Artificial Intelligence and Machine Learning (CS 337/335)

FALL 2023

Lecture 1a:

- 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*

Relationship between AI, ML, DL

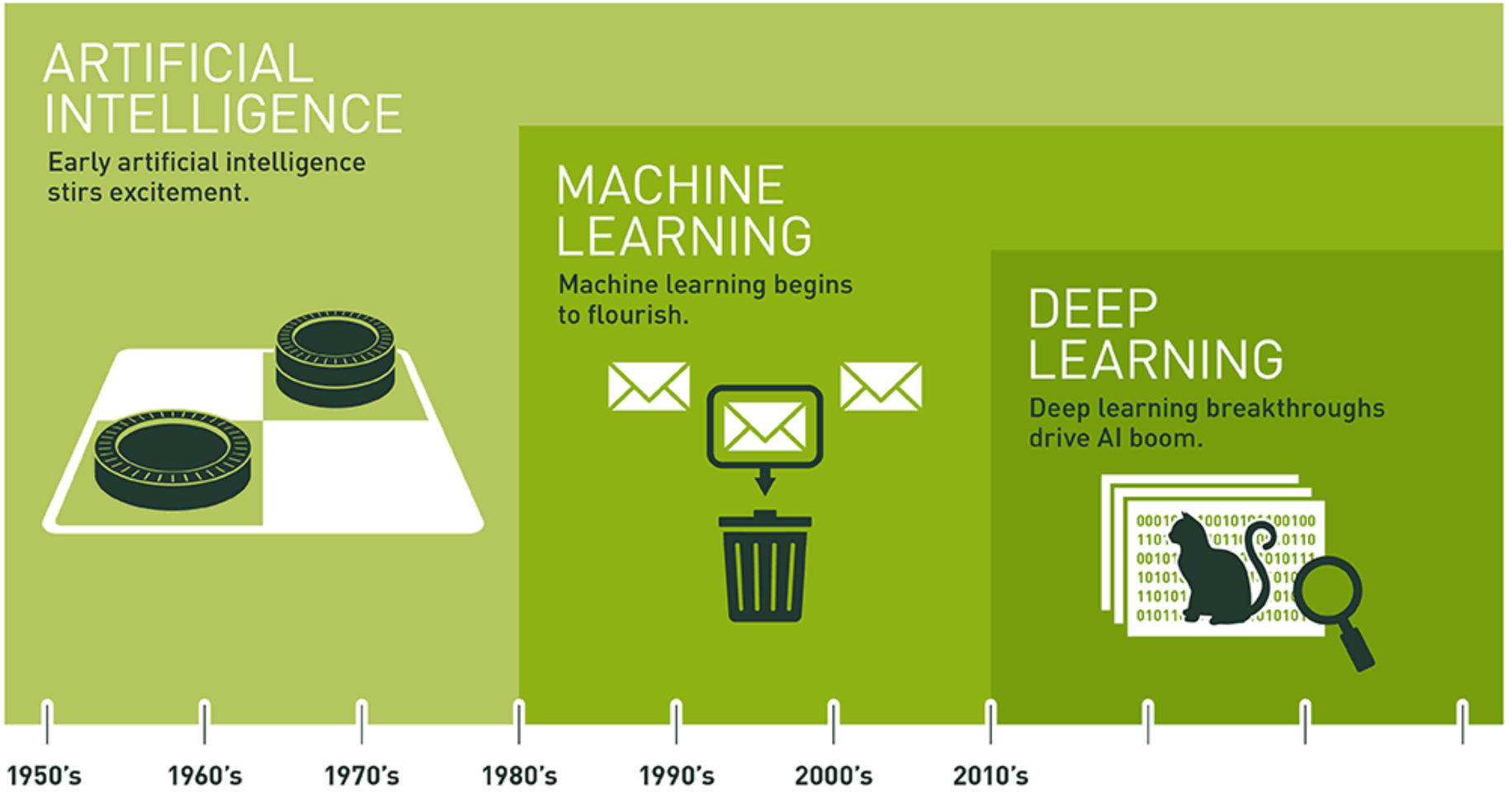


Image from: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

ML and Statistics?

Image source: Reddit

ML and Statistics?

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Glossary from: http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf

When do we need ML? (I)

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

Dog or Muffin?

Image from Reddit

- For tasks that a … e complex for computer syste…
  - **Vision**: Identi… …leo or still image, etc.
  - **Natural langu**… … English, question answ…
  - **Speech**: Rec… …ces naturally
  - **Game playin**… …ker, etc.
  - **Robotics**: Wa… …
  - Driving a car, …

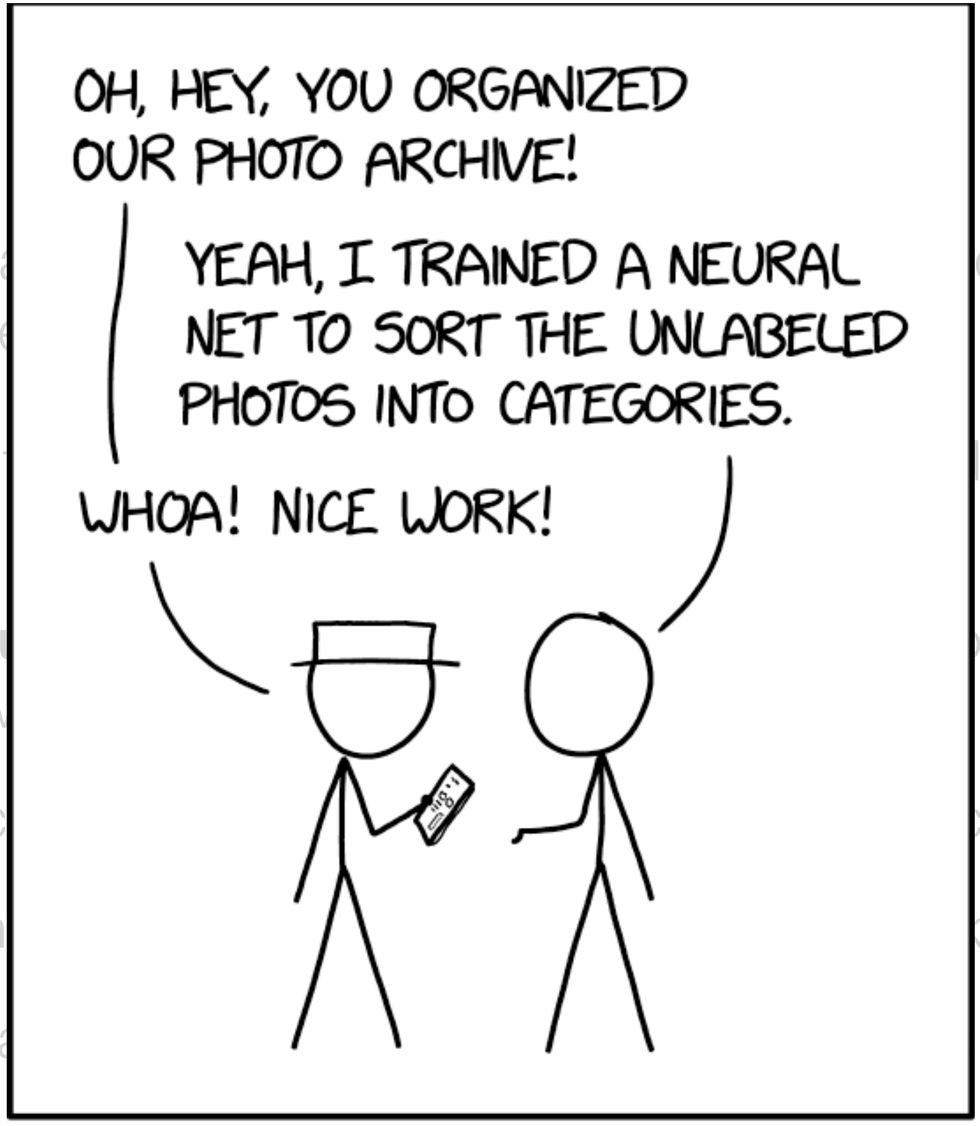


Image from:https://xkcd.com/2173/

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

Remarkable Progress in ML

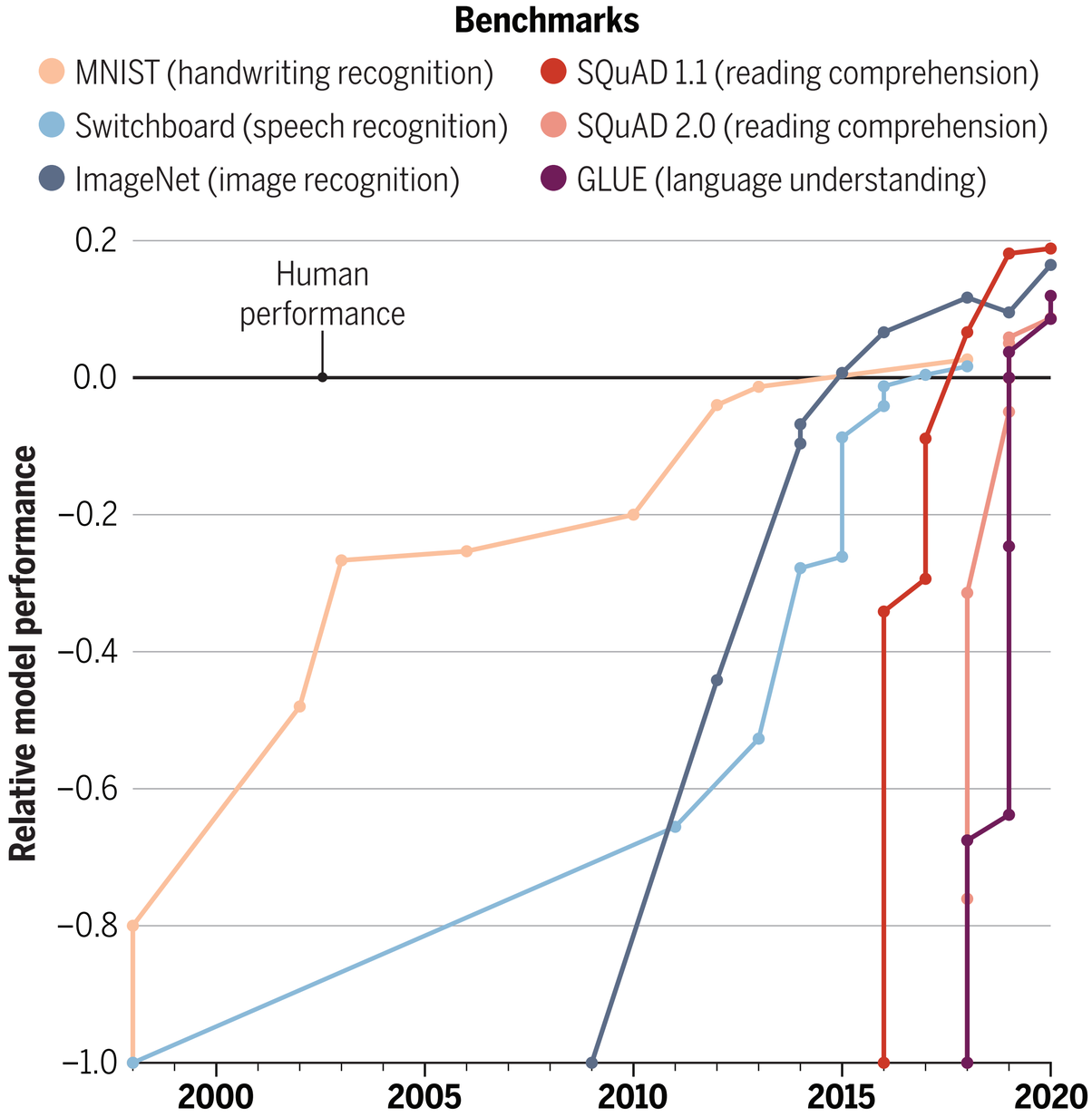
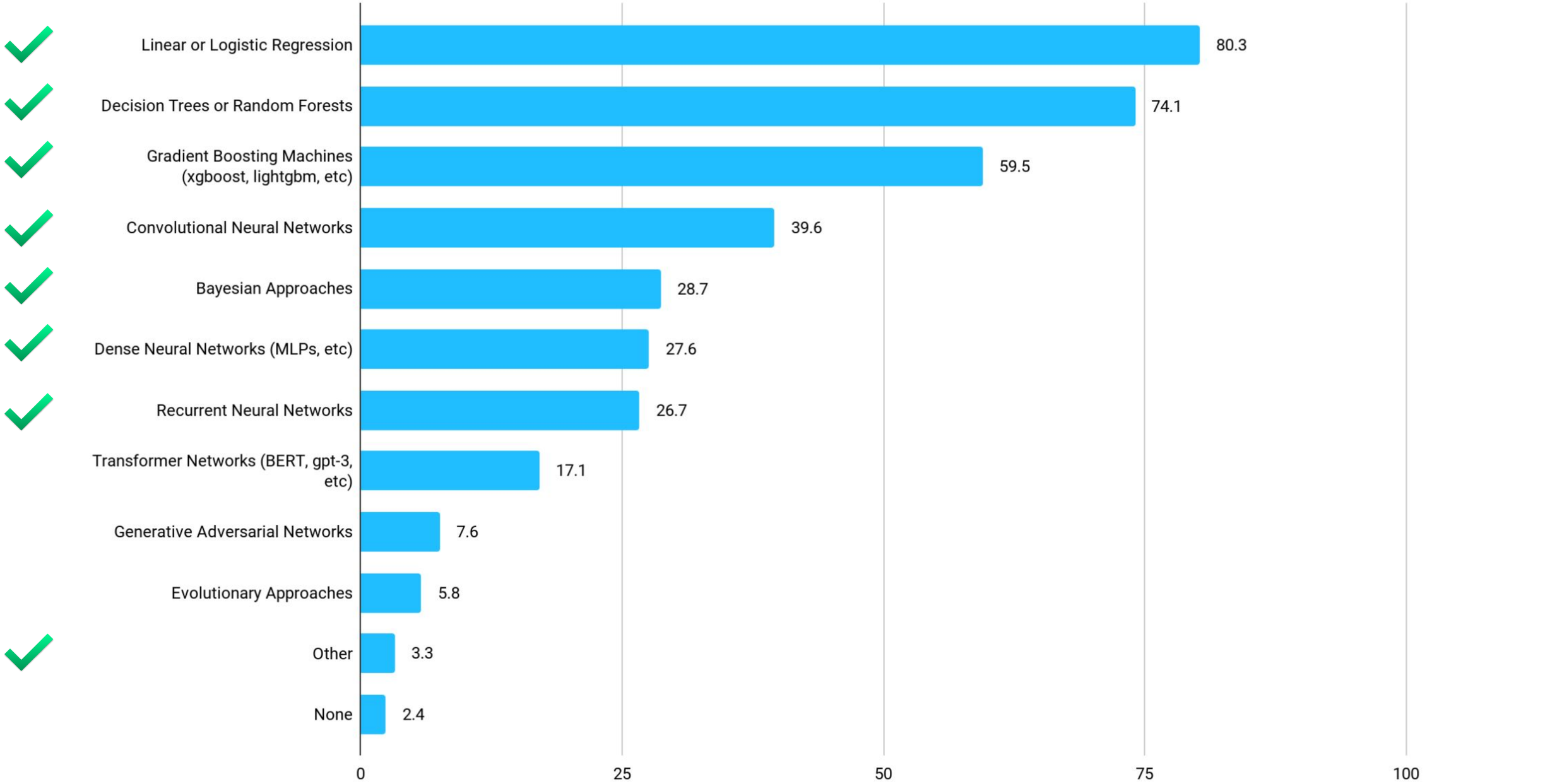


Image from “Taught to the Test” by Matthew Huston, Science, Volume 376, Issue 6593, pp 570—573, 2022

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

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

Machine Learning

- Ability of computers to “**learn**” from “**data**” or “past experience”
- **data**: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.
- **learn**: Make intelligent 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 computers to “**learn**” from “**data**” or “past experience”
- **data**: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.
- **learn**: Make intelligent 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 Logistics

Teaching Assistants

Barah Fazili (Ph.D.)
S Durga (Ph.D.)
Sanjeev Kumar (Ph.D.)

Ashish Sunil Agrawal (M.S.-2)
Tejpalsingh Baljeetsingh Siledar (M.S.-2)
P S V N Bhavani Shankar (M.Tech.-2)
Dhiraj Kumar Sah (M.Tech.-2)
Saurabh Kumar (M.Tech.-2)
Vishal Ashok Tapase (M.Tech.-2)

Mayank Jain (B.Tech.-4)
Vedang Dhirendra Asgaonkar (B.Tech.-4)
Ashwin Ramachandran (B.Tech.-4)
Govind Saju (B.Tech.-4)
Parshant Arora (B.Tech.-4)

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 Disciplinary Action Committee (DAC)[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].

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

CS 337 (Theory) Logistics

CS 337 Trivia

Class hours (venue: LH 102):
Monday (9:30 am - 10:25 am)
Tuesday (10:35 am - 11:30 am)
Thursday (11:35 am - 12:30 pm)

Attendance: No attendance criteria (for now)

Asynchronous Q&A: Moodle discussion forums

Class Announcements: Will all be made via Moodle announcements.

Preferred mode of Communication: Email with CS 337 and/or CS 335 in the subject line

Instructor office hours:
Tuesdays, 5 pm to 6 pm (1st half of the semester),
Wednesdays, 11 am to 12 pm (2nd half of the semester)

Basic Prerequisites: Should be comfortable with basic probability theory, linear algebra, multivariable calculus, programming in Python (for all CS 335 assignments)

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

Evaluation (subject to small changes)

Participation	**(05%)**
(Best of) Two In-class Quizzes	**(15%)**
Midsem exam	**(25%)**
Final exam	**(40%)**
Course Project	**(15%)**

Participation (5%):

- Every student should act as a scribe **twice** to help prepare class notes
- Thirty four 55-min lectures across the entire semester
- Two groups of students (comprising 2-4 students each) act as scribes for each lecture
- Typeset the notes using LaTeX template (will be shared on Moodle)
- Assigned TA will merge notes from both groups, proofread and publish on Moodle

Course Project

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

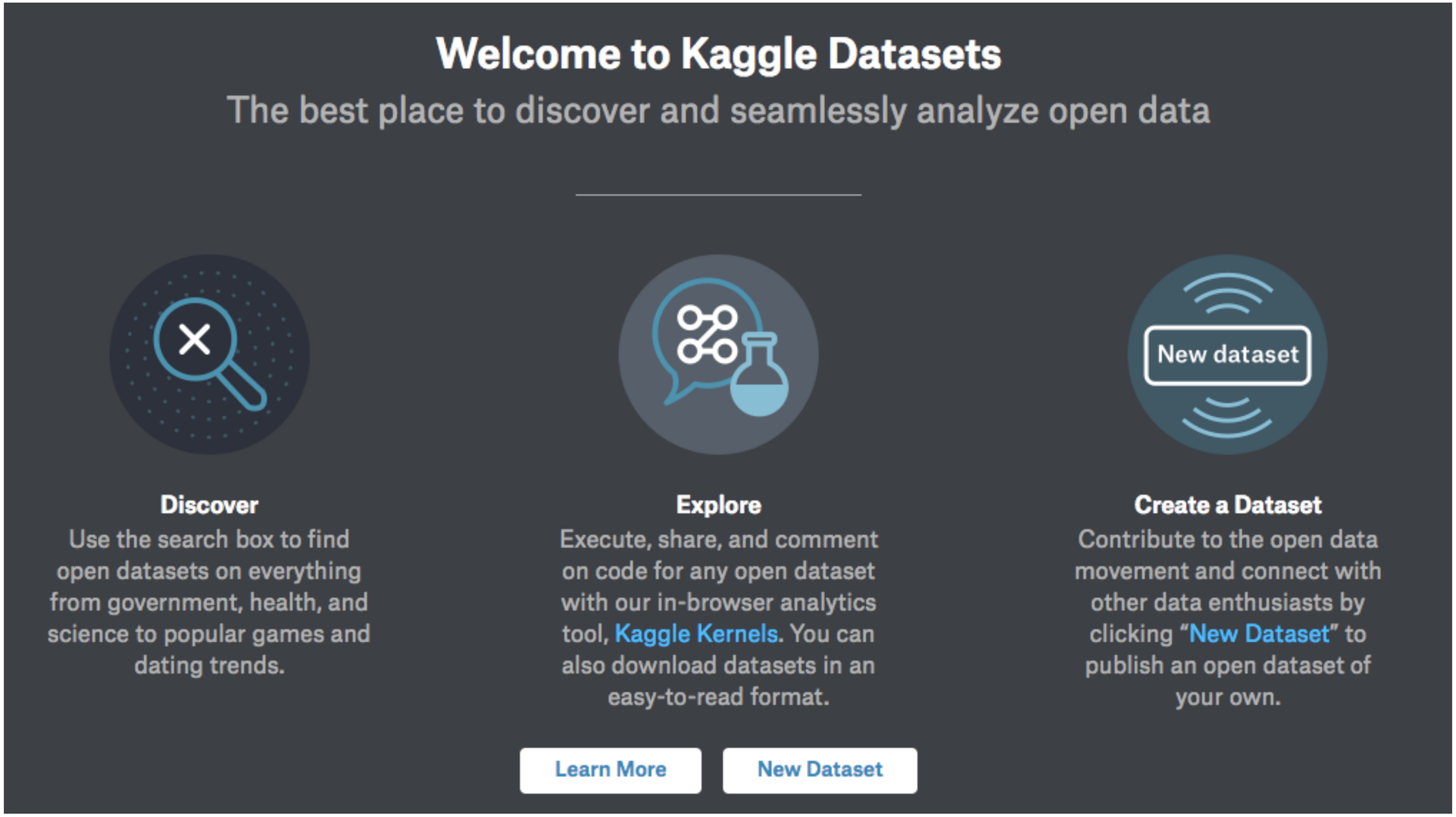
Evaluation: Submit a project report and make a short presentation after endsem exams

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.
- Project milestones will be posted on Moodle.
- Examples of project ideas: auto-complete code, help irctc predict ticket prices, image retrieval using captions, etc. Consult with me/TAs to check feasibility of project idea.

Many datasets…

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



Many datasets…

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

Good resource for NLP datasets: https://huggingface.co/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/

More speech and language resources:
www.openslr.org/

… and many ML libraries/toolkits

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

CS 335 (Lab) Logistics

Tentative Schedule

Aug 7: Probability/matrix-vector calculations, simple ML pipeline and Kaggle

Aug 14: Linear regression (closed form), gradient descent

Aug 21: Regularization, lasso/ridge regression

Aug 28: Logistic regression (using gradient descent), naive Bayes classifier

Sep 4: Decision Trees

Sep 11: Perceptron + SVM classifiers

Sep 18: QUIZ

Sep 25: Feedforward NNs + backprop

Oct 2: HOLIDAY

Oct 9: Regularizing NNs, optimizers, CNN-based classifier

Oct 16: Simple NLP, text-based classification using embeddings

Oct 23: PCA, dimensionality reduction

Oct 30: Clustering, k-means

Nov 6: Ensemble classifiers, boosting/bagging, random forests

Evaluation (subject to small changes)

In-lab problems	**(35%)**
Quiz 1	**(15%)**
Final exam	**(50%)**

- Program templates will all be in Python
- Read submission instructions very carefully and adhere to the specified format
- In-lab graded problems will be spread across different lab sessions and announced prior to the session
- Attendance is mandatory for labs