Foundations of Machine Learning (CS 725)

FALL 2024

Lecture 1:

- Introduction to Learning
- Course Administration and Trivia

Instructor: Preethi Jyothi

Machine Learning

- Ability of machines to “learn” from “data” or “past experience”

Machine Learning

- Ability of machines to “learn” from “**data**” or “**past experience**”

- **data/past experience:** Comes from various sources such as sensors, domain knowledge, experimental runs, etc.

Machine Learning

- Ability of machines to “**learn**” from “**data**” or “**past experience**”
- **data/past experience**: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.
- **learn**: Make accurate predictions or decisions based on data by optimizing a **model**

> “All models are wrong, but some are useful”, *George E. P. Box*

Pigeon Learning?

Video link: https://www.youtube.com/watch?v=TtfQlkGwE2U

Relationship between AI, ML, DL

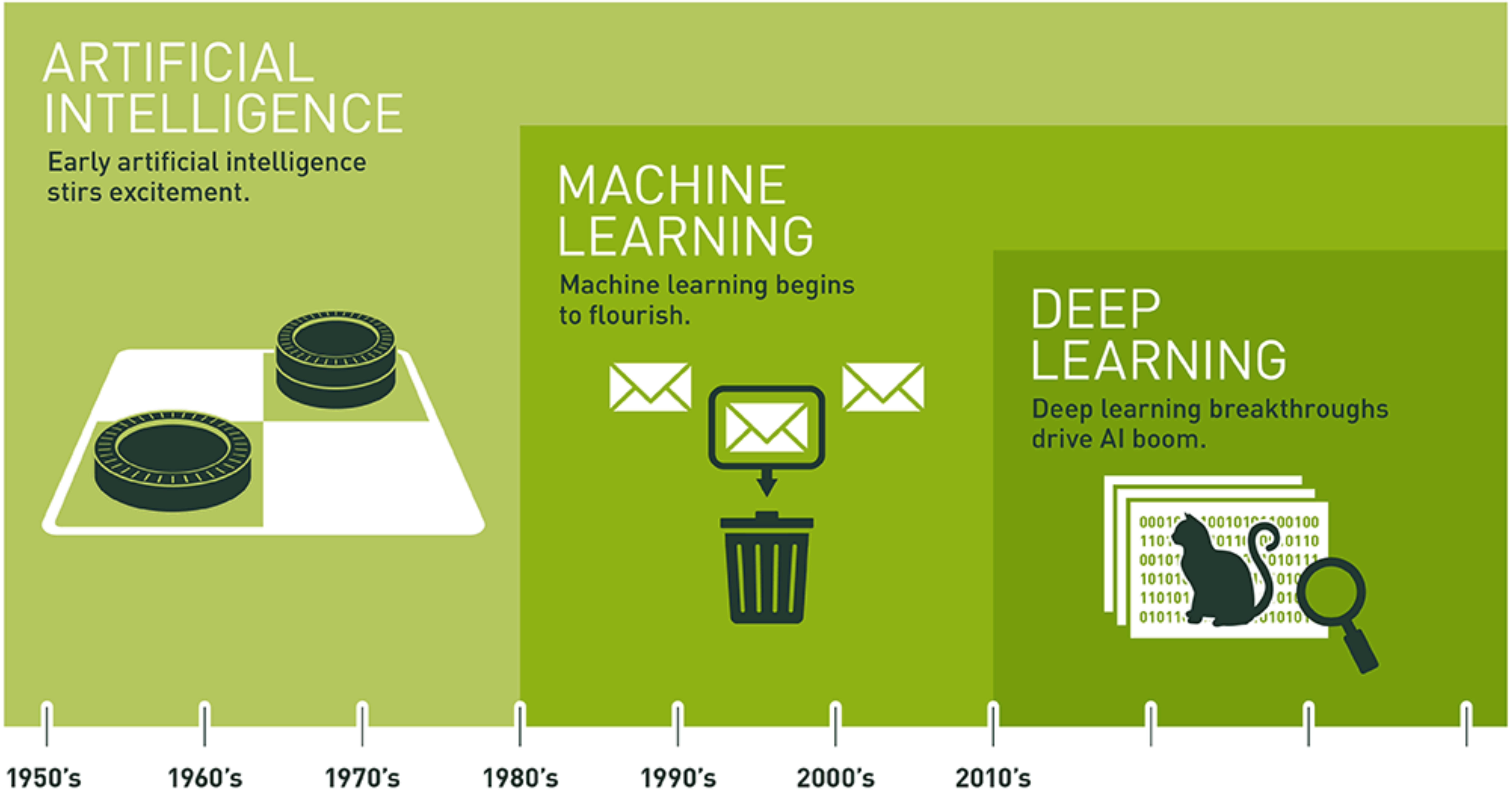
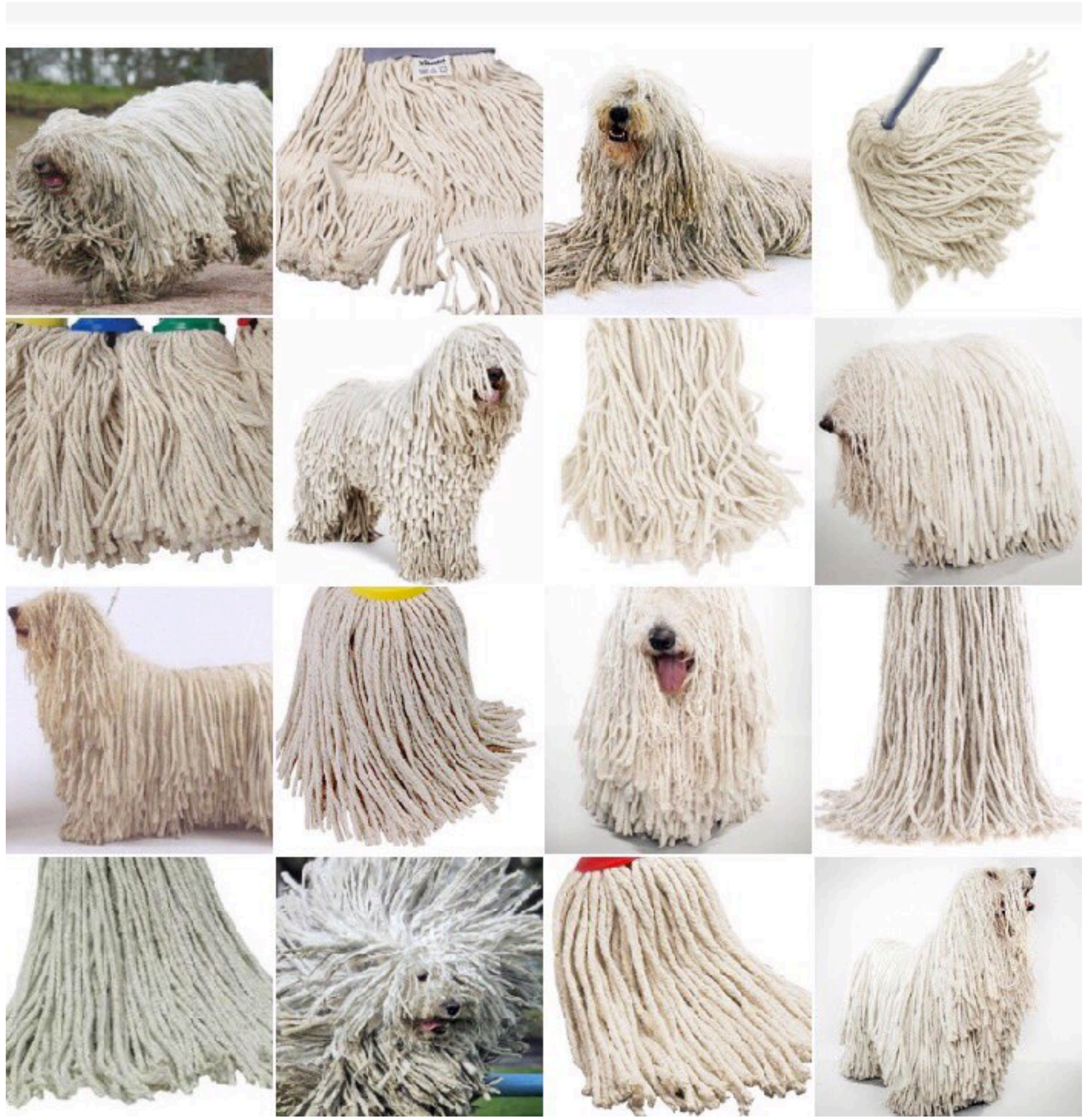

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Image from: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

When do we need ML? (I)

- For tasks that are easily performed by humans but are complex for computer systems to emulate



Sheepdog or mop?

Image from https://twitter.com/teenybiscuit/status/707670947830968320/photo/1 Photo courtesy: @teenybiscuit

When do we need ML? (I)

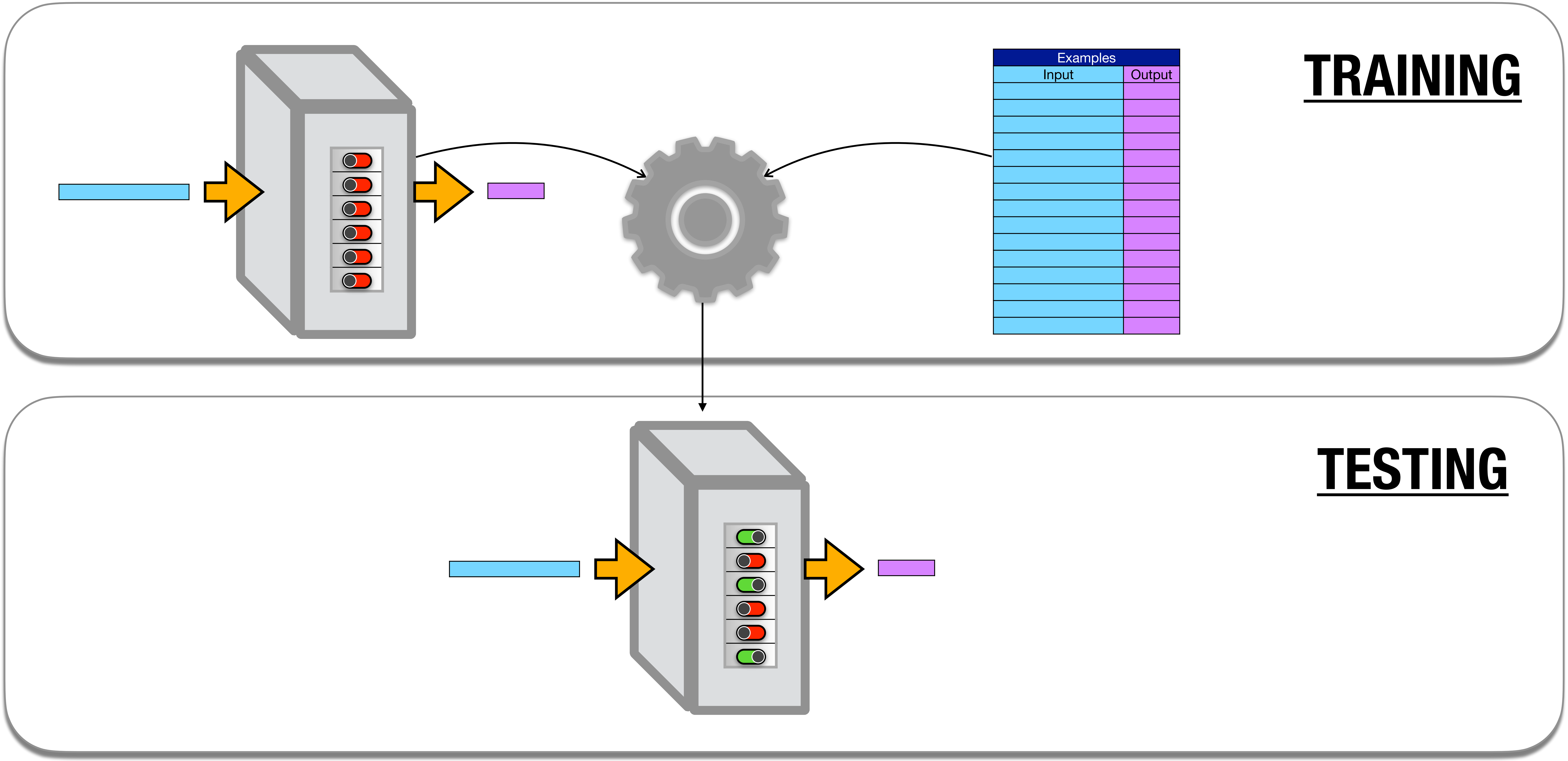
- For tasks that are easily performed by humans but are complex for computer systems to emulate
  - **Vision**: Identify faces in a photograph, objects in a video or still image, etc.
  - **Natural language**: Translate a sentence from Hindi to English, question answering, identify sentiment of text, etc.
  - **Speech**: Recognise spoken words, speaking sentences naturally
  - **Game playing**: Play games like chess, Go, Dota, Poker, etc.
  - **Robotics**: Walking, jumping, displaying emotions, etc.
  - Driving a car, navigating a maze, etc.

When do we need ML? (II)

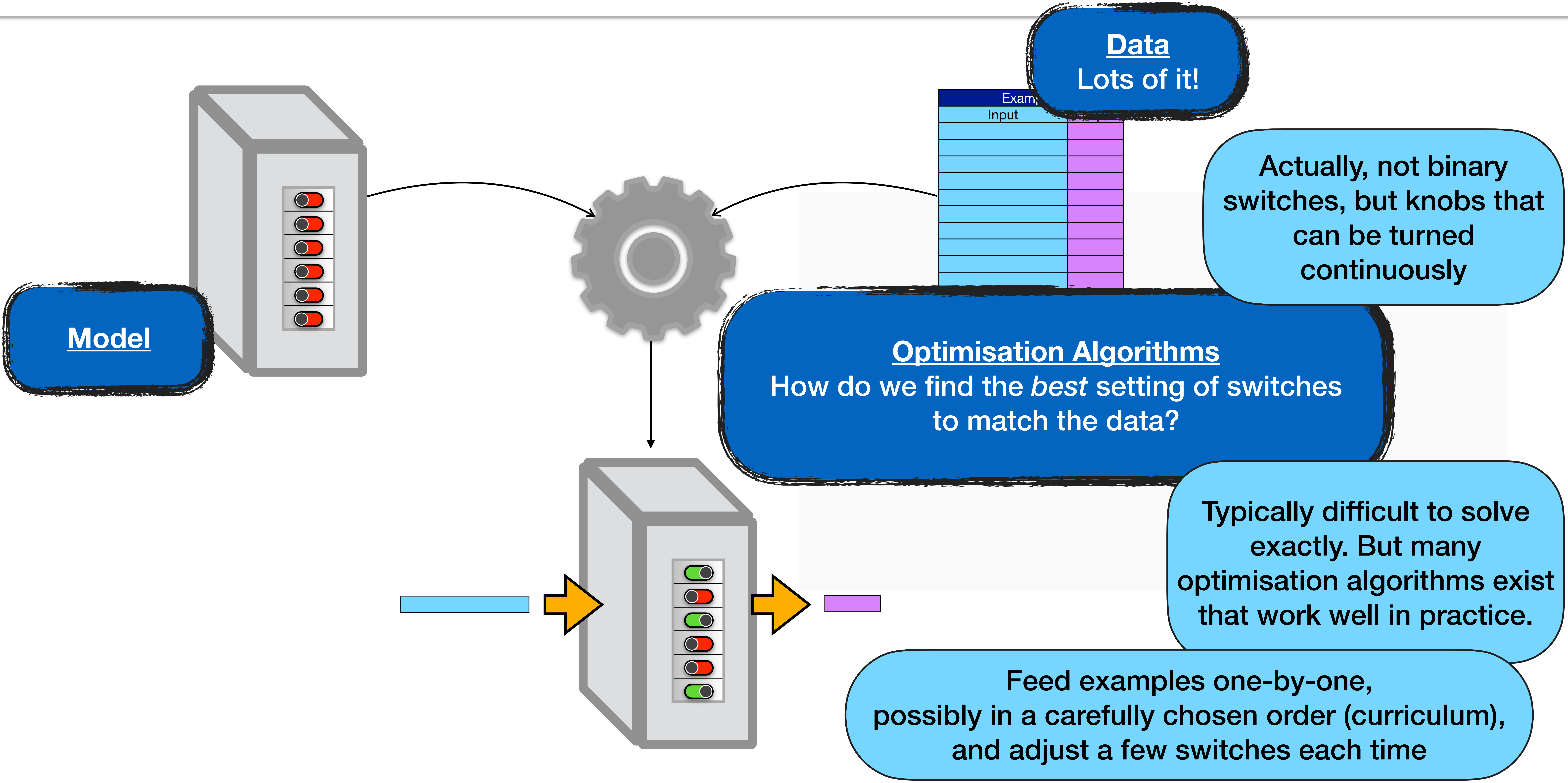
- For tasks that are beyond human capabilities
  - Analysis of large and complex datasets
  - E.g. IBM Watson’s Jeopardy-playing machine

Image credit: https://i.ytimg.com/vi/P18EdAKuC1U/maxresdefault.jpg

Components of Machine Learning



Components of Machine Learning



Data
Lots of it!

Model

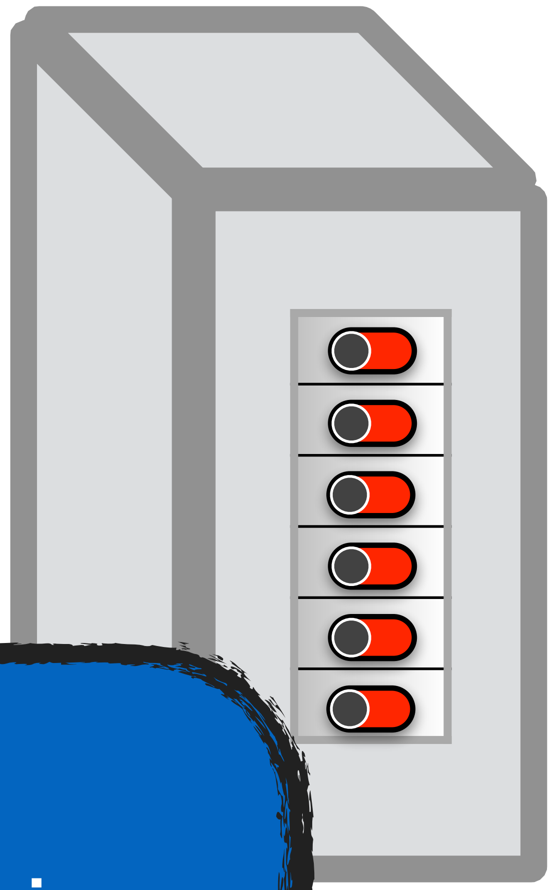
Actually, not binary switches, but knobs that can be turned continuously

Optimisation Algorithms
How do we find the *best* setting of switches to match the data?

Typically difficult to solve exactly. But many optimisation algorithms exist that work well in practice.

Feed examples one-by-one, possibly in a carefully chosen order (curriculum), and adjust a few switches each time

Components of Machine Learning



Models
What kind of machines can learn well?

Perils of Rote Learning

If a model *memorises* all the training examples,
it can achieve perfect score when tested on training data

But offers no advantage on *new* examples!

Moral: Good models should not overfit

Occam's Razor: models should be "simple"
(not have enough switches to record all the training data)

Remarkable Progress in ML

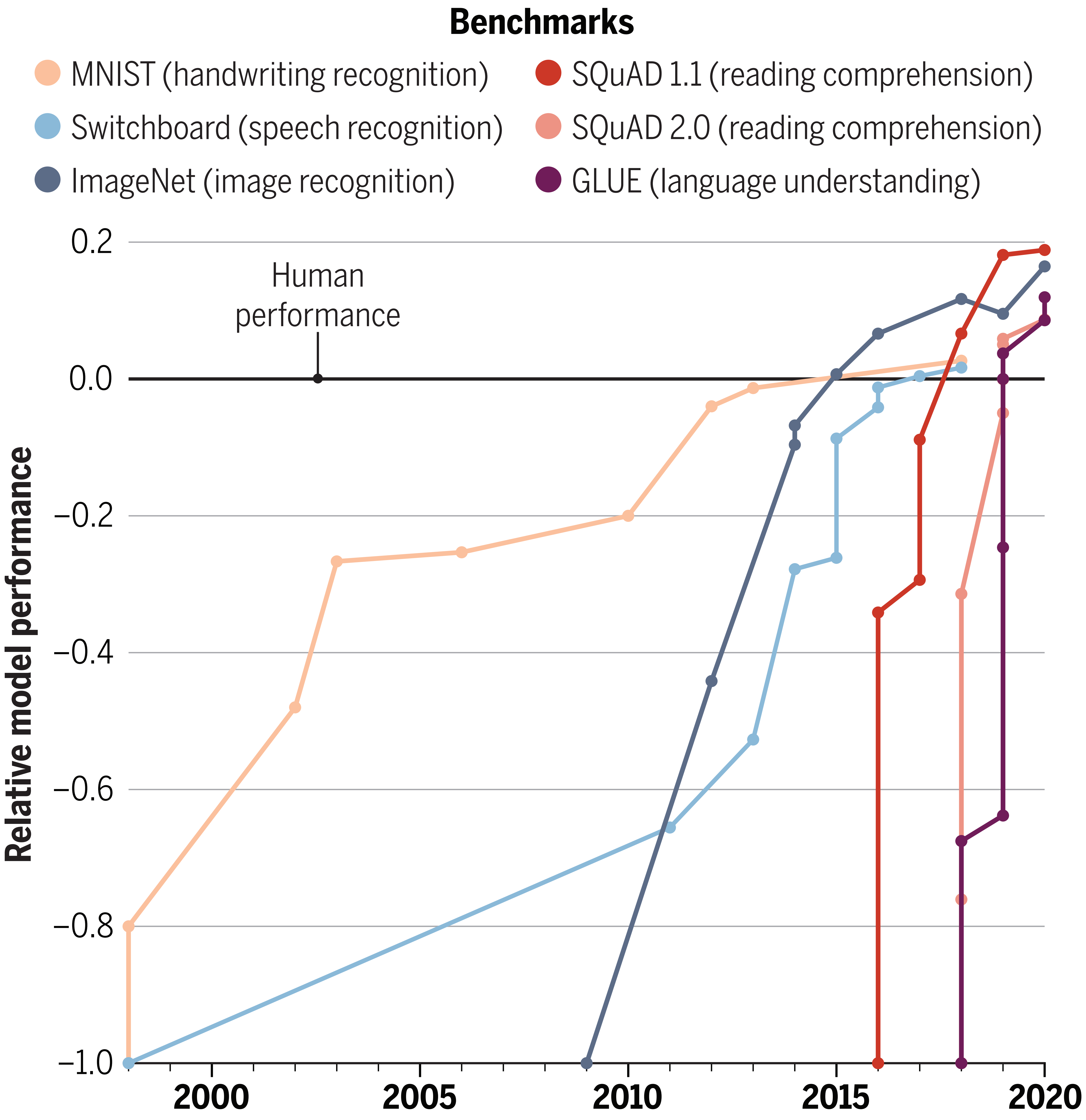
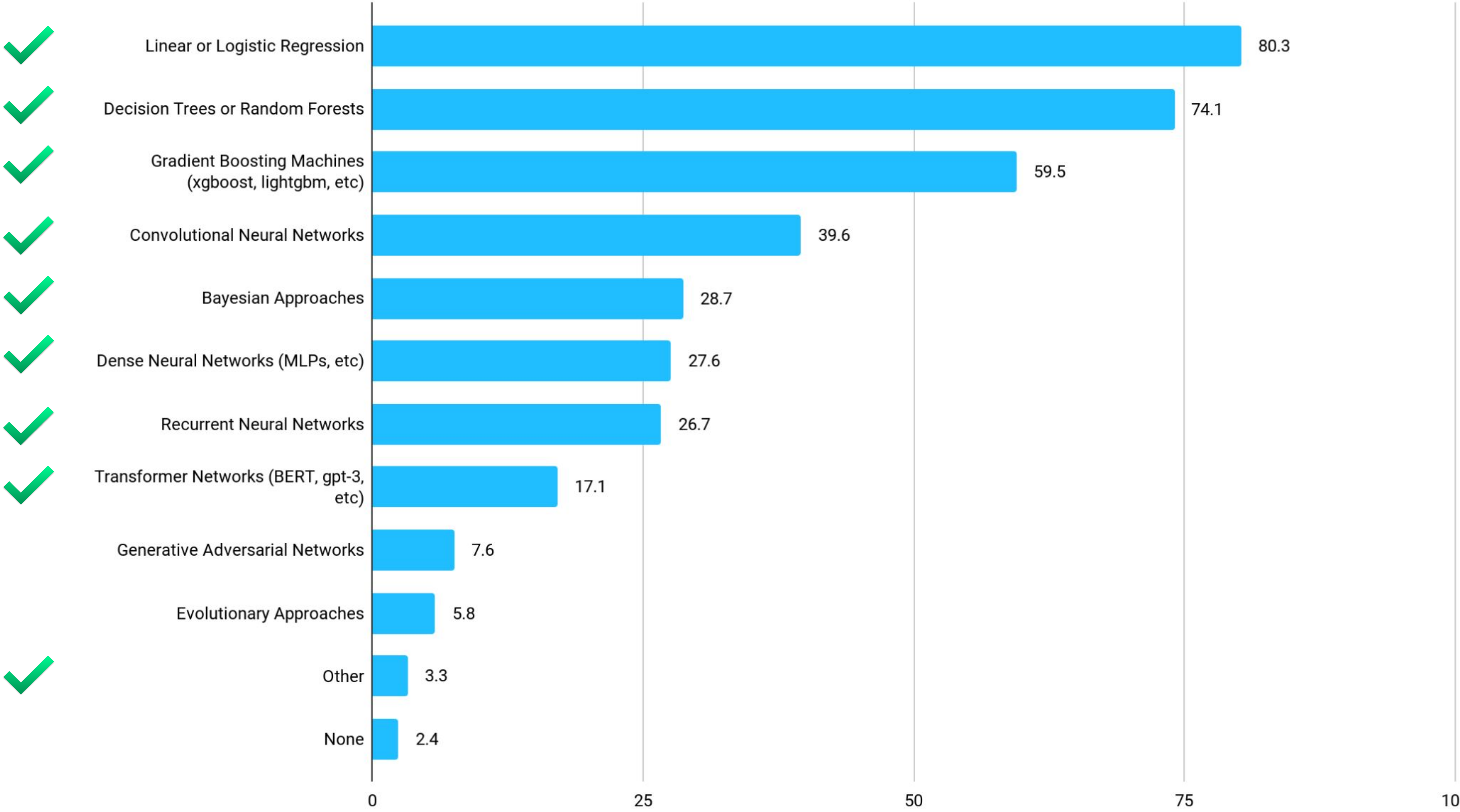


Image from “Taught to the Test” by Matthew Huston, Science, Volume 376, Issue 6593, pp 570—573, 2022

ML Algorithms



Graph from Kaggle's annual ML and Data Science Survey in 2021 linked at https://www.kaggle.com/c/kaggle-survey-2021/

Machine Learning

- Ability of machines to “**learn**” from “**data**” or “**past experience**”

- **data/past experience**: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.

- **learn**: Make accurate predictions or decisions based on data by optimizing a **model**

1. **Supervised learning**: Decision trees, neural networks, etc.

Machine Learning

- Ability of machines to “**learn**” from “**data**” or “**past experience**”
- **data/past experience**: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.
- **learn**: Make accurate predictions or decisions based on data by optimizing a **model**

1. **Supervised learning**: Decision trees, neural networks, etc.
2. **Unsupervised learning**: k-means clustering, PCA, mixture models, etc.

Machine Learning

- Ability of machines to “**learn**” from “**data**” or “**past experience**”
- **data/past experience**: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.
- **learn**: Make accurate predictions or decisions based on data by optimizing a **model**

1. **Supervised learning**: Decision trees, neural networks, etc.
2. **Unsupervised learning**: k-means clustering, PCA, mixture models, etc.
3. **Reinforcement learning**: Not covered in this course

ML Pipeline

- **Data:** Collect data for your problem.

  Labeled or unlabelled? What annotations?

- **Representation:** Choose features that represent your data.

  Raw? Expert-derived? Learned?

- **Modeling:** Choose a model for the task.

  Linear? Non-linear? What are the computational overheads?

- **Training/Learning:** Model will (likely) be parameterized and the parameters are learned using data. Choose an objective function to optimize.

  Which loss function? How best to optimize?

- **Prediction/Inference:** Given a model, assign labels to unseen test instances. Choose an evaluation metric.

  Automatic/manual evaluation?

Course Specifics / Administration / Trivia

Prerequisites

No official prerequisites

Should be comfortable with

- basic probability theory
- linear algebra
- multivariable calculus
- programming in Python (for assignments and project)

Course Syllabus

CS725 aims at providing an overview of ML and well-known ML techniques.

- Basic ML concepts (bias/variance/regularization), maximum likelihood/MAP estimates, linear and logistic regression
- Supervised learning: Decision trees, perceptron models, support vector machines, (deep) neural networks, convolutional/recurrent networks
- Unsupervised learning: k-means clustering, Gaussian mixture models
- Feature selection, dimensionality reduction, ensemble learning (boosting, bagging)
- Brief introduction to ML applications in computer vision, speech and natural language processing

No fixed textbook. Reference books (all available for free online) are listed on Moodle.

Course logistics

Class hours: Monday (2 pm - 3:30 pm), Thursday (2 pm - 3:30 pm)

Reading: All mandatory reading will be freely available online and linked on Moodle

Lecture Venue: LH 101 (exam venues will be announced close to the exam dates)

Attendance: Nothing mandatory

Course logistics

Class hours: Monday (2 pm - 3:30 pm), Thursday (2 pm - 3:30 pm).

Reading: All mandatory reading will be freely available online and linked on Moodle.

Lecture Venue: LH 101 (exam venues will be announced close to the exam dates).

Attendance: No fixed policy.

Asynchronous Q&A: Moodle discussion forum. Office hours (up on Moodle).

Class Announcements: Will be made on Moodle. Make sure you check these emails.

Frequent Moodle quizzes: Short quizzes consisting of 3-4 Qs and to be attempted offline (typically over the course of 2 days). Will count overall towards 5 participation points.

Quiz 0: First (ungraded) quiz will be live on Moodle at 11 pm on July 30th, 2024.
You will have until 11:59 pm on August 1, 2024 (Thursday) to complete the quiz.

Course TAs

Tejomay Kishor Padole (Ph.D., CSE)

Prateek Chanda (Ph.D., CSE)

Poulami Ghosh (Ph.D., CSE)

Darshan Prabhu (Ph.D., CSE)

Sona Elza Simon (Ph.D., CMInDS)

Sameer Anil Pimparkhede (M.S., CSE)

Soumen Kumar Mondal (M.S., CMInDS)

A. Snegha (M.S., CMInDS)

Evaluation (subject to small changes)

Two Programming Assignments	**(2 x 10 = 20%)**
Project	**(15%)**
Midsem exam	**(25%)**
Final exam	**(35%)**
Moodle quizzes (participation)	**(05%)**

Audit requirements:

Score 40% or above on either both assignments or assignment + project

Academic Integrity

Code of conduct:

Abide by an honour code and do not be involved in any plagiarism. No leeway here. If caught for copying or plagiarism, name of both parties will be handed over to the Department Disciplinary Action Committee (DDAC)[1].

- Write what you know.
- Use your own words.
- If you refer to *any* external material, ***always*** cite your sources. Follow proper citation guidelines.
- If you’re caught for plagiarism or copying, penalties are very high.

[1] http://www1.iitb.ac.in/newacadhome/punishments201521July.pdf

Final Project

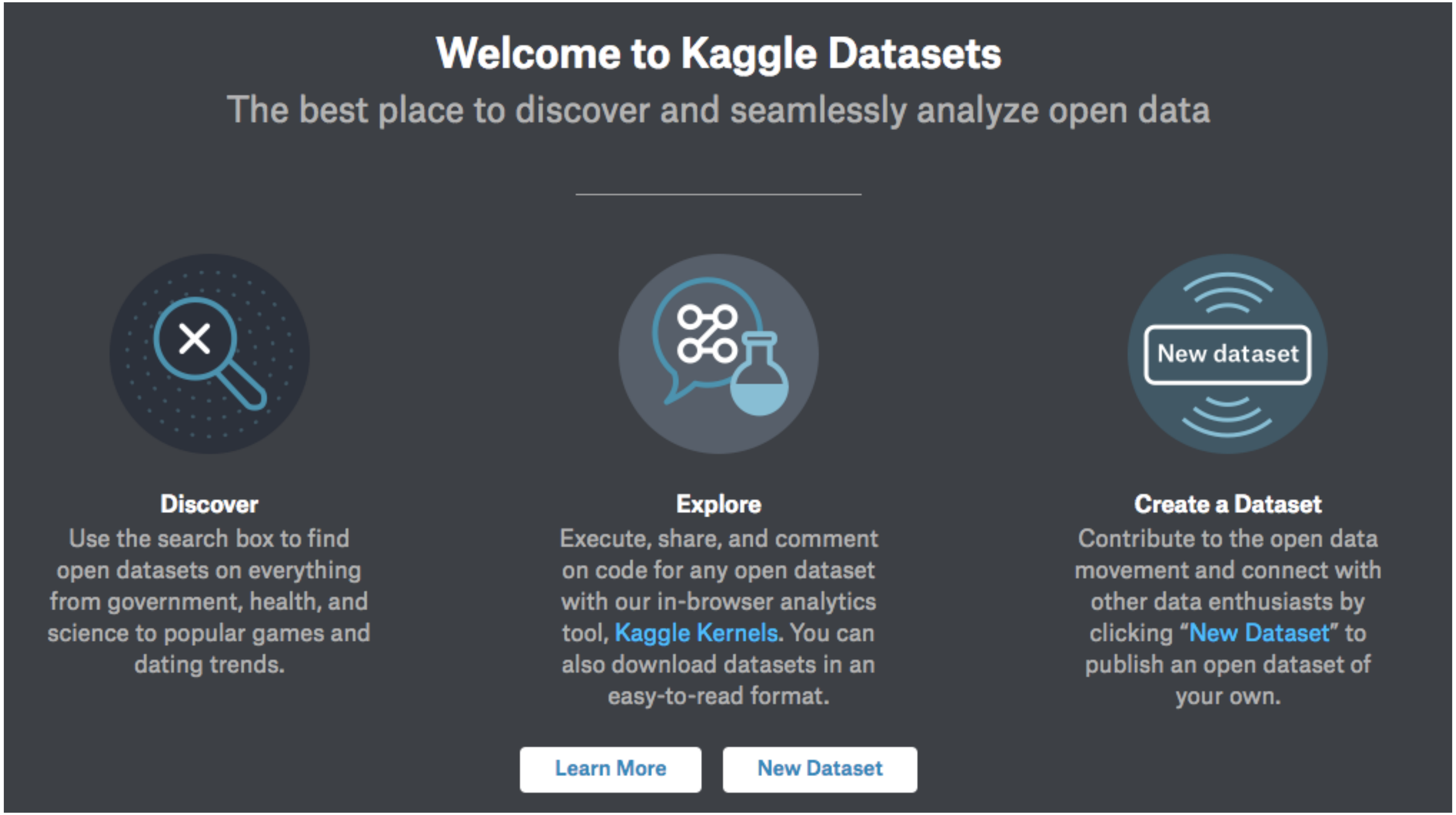
Team: 3-4 members. Find your team early!

Project details:

- Apply the techniques you studied in class to any interesting problem of your choice
- Think of a problem early and work on it throughout the course. Stick to existing datasets as far as possible. Project milestones will be posted on Moodle.
- Examples of project ideas: auto-complete code, help irctc predict ticket prices, sentiment analysis of tweets, image retrieval using captions, etc.
- Consult with me/TAs to check feasibility of project idea. Do not choose compute-intensive problems.

Datasets abound…

Kaggle: https://www.kaggle.com/datasets



Welcome to Kaggle Datasets

The best place to discover and seamlessly analyze open data

Discover

Use the search box to find open datasets on everything from government, health, and science to popular games and dating trends.

Explore

Execute, share, and comment on code for any open dataset with our in-browser analytics tool, Kaggle Kernels. You can also download datasets in an easy-to-read format.

Create a Dataset

Contribute to the open data movement and connect with other data enthusiasts by clicking "New Dataset" to publish an open dataset of your own.

Datasets abound…

Kaggle: https://www.kaggle.com/datasets

Another good resource: http://deeplearning.net/datasets/

Popular resource for ML beginners:
http://archive.ics.uci.edu/ml/index.php

Interesting datasets for computational journalists:
http://cjlab.stanford.edu/2015/09/30/lab-launch-and-data-sets/

Speech and language resources:
www.openslr.org/

… and so do ML libraries/toolkits

scikit-learn, openCV, Keras, PyTorch, Tensorflow, etc.

Next Class: Introduction to Linear Regression