Introduction to Machine Learning (CS419M)

Lecture 1:

- What is learning?
- Supervised vs. unsupervised learning
- Basic course administration and trivia

Jan 10, 2020

What is Machine Learning?

- Machine Learning (ML) is a sub-field of computer science that evolved from the study of **pattern recognition** and **computational learning theory** in artificial intelligence.

- Using *algorithms* that iteratively learn from *data*

- Allowing computers to discover *patterns* without being explicitly programmed where to look

Relationship between AI, ML, DL

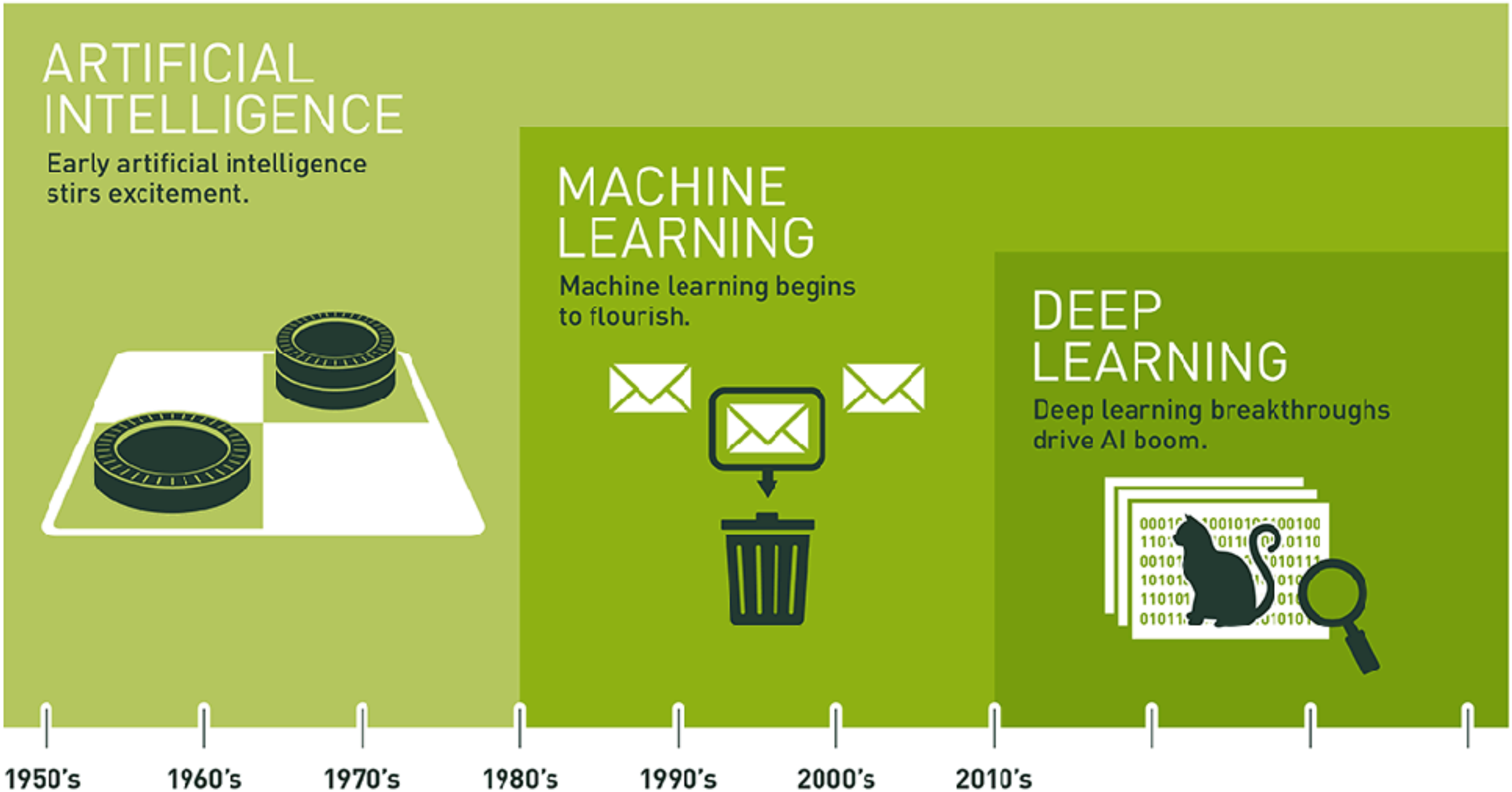


Image from: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

ML and Statistics?

Glossary

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
  - **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.
  - **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

Machine Learning

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

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.

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

Example: Image Recognition

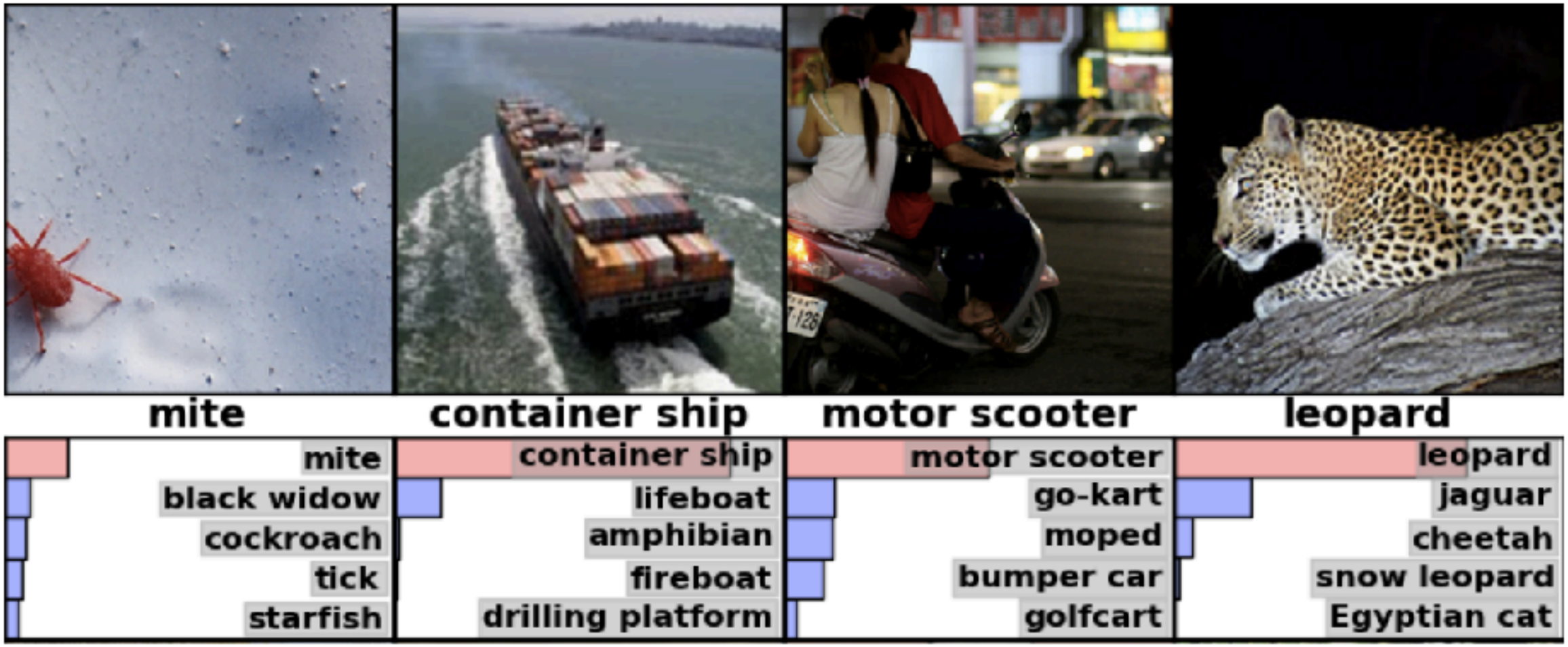


Image from "ImageNet classification with deep CNNs", Krizhevsky et al.

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, 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, etc.
    3. *Reinforcement learning: Not covered in this course.*

Course Specifics / Administration / Trivia

Prerequisites

No official prerequisites.

Should be comfortable with

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

Course Webpage

https://www.cse.iitb.ac.in/~pjyothi/cs419/

Introduction to Machine Learning, Spring 2020

Course Description

Welcome to *"Introduction to Machine Learning 419(M)"*. In this undergraduate-level course, you will be introduced to the foundations of machine learning along with a slew of popular machine learning techniques. This will also give you insights on how to apply machine learning to solve a new problem.

This course is open to any non-CSE undergraduate student who wants to do a minor in CSE. There are no prerequisites.

Course Info

Time: Wednesdays, Fridays, 9.30 am to 10.55 am
Venue: CC 103
Instructor: Preethi Jyothi. You can email me at pjyothi [at] cse [dot] iitb [dot] ac [dot] in

---

TAs (expand [at cse] as in the instructor's email ID above):
Srijon Sarkar (email: srijon [at cse])
Navya Muttineni (email: mnavya [at cse])
Shivam Sood (email: ssood [at cse]
Mayur Warialani (email: mayurwarialani [at cse])
Achari Rakesh Prasanth (email: rakeshprasanth [at cse])
Rishabh Kumar (email: krrishabh [at cse])

Instructor office hours (in CC 221): 4 to 5 pm on Wednesdays

TA office hours: TBA

Course logistics

Reading: All mandatory reading will be freely available online and posted on the course website.

Textbooks (available online):

1. Understanding Machine Learning. Shai Shalev-Shwartz and Shai Ben-David. Cambridge University Press. 2017.
2. The Elements of Statistical Learning. Trevor Hastie, Robert Tibshirani and Jerome Friedman. Second Edition. 2009.

Attendance: 60% minimum attendance. Counts towards participation points. Strongly advised to attend class. Lot of material will be covered in class, which will not be on the slides.

Personnel and Academic Integrity

Course TAs: Srijon Sarkar, Navya Muttineni, Shivam Sood, Mayur Warialani, Achari Rakesh Prasanth, Rishabh Kumar

Communication:

We will use Moodle for all course-related announcements.
My office hours: 4 pm to 5 pm on Fridays
TA's office hours: TBA. Will be spread out over the week.

Code of conduct:

Abide by an honour code and not be involved in any plagiarism. If caught for copying or plagiarism, name of both parties will be handed over to the Disciplinary Action Committee (DAC)[1].

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

Course Syllabus

Provide an overview of machine learning and well-known ML techniques. We will briefly cover some ML applications as well.

Some Topics:

- Basic foundations of ML, classification/regression, Naive Bayes' classifier, linear and logistic regression
- Supervised learning: Decision trees, perceptron, support vector machines, neural networks.
- Unsupervised learning: k-means clustering, EM algorithm.
- Other topics: feature selection, dimensionality reduction, boosting, bagging.
- Brief introduction to ML applications in computer vision, speech and natural language processing.

Evaluation (subject to minor changes)

Two programming assignments	(**20%**)
Two quizzes	(**20%**)
Midsem Exam	(**20%**)
Final Exam	(**25%**)
Project	(**10%**)
Participation	(**05%**)

Audit requirements:

Both assignments, both quizzes, participation points.
Score 50% or above to successfully audit the course.

Final Project

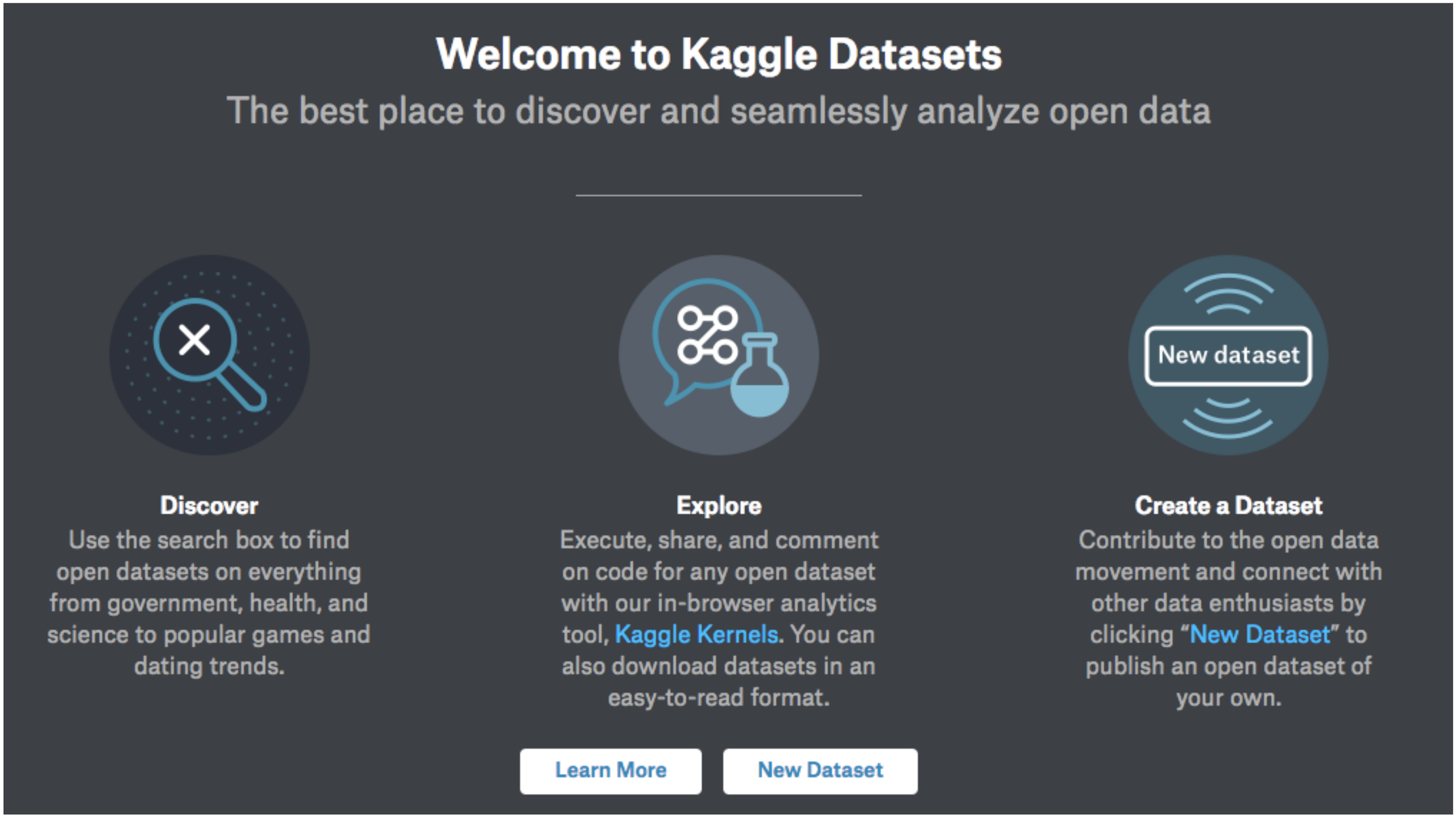
Team: 2-3 members. Individual projects are highly discouraged.

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. Project milestones will be posted on Moodle.
- Examples of project ideas: auto-complete code, generate song lyrics, help irctc predict ticket prices, etc.
- Feel free to be creative; consult with TAs/me if it’s feasible

Datasets abound…

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



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, Tensorflow, NLTK, etc.

Typical ML approach

- How do we approach an ML problem?
  - **Modeling**: Use a model to represent the task
  - **Decoding/Inference**: Given a model, answer questions with respect to the model
  - **Training**: The model could be parameterized and the parameters are estimated using data

How do we know if our model's any good?

- **Generalization**: Does the trained model produce good predictions on examples beyond the training set?
- We should be careful not to *overfit* the training data
    - Occam's Razor: All other things being equal, pick the simplest solution
- These concepts will be made more precise in later classes