THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

CS 4364/6364

Machine Learning

Fall Semester 8/24/2023

Lecture 1.

Introduction to ML

John Sipple

jsipple@gwu.edu

Today’s Topics

Welcome!

Introduction to Machine Learning

Fundamental Concepts of ML

Course Roadmap

Homework 1 (due 9/12)

Readings

Instructor

John Sipple (jsipple@gwu.edu)

Since 2003: Software Engineer, Data Scientist and ML Engineer

Staff Software Engineer and Tech Lead at Google's Core Enterprise ML team

Technology Consultant for multiple AI startups

Research interests:

- Reinforcement Learning in the Real World
- Intelligent Diagnostics
- Generative Language Models
- Internet of Things

Grader

Mike Rossetti (rossetti@gwu.edu)

Data scientist and software engineer. He has worked as a data analyst for a winning US Presidential campaign, an analytics director for a Silicon Valley startup, and a technology consultant for the US Government. He teaches courses in data science, computer science, and software development, and conducts research in applied machine learning.

Research interests:

- Applied Machine Learning
- Unsupervised Learning
- Natural Language Processing
- Systems Development Methodologies
- Software Testing and Verification
- Information Security and Privacy
- Social and Information Networks

Welcome

Purpose: To provide the important theoretical basis for continued research and development skills for applying Machine Learning to real-world problems.

Course subdivided into four phases:

I. Fundamentals
II. Supervised Methods
III. Unsupervised Methods
IV. Advanced Topics

Lectures

Lectures will be in-person (or Zoom):

- Tuesday, Thursday 9:35 - 10:50, Tompkins 208
- Slides & Recording will be posted to the course website next day
- Please complete the recommended readings listed on the schedule before

ML Office Hours:

- Wednesdays: 9:00 - 10:00 am (via Zoom)
- By appointment (24hr notice, please)

Homework Office Hours

- Wednesdays: 7:00 - 8:30 pm (via Zoom)

Prerequisites

Mathematics and Statistics

- Calculus
- Linear Algebra
- Probability and Statistics

Programming/Software Development

- Python experience not required, but must have programming skills in at least one object-oriented programming language
- Software patterns and algorithms

Course at Glance

- Fundamentals (5 lectures)
- Supervised Machine Learning (8 lectures)
- Unsupervised Machine Learning (3 lectures)
- Advanced Topics (10 lectures)

Assignments and Grading

25%	Six Assignments, each worth 5%, focused on practical application. No late submissions. Top 5 graded.
20%	Midterm exam, covering Fundamentals and Supervised Methods
30%	Group or Individual Project, technical paper, demo and presentation
25%	Final exam, covering Fundamentals, Supervised Methods, Unsupervised Methods and Advanced Topics

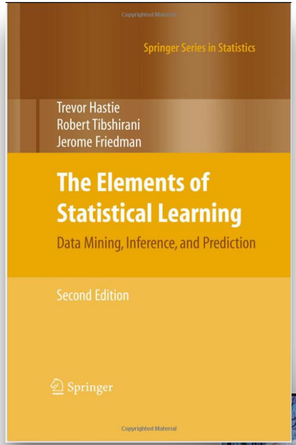
Required Textbooks

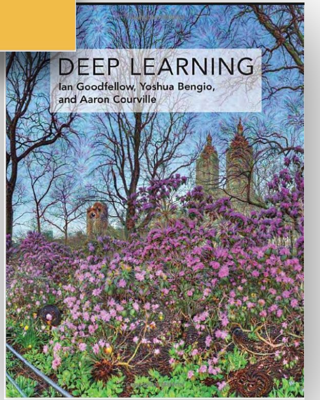
The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) 2nd Edition by Hastie, Tibshirani, Friedman

- Available on Amazon (hardcover or eTextbook)
- Online (free) at http://www-stat.stanford.edu/ElemStatLearn

Deep Learning (Adaptive Computation and Machine Learning series) by Goodfellow, Bengio, Courville

- Available on Amazon or at the GW Bookstore
- Online (free) at https://www.deeplearningbook.org/





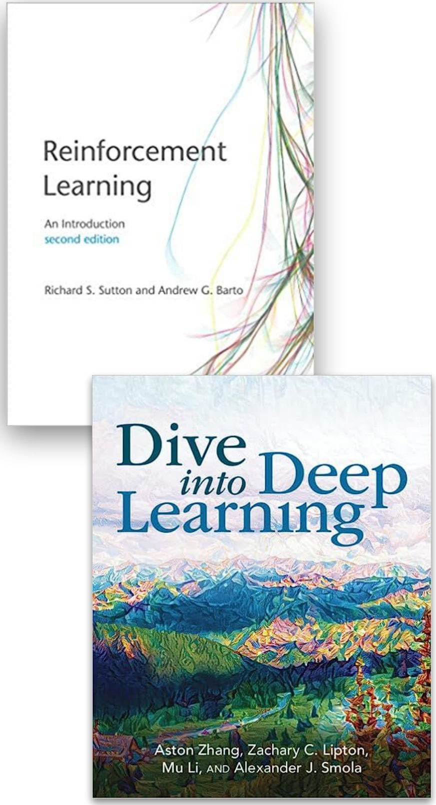
Required Textbooks

Reinforcement Learning, 2nd edition, by Sutton and Barto

- Available on Amazon as hard copy or eBook
- Online (free) at: http://incompleteideas.net/book/RLbook2020.pdf

Dive into Deep Learning

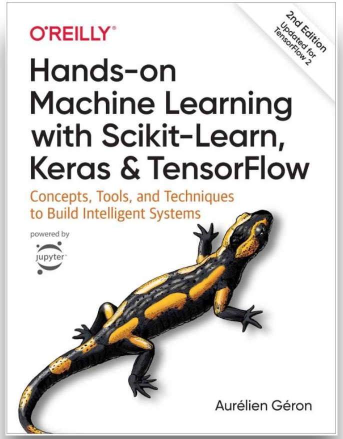
- Available on Amazon as paperback
- Online (free) at: https://d2l.ai/



Supplemental Textbooks

Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, by Aurélien Géron

- Available on Amazon as hard copy or eBook
- Online (free) at: https://www.knowledgeisle.com/wp-content/uploads/2019/12/2-Aur%C3%A9lien-G%C3%A9ron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O%E2%80%99Reilly-Media-2019.pdf



Homework Assignments

In this course we will use fundamental ML open-source ML Libraries

- Python 3
- Google Colaboratory Pro/Pro Plus (subscription info)
- Scikit-Learn
- Tensorflow 2
- Keras
- and possibly others open source tools…Tensor2Tensor, TF-Agents,...

Graded on execution (50%), correctness (30%), readability (20%)

Project Assignment

Apply ML to a unique, interesting problem in a novel way:

- **Applied**: Novel problem, familiar ML techniques
- **Research**: Familiar problem, novel ML techniques

Deliverables:

- Team Assignments: (Week 4)
- Project Proposal: (Week 5)
- Demo and Working Code in Colab Notebook (Week 17)
- Project Paper: Abstract, Intro, Related Work, Method, Results and Conclusion (Week 17)

Individual Project: For those interested pursuing novel research in ML

Team Project: For those interested in collaborating on an “industrial” application

Feedback from Last Class

The workload was very significant, more than many other classes combined…

This course was one of the best courses that I have taken during my studies.

Exams are one of the best I had as they are actually very fun and they focus on real-world scenarios rather than endless theories and mathematical equations that you would have to memorize.

The professor is amazingly knowledgeable. I also like how we covered a spectrum / broad survey of ML related concepts and techniques. The professor should definitely be asked to teach more classes in the future! I would take them all.

This is a hard course, please be prepared.

Introduction to Machine Learning

THE WALL STREET JOURNAL.

World | Business | U.S. | Politics | Economy | Tech | Finance | Opinion | Arts & Culture | Lifestyle | Real Es

TECH

The $900,000 AI Job Is Here

Salaries rise as employers such as Netflix and Walmart seek candidates with artificial-intelligence skills

By Chip Cutter Follow

Updated Aug. 14, 2023 11:46 am ET

Gift unlocked article | Listen (7 min)

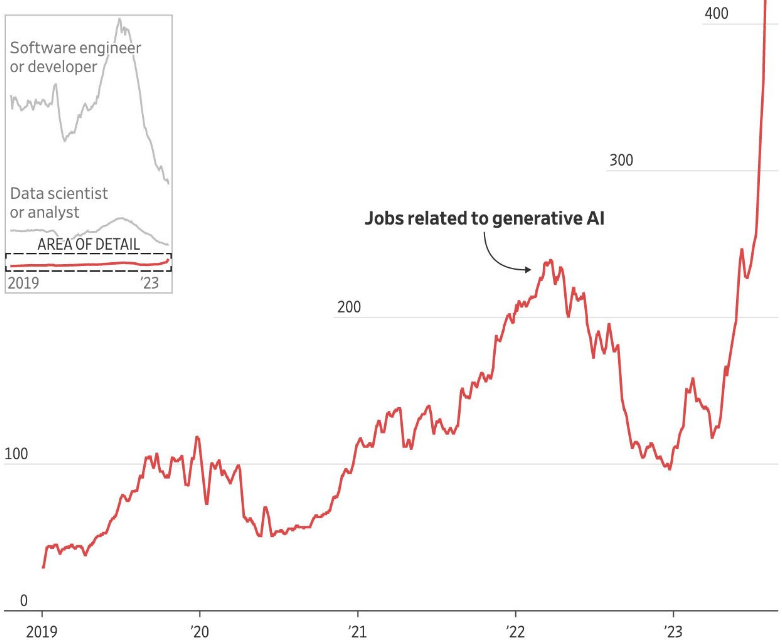
Netflix caused a stir when it listed an exceptionally high-paying job for its machine-learning platform.
PHOTO: BING GUAN/BLOOMBERG NEWS

American companies are in the midst of an AI recruiting frenzy, and some are willing to pay salaries approaching seven figures to hire top talent.

Firms in industries such as entertainment and manufacturing are racing to seize on the potential of artificial intelligence by wooing data scientists, machine-learning specialists and other practitioners skilled at deploying the technology.

The rush to hire is pushing up pay for technical professionals and prompting companies to hone their pitches to applicants to avoid losing out to rivals. Some companies, including Accenture, are building their AI expertise through individual hires and internal training programs. Others, including the technology company ServiceNow, say they are open to acquiring smaller AI startups as a way to scoop up talent.

Number of job listings per one million postings on Indeed



Note: U.S. only. As of Aug. 3.
Source: Indeed

Artificial Intelligence

"The science of making machines smart."

Demis Hassabis, Founder of DeepMind

Learning with Logic and Symbols

- **Expert Systems**

Learning by Observation and Exploration

- **Machine Learning**

Source: https://www.youtube.com/watch?v=zYfzux7JKHE

Learning with Logic and Symbols

Expert systems developed in the 1960-1990s expressed complex knowledge using rules (i.e., nested conditionals)

For any challenging tasks, Expert Systems failed to generalize adequately

Huge amount of manual effort generating rules

Consistency of rules also posed a huge challenge - Boolean satisfiability is NP Hard, and search space can grow exponentially

Prone to **Overfitting** and **Overgeneralization**

A Symbolics Lisp Machine: an early platform for expert systems

Source: https://en.wikipedia.org/wiki/Expert_system

Learning by Observation and Exploration

Natural intelligent species learn by observation and interacting with the environment.

Machine Learning broadly encompasses the concept of learning from historical data in order to make useful predictions on novel data.

Memorization of historic data leads to overfitting, and predictions fail on novel data.

Generalization is the process of learning robust decision boundaries that work on novel data.

“A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P improves with experience E.”

Tom Mitchell, (1997) Machine Learning (McGraw-Hill International Editions Computer Science Series)

Fundamental Concepts

Types of Machine Learning

Supervised Machine Learning

Semi-Supervised Machine Learning

Unsupervised Machine Learning

Reinforcement Learning

Example Data

Features

Examples or Instances

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

Source: https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_california_housing.html

Supervised ML

Learning with a *teacher*

Given labeled examples $\mathbf{X}$, $\boldsymbol{y}$, learn to predict y from a new, unlabeled instance $\boldsymbol{x}$

- **Classification:** $\boldsymbol{y}$ is categorical (e.g., spam, not spam)
- **Regression:** $\boldsymbol{y}$ is continuous

$\mathbf{X}$

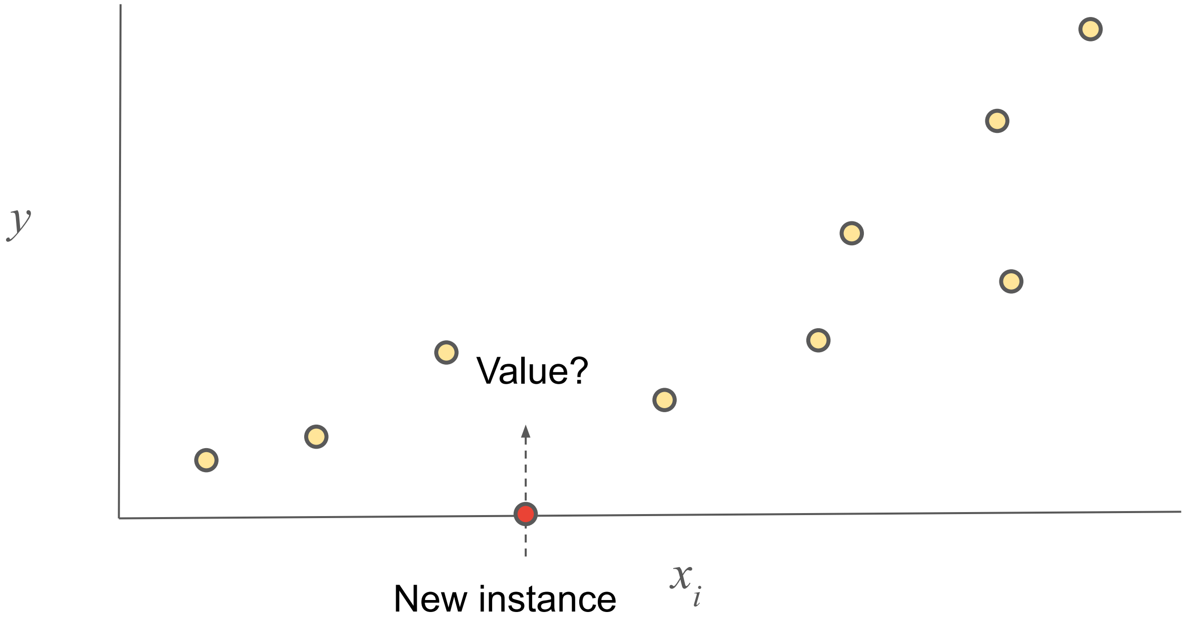
y

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

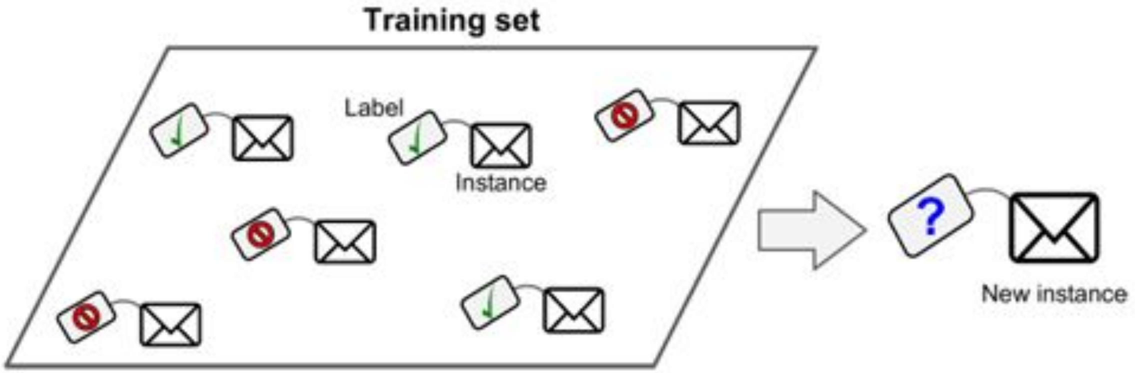
$\mathbf{x}_i$

y_i

Supervised Machine Learning (Regression)



Supervised Machine Learning (Classification)



Source Hands-On Machine Learning, Géron A.

Supervised Machine Learning

$\mathbf{x}$: Example with one or more features x_i

y: Label, Categorical C or Real value $\mathbb{R}^D$, $D \in \mathbb{Z}^+$

F: Prediction function or model

θ: Model Parameters

Regression:

$F(\mathbf{x}|\theta) = y$: Regression if $y \in \mathbb{R}^D$

Predict the median house value

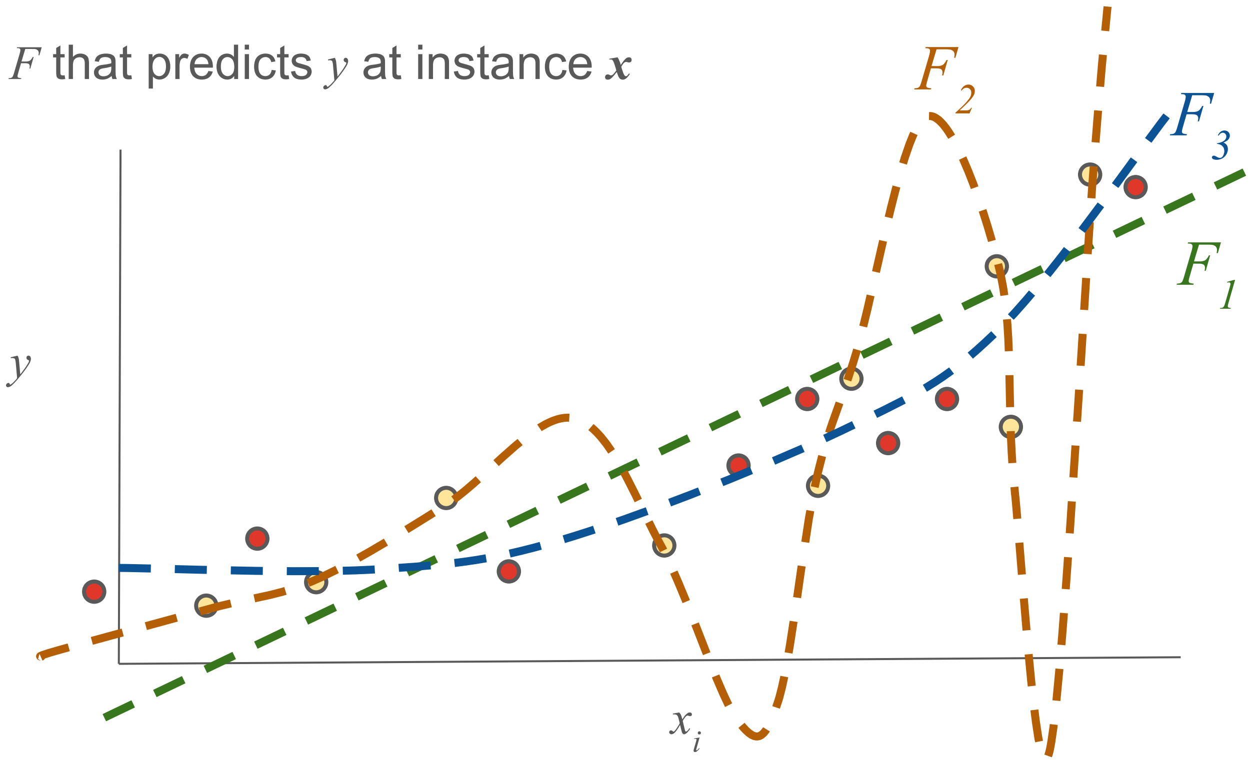
Classification:

$F(\mathbf{x}|\theta) = P(y|\mathbf{x})$ if $y \in \{C\}$

Predict "spam" or "not spam" class

Prediction

Find a function F that predicts y at instance $\boldsymbol{x}$



Generalization, not Memorization

A good hypothesis will generalize well (i.e., predict unseen examples correctly)

Ockham's razor: Prefer the simplest hypothesis that's consistent with the data

Supervised ML Algorithms

Linear Regression

Logistic Regression

Support Vector Machines

Decision Trees and Random Forests

Neural Networks

Supervised Machine Learning

Detecting tumors in brain scans: Image segmentation of labeling pixels

Detecting Credit Card Fraud: Location, amount, vendor

Value Prediction: Predicting the value of stocks and real estate

Email Spam Classification: Classifying an email as spam/not spam from text

Sequence-to-Sequence: Language translation and summarization

Unsupervised Machine Learning

Learning without a *teacher*

Groups. similarity, and distance

$\mathbf{X}$

$\mathbf{x}_i$

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

Unsupervised Machine Learning

Clustering:

- k-Means
- DBSCAN
- Hierarchical Clustering

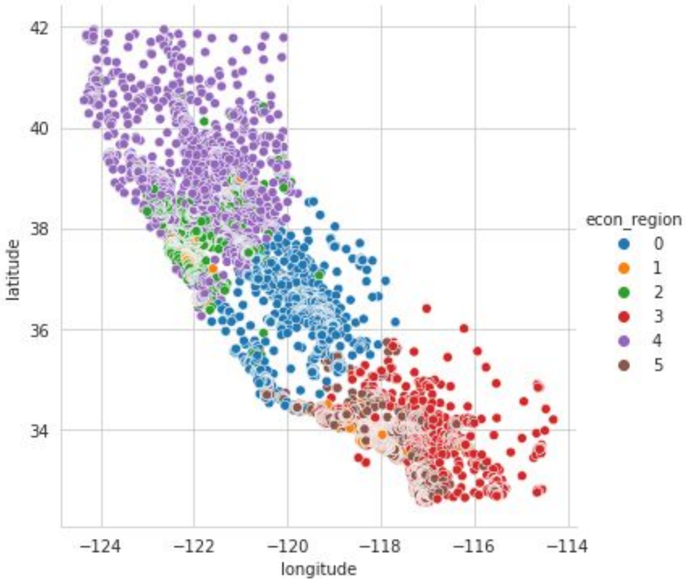
Anomaly (Novelty) Detection

- OC-SVM
- IsoForest
- MADI

Visualization and Dimensionality Reduction

- PCA
- t-SNE
- Autoencoders

k-Means clustering with $k = 6$



Source: https://www.kaggle.com/code/benherbertson/california-housing-k-means-clustering

Semi-Supervised Machine Learning

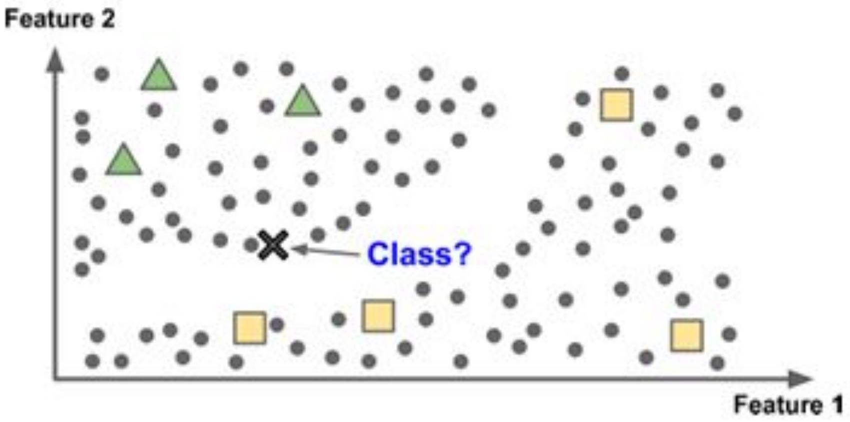
Getting lots of labeled data is time consuming and costly

Often we have many unlabeled instance

Hybrid approach → few labeled instances + many in labeled instances

Examples:

- Deep Belief Networks
- Restricted Boltzmann Machines



Reinforcement Learning

Reinforcement learning used to train Animals

Negative reinforcement

- Pain and hunger

Positive reinforcement

- Pleasure and food

Apply this idea to computers!

Reward signal - indicates how good an action is

Reinforcement Learning

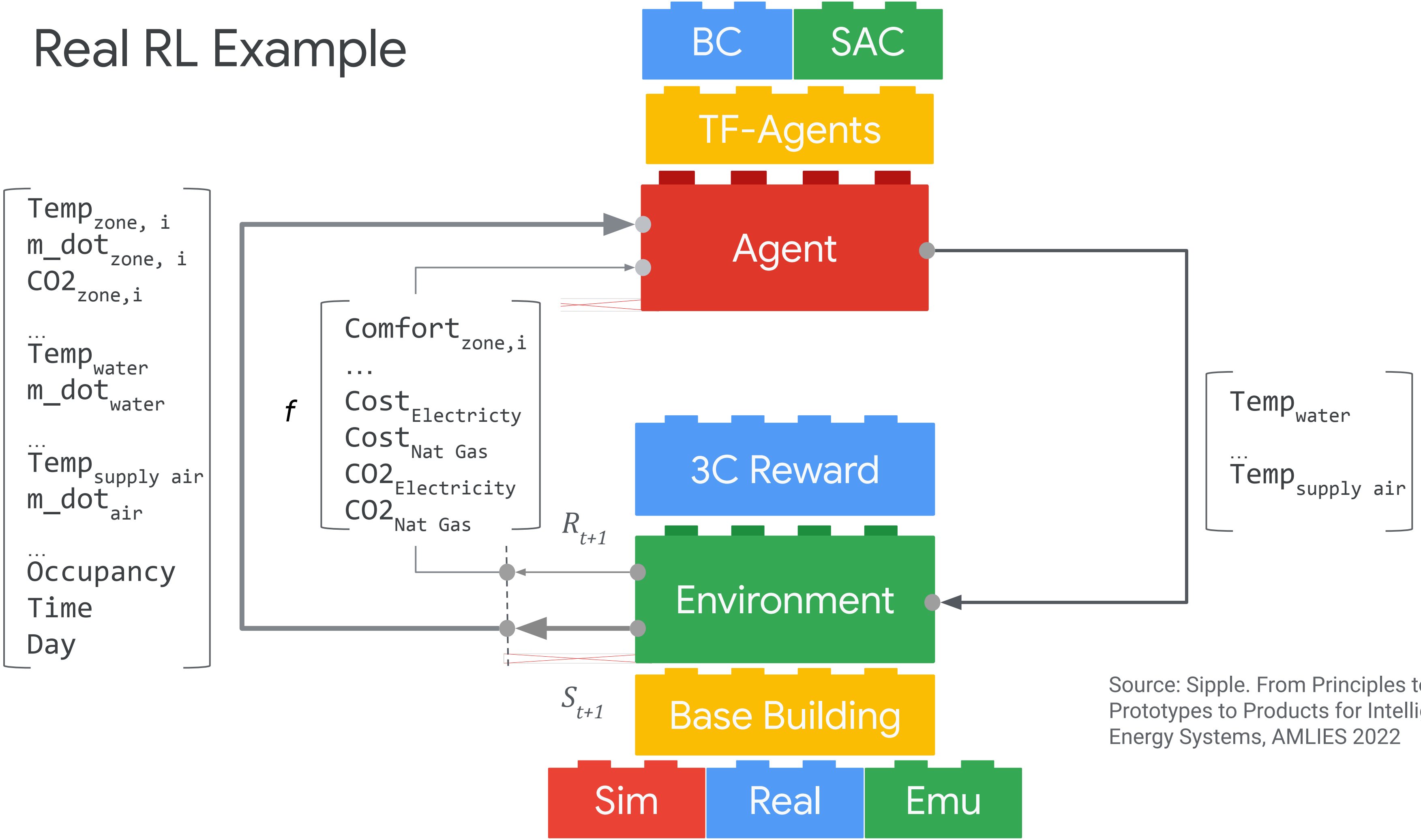
Environment

- An interactive game, universe with state
- Provides observations of the state
- Accepts actions may change the Environment's state
- Returns feedback, or reward

Agent

- Machine learning model with a policy
- At each step, gets observations of the current state, chooses an action
- Learns a policy to maximize long-term cumulative reward, choosing the best action given the current state

Real RL Example



Source: Sipple. From Principles to Prototypes to Products for Intelligent Energy Systems, AMLIES 2022

Google

RL Algorithms

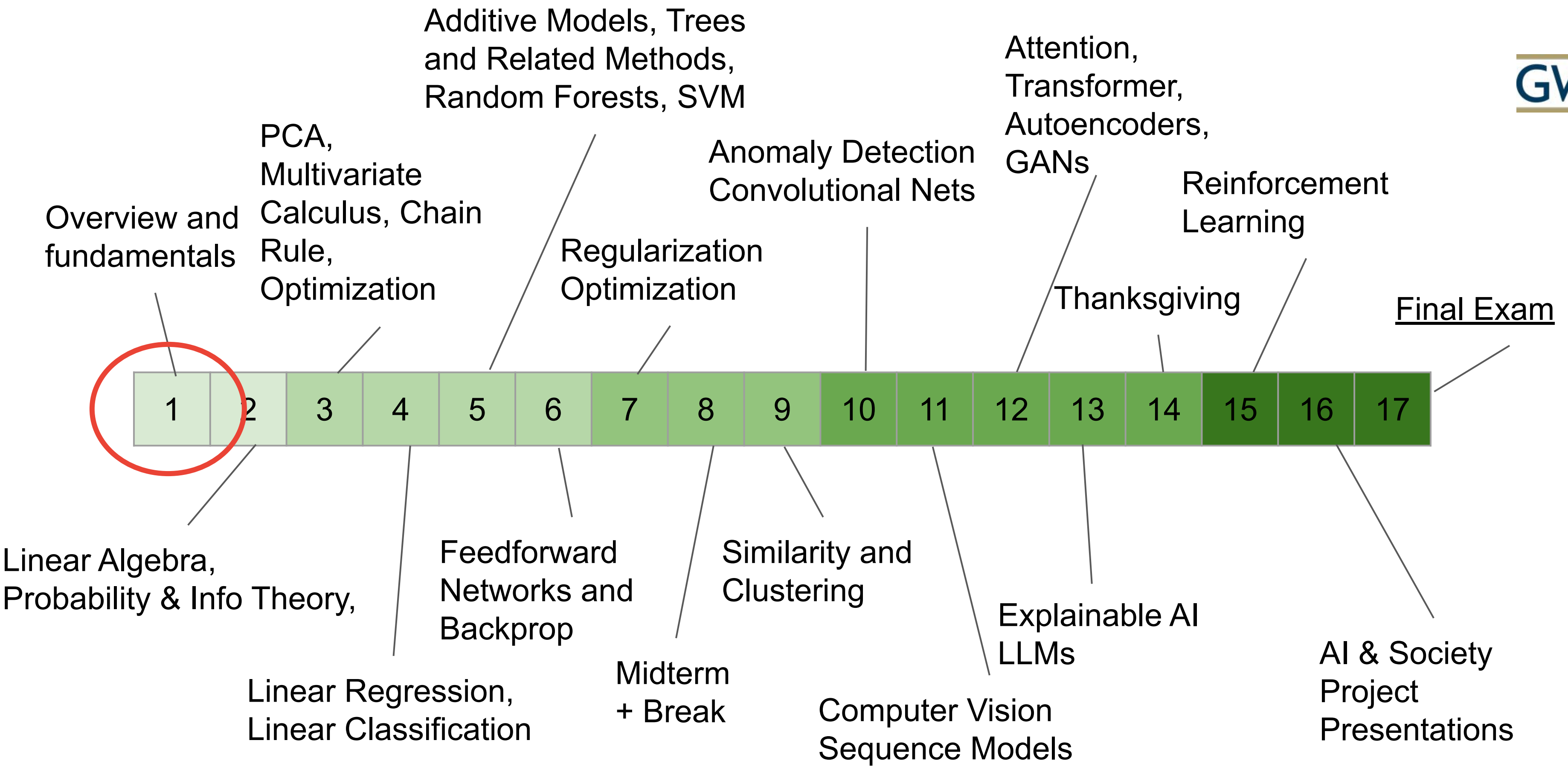
Model-Free RL relies on trial-and-error experience for setting up the optimal policy, associating the best action for any given state

- Q-Learning and value iteration
- Policy Gradient
- Actor-Critic

Model-Based RL models the environment and chooses an optimal policy based on the learned model

- Learn the model
- Given a Model

Course Roadmap


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Overview and fundamentals
Linear Algebra, Probability & Info Theory,
PCA, Multivariate Calculus, Chain Rule, Optimization
Linear Regression, Linear Classification
Additive Models, Trees and Related Methods, Random Forests, SVM
Feedforward Networks and Backprop
Regularization Optimization
Midterm + Break
Similarity and Clustering
Anomaly Detection Convolutional Nets
Computer Vision Sequence Models
Attention, Transformer, Autoencoders, GANs
Explainable AI LLMs
Thanksgiving
Reinforcement Learning
AI & Society Project Presentations
Final Exam

Homework 1

Homework 1

Due date: 9/12/2023

Familiarization with the environment:

- Python Language and Programming Style Guide
- ML Libraries: Tensorflow, Keras, Scikit-Learn
- Google Colaboratory Notebook
- Tensorboard

Training and Evaluating Binary Classifiers

- Cross-fold validation

Hyperparameter Tuning

Comparing Linear Regression against Neural Network

Readings

Readings

Goodfellow: Chapter 1 Introduction, 2 Linear Algebra

Questions?

Python Tutorial

https://docs.python.org/3/tutorial/

Intro to Colaboratory

Remote Runtime

Local Runtime https://research.google.com/colaboratory/local-runtimes.html

Colab Tutorial https://colab.research.google.com/?utm_source=scs-index

Pandas and Numpy

Pandas Tutorial
https://colab.research.google.com/notebooks/mlcc/intro_to_pandas.ipynb

Pandas, Numpy, Matplotlib Tutorial

https://colab.research.google.com/github/cs231n/cs231n.github.io/blob/master/python-colab.ipynb#scrollTo=9t1gKp9PL9eV

Hyperparameter Tuning

https://colab.research.google.com/github/tensorflow/tensorboard/blob/master/docs/hyperparameter_tuning_with_hparams.ipynb#scrollTo=mVtYvbbIWRkV