE
PYTHON
R
MAVE RICK
IT'S THE PILOT
imgflip.com


MACHINE LEARNING
MACHINE LEARNING
EVERYWHERE
makeameme.org