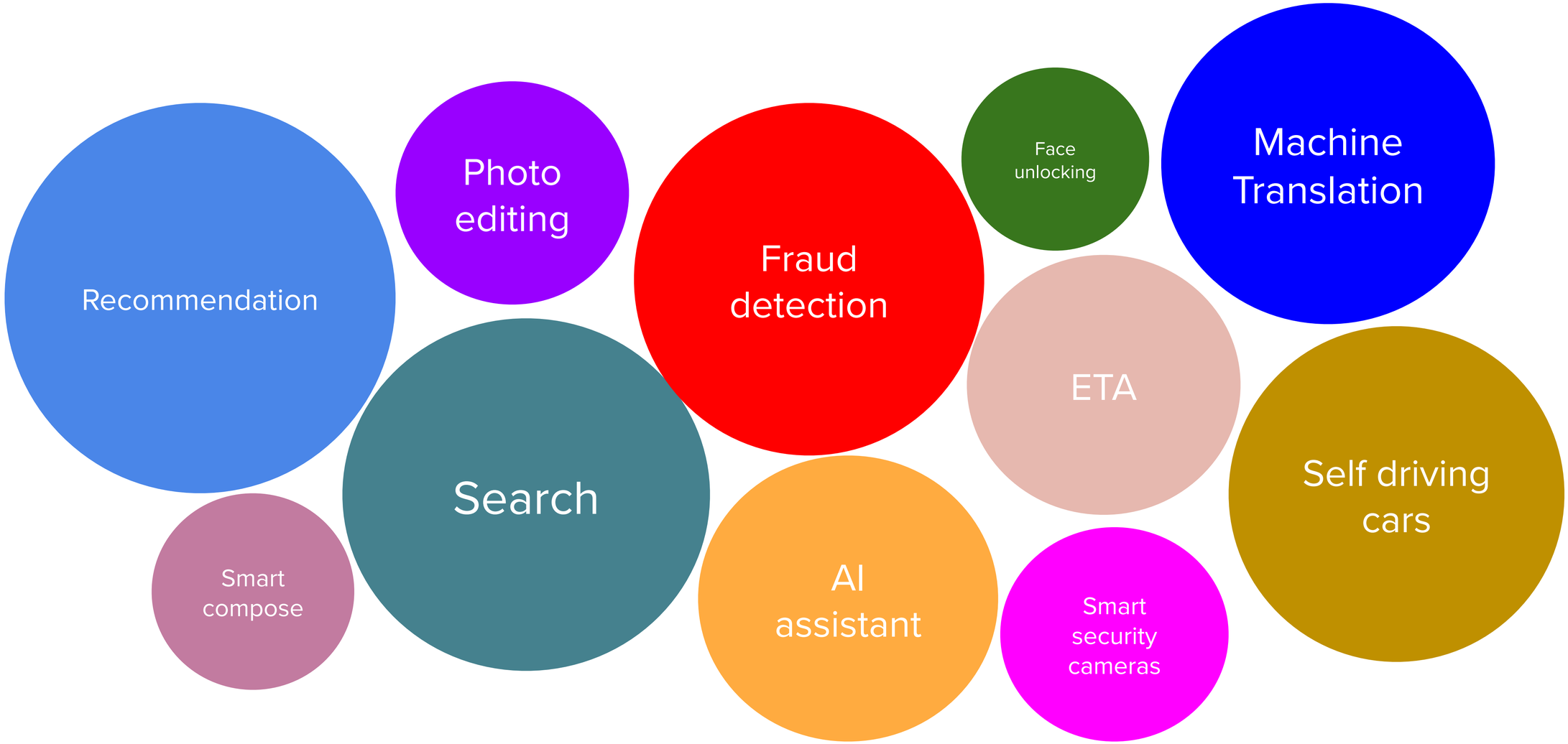
Machine Learning Systems

Lecture 1: Course Overview

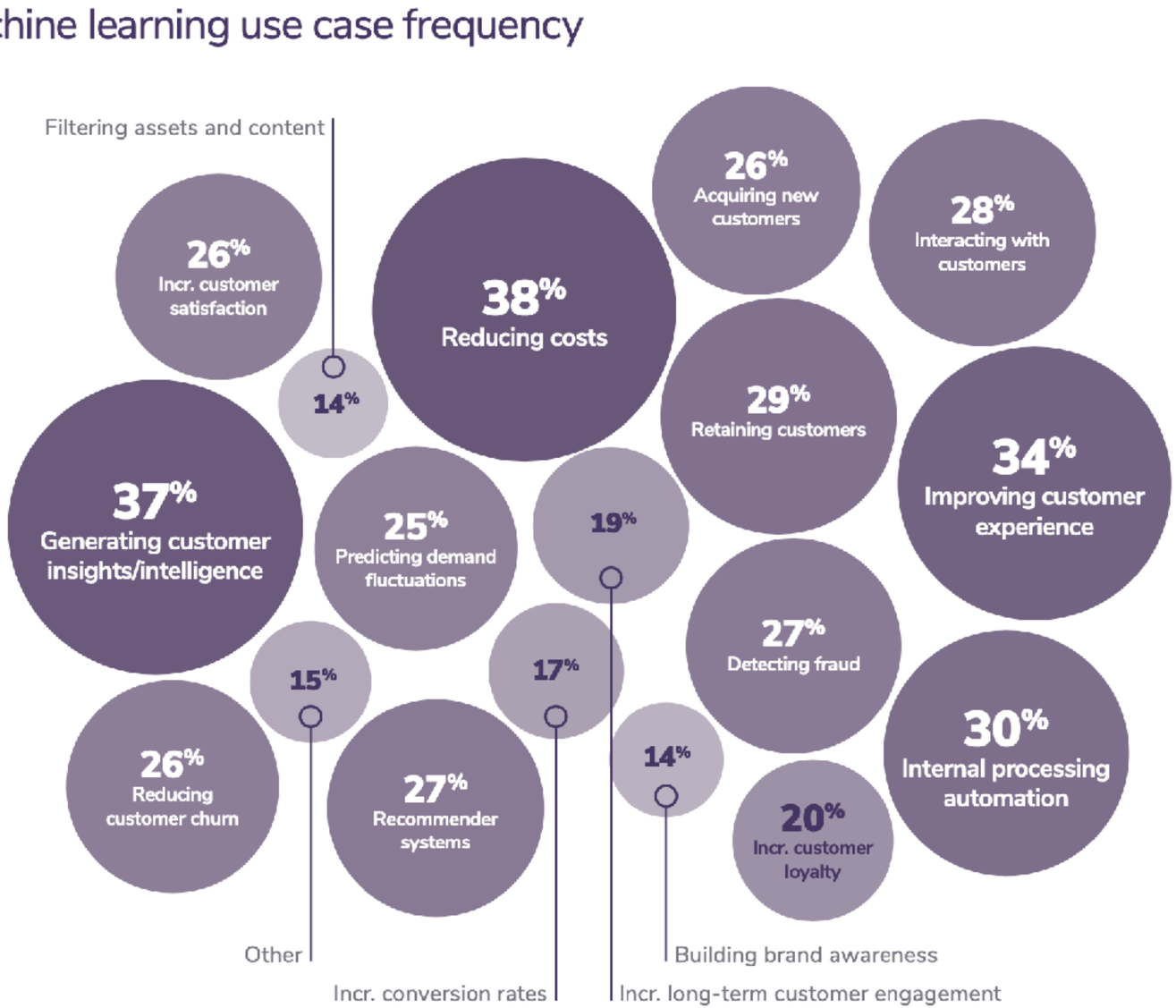
Pooyan Jamshidi

Course Overview

2022: ML is in almost every aspect of our lives



Enterprise use cases



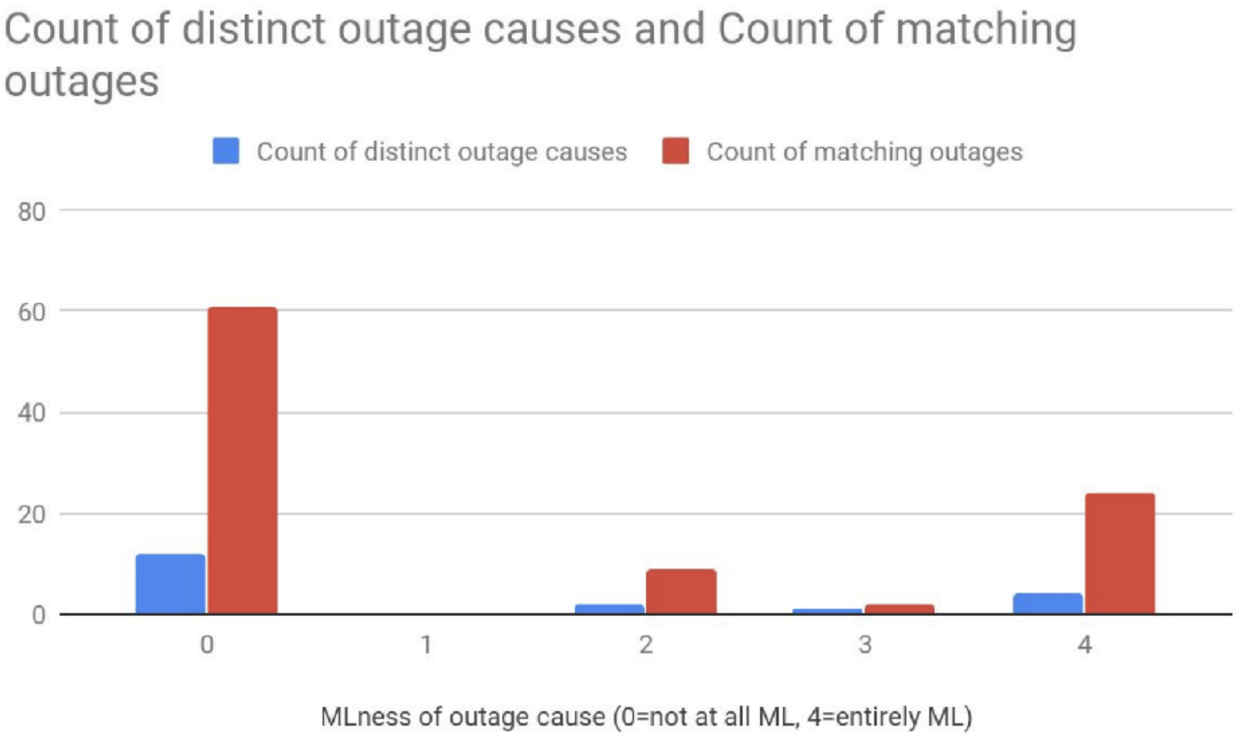
Why ML Systems instead of ML algorithms?

- ML algorithms is the less problematic part.
- The hard part is to **how to make algorithms work with other parts to solve real-world problems**.

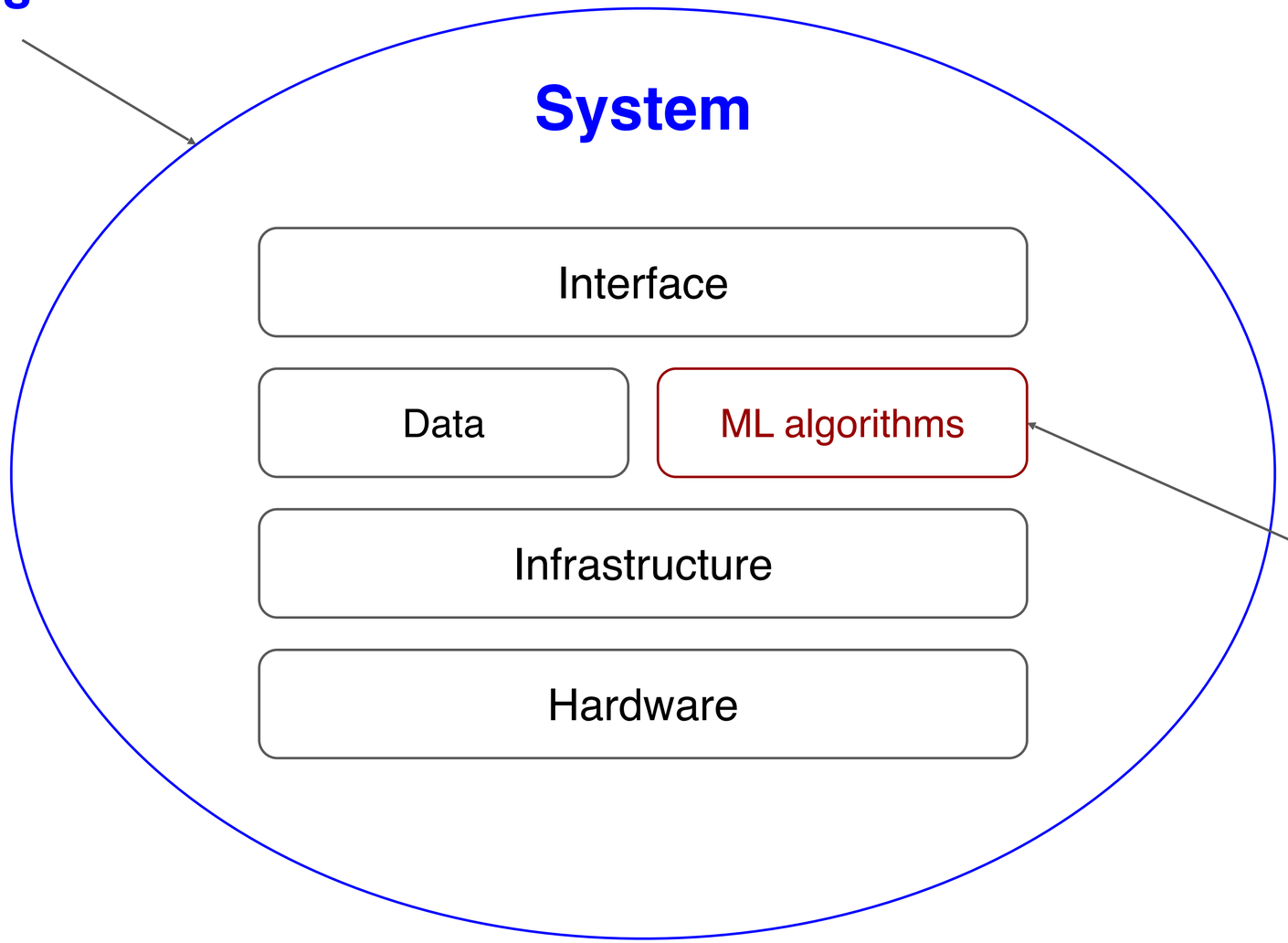
Why ML Systems instead of ML algorithms?

- ML algorithms is the less problematic part.
- The hard part is to **how to make algorithms work with other parts to solve real-world problems**.
- 60/96 failures caused by non-ML components

More on ML systems failures later!

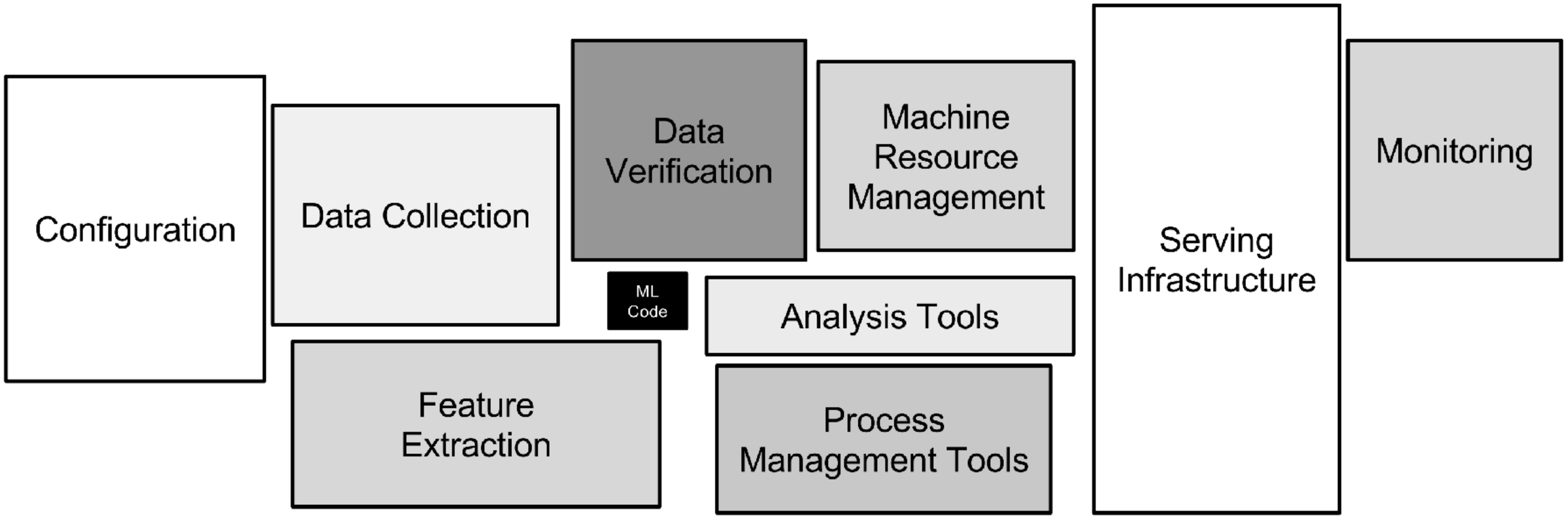


CSCE 585:
ML Systems



Most ML courses/ books

ML in Production



What is machine learning systems design?

The process of defining the **interface, algorithms, data**, **infrastructure**, and **hardware** for a machine learning system to satisfy **specified requirements**.

What is machine learning systems design?

The process of defining the **interface, algorithms, data**, **infrastructure**, and **hardware** for a machine learning system to satisfy **specified requirements**.

reliable, scalable, maintainable, adaptable

The questions this class will help answer …

- You've trained a model, now what?
- What are different components of an ML system?
- How to do data engineering?
- How to engineer features?
- How to evaluate your models, both offline and online?
- What's the difference between online prediction and batch prediction?
- How to serve a model on the cloud? On the edge?
- How to continually monitor and deploy changes to ML systems?
- …

This class will cover ...

- ML production in the real world from software, hardware, and business perspectives
- Iterative process for building ML systems at scale
  - project scoping, data management, developing, deploying, monitoring & maintenance, infrastructure & hardware, business analysis
- Challenges and solutions of ML engineering

This class will not teach ...

- Machine learning/deep learning algorithms
  - Machine Learning
  - Deep Learning
  - Convolutional Neural Networks for Visual Recognition
  - Natural Language Processing with Deep Learning
- Computer systems
  - Principles of Computer Systems
  - Operating systems design and implementation
- UX design
  - Introduction to Human-Computer Interaction Design
  - Designing Machine Learning: A Multidisciplinary Approach

Machine learning: expectation

This class won't teach you how to do this

Machine learning: reality

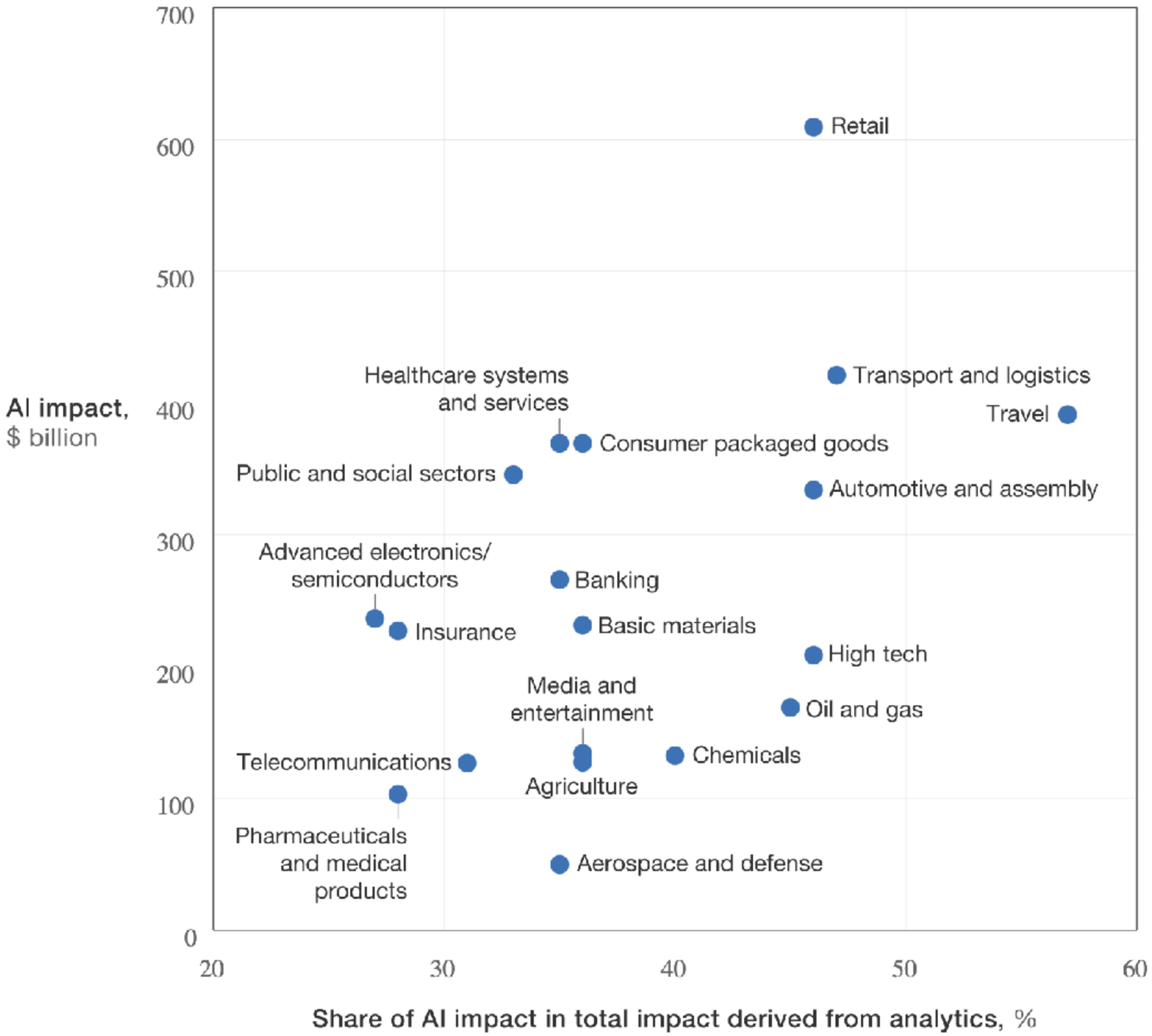
You'll likely build something like this (buggy but cool)

Prerequisites

- Knowledge of CS principles and skills
- Understanding of ML algorithms
- Familiar with at least one framework such as TensorFlow, PyTorch, JAX
- Familiarity with basic probability theory.

You will be fine and would take away lots of good things if you are eager to learn :)

Artificial intelligence (AI) has the potential to create value across sectors.



AI value creation by 2030

13 trillion USD

Most of it will be outside the consumer internet industry

We need more people from non-CS background in AI!

McKinsey&Company | **Source:** McKinsey Global Institute analysis

Evaluations

ML Systems course is project-based

- Build an ML-powered application
- Must work in groups of 2-3
- Demo + report (creative formats encouraged)

See course website: https://pooyanjamshidi.github.io/mls/

Grading

- **Project Proposal:** 10%
- **Milestones 1, 2, 3:** each 10 %
- **Final Deliverables:** 40%
- **Project Demonstration:** 20%

Grading

- A [90 – 100]
- B+ [86 – 90)
- B [75 – 86)
- C+ [70 – 75)
- C [60 – 70)
- D+ [55 – 60)
- D [40 – 55)
- F [0 – 40)

Office hours

- When? TR 13 pm – 14 pm
- Where? Innovation Building 2212
- What if this time does not work for me? Please drop me an email and I will find a time to meet with you.

Discussions

- Piazza: you have already been added!
- Ask questions
- Answer others’ questions
- Learn from others’ questions and answers
- Find teammates

Project Proposal

- What is the **problem** that you will be investigating? Why is it interesting?
- What **reading** will you examine to provide context and background?
- What **data** will you use? If you are collecting new data, how will you do it?
- What **method** or algorithm are you proposing? If there are existing implementations, will you use them and how? How do you plan to improve or modify such implementations? You don't have to have an exact answer at this point, but you should have a general sense of how you will approach the problem you are working on.
- How will you **evaluate** your results? Qualitatively, what kind of results do you expect (e.g. plots or figures)? Quantitatively, what kind of analysis will you use to evaluate and/or compare your results (e.g. what performance metrics or statistical tests)?

How projects will be evaluated

- You can work in teams of up to 2 or 3 people.
- Every team member should be able to demonstrate her/his contribution(s)
- The outcome will be evaluated based on the quality of the deliverables (code, results, report) and presentations/ demonstrations.
- The final report is an iPython notebook with documentation, results, comparisons, discussions, and related work.

Honor code: permissive but strict - don’t test us ;)

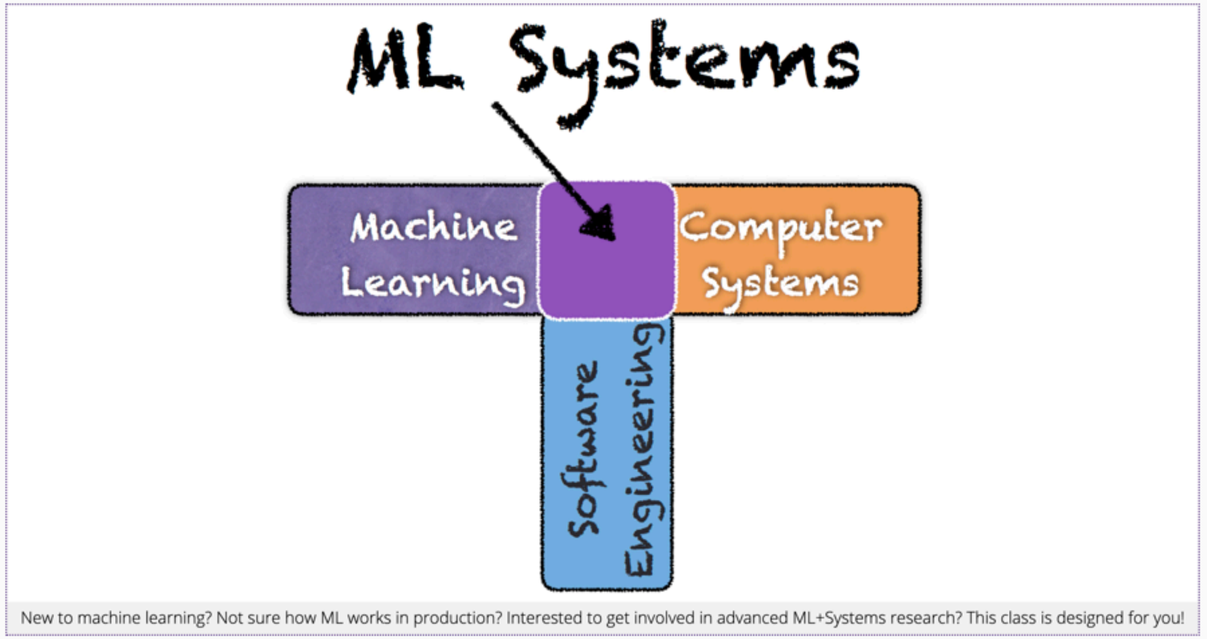
- OK to search about the systems we’re studying.
- Cite all the resources you reference.
  - E.g., if you read it in a paper, cite it.
- NOT OK to ask someone to do assignments/projects for you.
- OK to discuss questions with classmates. Disclose your discussion partners.
- NOT OK to copy solutions from classmates.
- OK to use existing solutions as part of your projects/assignments. Clarify your contributions.
- NOT OK to pretend that someone’s solution is yours.
- OK to publish your final project after the course is over (we encourage that!)
- NOT OK to post your assignment solutions online.
- **ASK the course instructor if unsure!**

Important Dates

- Project proposal: due September 6.
- Project milestone 1: due September 29.
- Project milestone 2: due October 20.
- Project milestone 3: due November 10.
- Final report and all deliverables: due December 2.

CSCE 585 LECTURES PROJECTS RESOURCES POLICIES PIAZZA

Machine Learning Systems



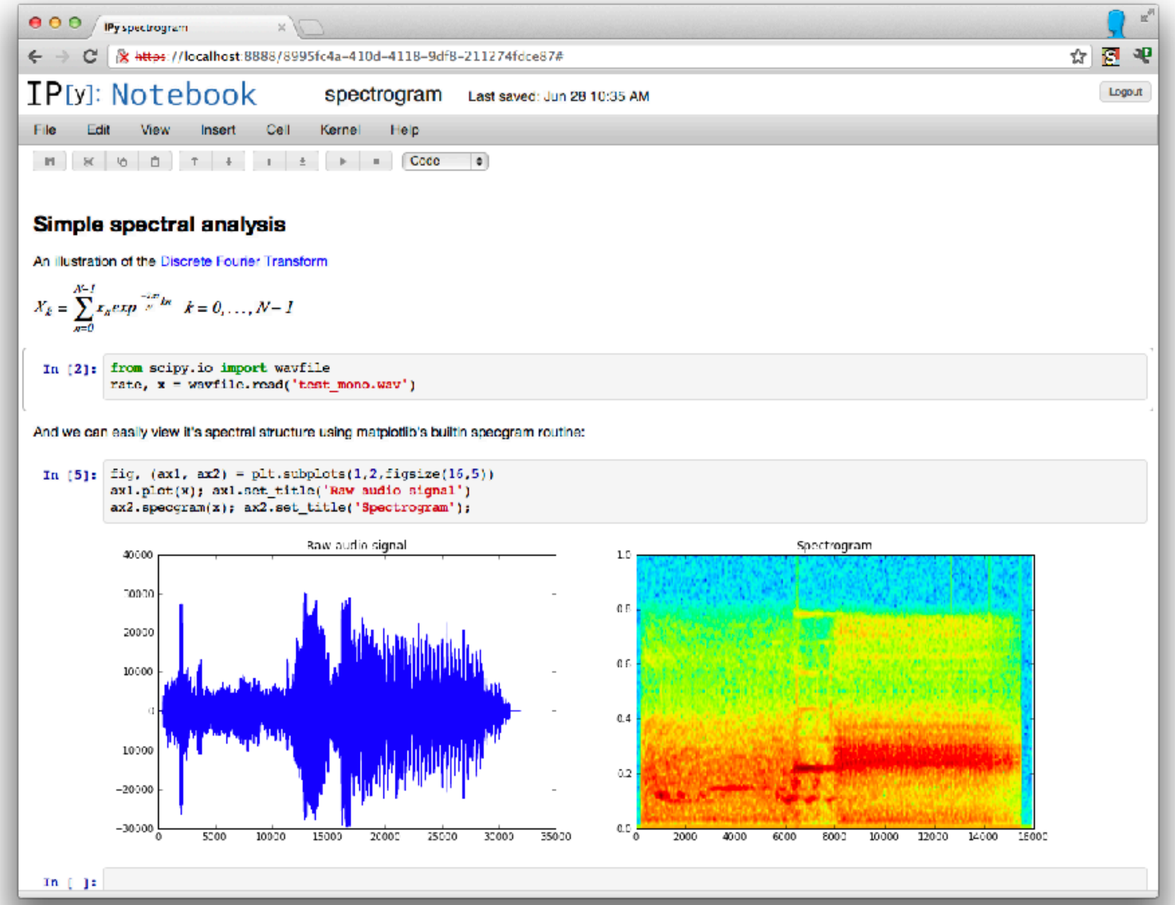
New to machine learning? Not sure how ML works in production? Interested to get involved in advanced ML+Systems research? This class is designed for you!

When we talk about Artificial Intelligence (AI) or Machine Learning (ML), we typically refer to a technique, a model, or an algorithm that gives the computer systems the ability to learn and to reason with data. However, there is a lot more to ML than just implementing an algorithm or a technique. In this course, we will learn the fundamental differences between AI/ML as a model versus AI/ML as a system in production.

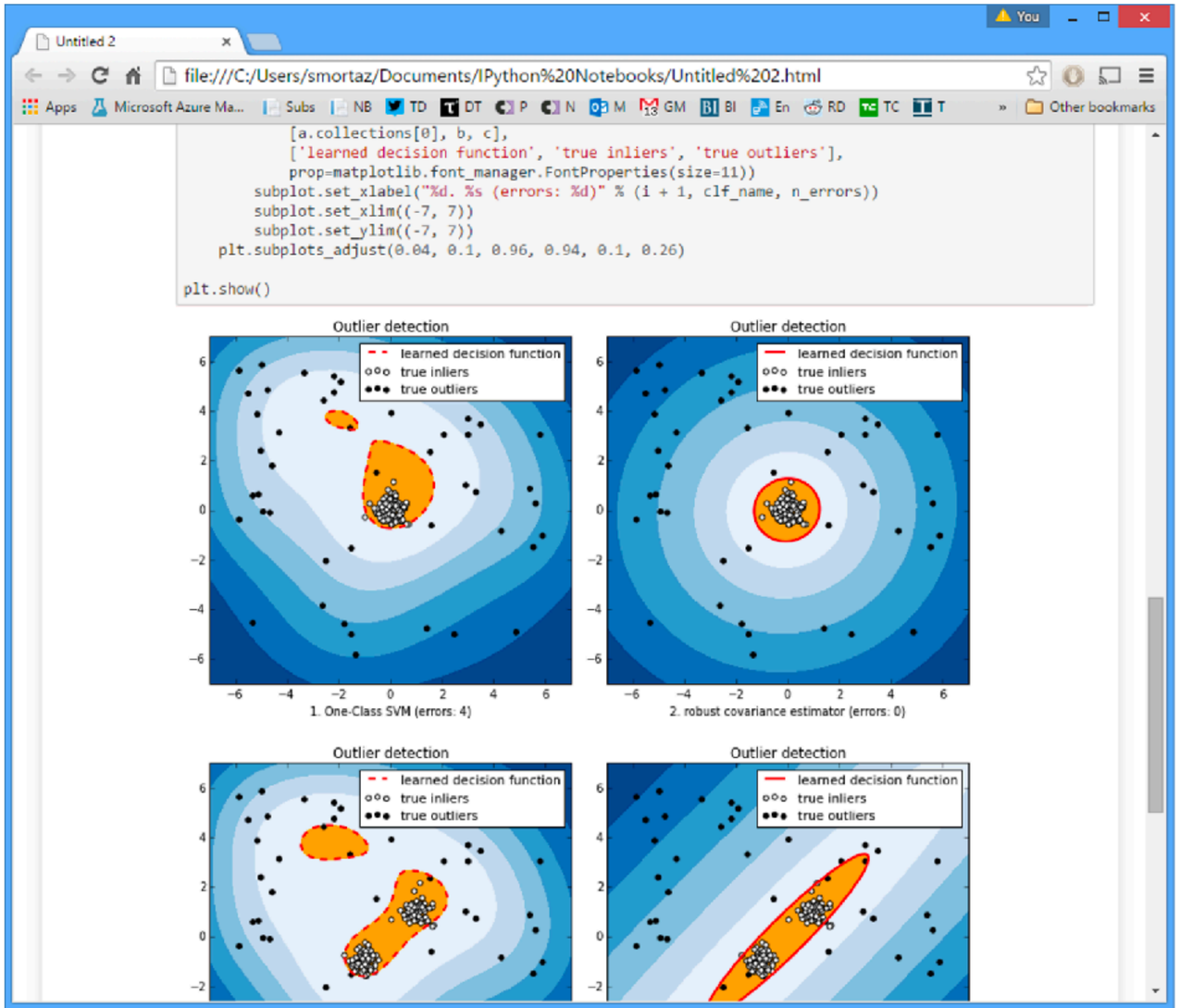
https://pooyanjamshidi.github.io/mls/

Examples, Tips, Suggestions

How the project report should looks like?



How the project report should looks like?



How the project report should looks like?

IP[y]: Notebook GDP_CO2_Example Last saved: Feb 26 12:33 PM

File Edit View Insert Cell Kernel Help

Code

Andy Wilson has a nice more tightly intergration of d3.js and ipython notebook. See https://github.com/wilsaj/ipython/tree/d3plots

some other examples (mostly experimental, need various different setups.)

https://github.com/csohn/IPython-Notebook---d3.js-mashup

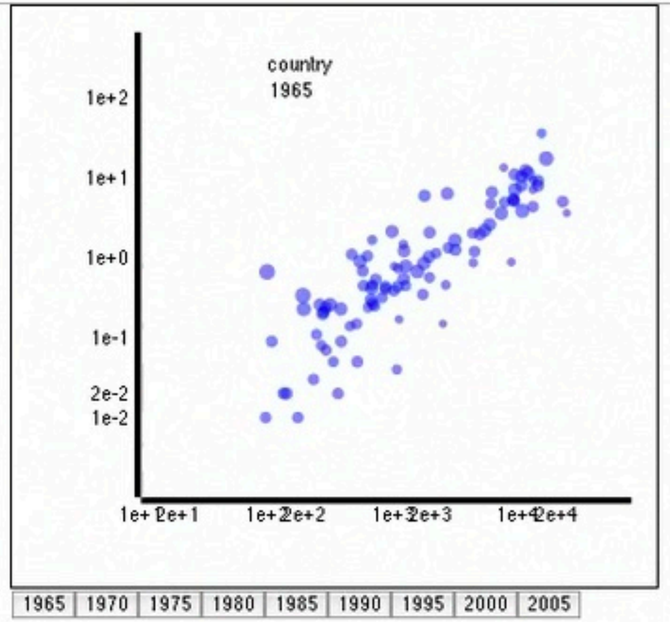
Why?

The whole exercise here is mostly on exploring the possibility to have really dynamic frontend for developing visulizations or demostrations. The ipython notebook provides a really nice way to integrate web technologies with the powerful backend python processes. This will make dynamic data exploratory work with python easier in the furture using mostly open-source software. We can eventually integrate a lots of other cool web technologies (e.g. webGL, html5 video, canvas) together.

What's next

In this example, I use bare-bone python functions / javascript functions for the work. I think the reasonable next step is to see what is the right kind of framework for mapping the javascript objects and python objects (e.g. something like https://github.com/mikedewar/d3py for ipython notebook or Andy Wilson's d3plots approach.) Eventually, we may develope a standard set of widgets or integrate some concept of the "Grammar of Graphics" (http://www.amazon.com/Grammar-Graphics-Leland-Wilkinson/dp/0387987746) and ggplot2-like features (http://had.co.nz/ggplot2/) as python notebook libraries.

--Jason Chin, Feb 26, 2012



In [35]:
```
# Here we show we can re-define the function and have the javascript calls
# the re-defined function immediately
# The code below plots the circles using the sizes propotional to the log o
# population of each country
```

Design think it a lil

- Have each member of your team flesh out 20 quick ideas down on paper before meeting. Don’t be afraid to get creative
- Filter out list by doing quick Google searches on data a. Anything below GB scale of data...good luck. Vision = big datasets b. If you have an idea, Google it first! Don’t want to “just” reproduce the same result. There’s probably a Github with your project already
- Pay attention to how long and much data the models you see are trained on
- Find pattern in data+architecture combos
- Ask are there little tweaks or other experiments that haven’t been done yet?
- Can you extend the idea in one paper with another?
- Which idea gives you more things to experiment with? 8. How can you get pretty images / figures?

Try to avoid

- Nothing special in data pipeline. Uses prepackaged source
  Team starts late. Just instance and draft of code up by milestone
- Explore 3 architectures with code that already exists a. One RES-net, then a VGG, and then some slightly different thing
- Only ran models until they got ~65% accuracy 5. Didn’t hyperparameter search much
- A few standard graphs: loss curves, accuracy chart, simple architecture graphic
- Conclusion doesn’t have much to say about the task besides that it didn’t work

Aim for this

- Setup your workflow
- Have running code and have baseline model running and fully-trained
- Formulate hypotheses and write down how you are going to test them
- Mixing knowledge from different aspects in ML
- Have a meaningful graphic (pretty or info-rich)
- Conclusion and Results teach me something
- ++ interactive demo
- ++ novel / impressive engineering features
- ++ Surprising results

Milestone Goals

- We want to see you have code up and running
- Data source explained correctly a. Give the true train/test/val split b. Number training examples c. Where you got the data
- What Github repo, or other code you’re basing off of
- Ran baseline model have results a. Points off for no model running, no results
- Data pipeline should be in place
- Brief discussion of initial, preliminary results
- Reasonable literature review (3+ sources)
- 1-2 page progress report. Not super formal

ML Systems Project Ideas

Checkout

Projects

Coursework Template

Please use this template for submitting your written assignments.

Topics

The course project is an opportunity to apply what you have learned in class to a problem of your interest. Potential projects must have these two components:

- `Machine Learning` algorithm: Any ML model class including neural networks or any good old-fashioned ML/AI.
- `Computer Systems` : The project should have at least one computer systems component: (i) Platform: Embedded, Realtime, Cloud, IoT, Edge; (ii) Systems issues such as scalability, performance, reliability; (iii) On-device ML: e.g., TinyML, AI on Edge; (iv) Trustworthy AI: Bias, Fairness, Robustness, Privacy, Security, Explainability, Interpretability, Interoperability; (v) Robot Learning, any project that makes robots more intelligent!

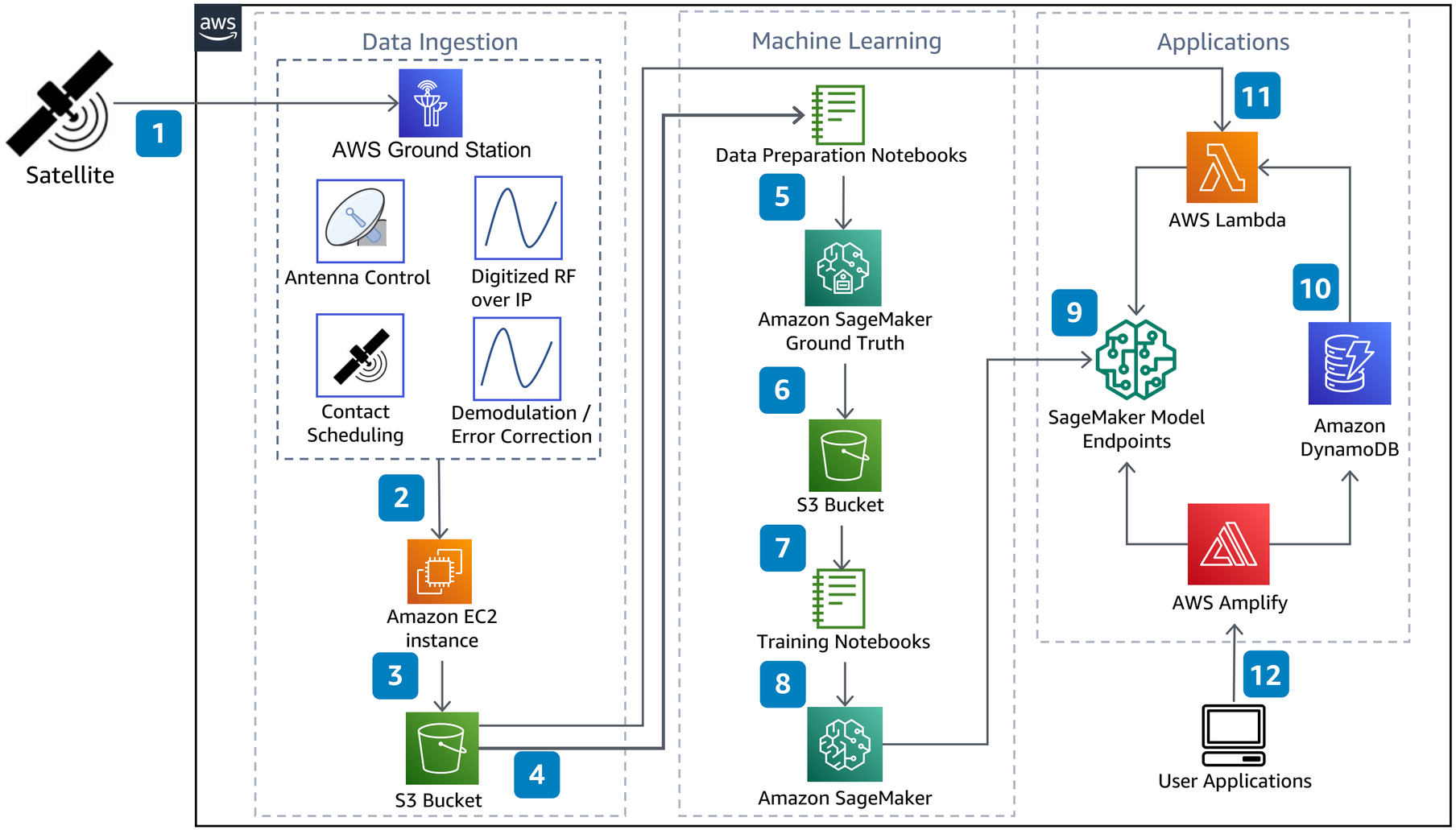
The following categories also fit within the scope, and I highly encourage students to consider such projects:

- Building on top of an existing ML system that you can find on GitHub, e.g., you can develop a tracking algorithm and develop a plugin for DeepStream
- AI competitions and challenge problems, these are great project ideas for non-CS students, e.g., EvalAI hosts many interesting competitions with prizes suitable for students from all backgrounds, e.g., (i) Open Catalyst Challenge for Chemical Engineering students; (ii) Neural Latents Benchmark for Neuroscience students; (iii) The Robotic Vision Challenges for students interested in robotics.
- Hackathons, e.g., PyTorch Annual Hackathon 2021, AWS BugBust, Kaggle
- Systematic study of open source ML Systems via (i) interview study (please make sure you design the interview study correctly before conducting the interviews) and/or (ii) Formulating interesting research questions about building ML systems, for example, contrasting testing practices for ML systems vs. traditional software systems, collecting data from software repositories, and systematically extracting info from these repositories that answers your research questions.
- TinyML projects, you can find many ideas on GitHub and TinyML community forum
- And, in general, any interesting ML systems project ideas, try GitHub, Reddit

If you are unsure whether the project you have defined fits within the scope, please talk with me after class. If you believe that might be helpful for other students, please ask your question on Piazza or during class hours.

Run Machine Learning Algorithms with Satellite Data

Use AWS Ground Station to ingest satellite imagery, and use Amazon SageMaker to label image data, train a machine learning model, and deploy inferences to customer applications.



1. Satellite sends data and imagery to the **AWS Ground Station** antenna.
2. **AWS Ground Station** delivers baseband or digitized RF-over-IP data to an **Amazon EC2** instance.
3. The **Amazon EC2** instance receives and processes the data, and then stores the data in an **Amazon S3** bucket.
4. A Jupyter Notebook ingests data from the **Amazon S3** bucket to prepare the data for training.
5. **Amazon SageMaker Ground Truth** labels the images.
6. The labeled images are stored in the **Amazon S3** bucket.
7. The Jupyter Notebook hosts the training algorithm and code.
8. **Amazon SageMaker** runs the training algorithm on the data and trains the machine learning (ML) model.
9. **Amazon SageMaker** deploys the ML models to an endpoint.
10. The SageMaker ML model processes image data and stores the generated inferences and metadata in **Amazon DynamoDB**.
11. Image data received into **Amazon S3** automatically triggers an **AWS Lambda** function to run machine learning services on the image data.

master ▾ | 2 branches | 0 tags | Go to file | Add file ▾ | Code ▾

MENG2010 Update README.md — d150afd on Dec 15, 2020 — **161** commits

data	Update README.md	11 months ago
documents	Create README.md	11 months ago
models	updated models/svm/README.md	10 months ago
notebooks	Merge remote-tracking branch 'origin/master'	10 months ago
src	updated models/svm/README.md	10 months ago
.gitignore	update gitignore	9 months ago
LICENSE	Initial commit	11 months ago
README.md	Update README.md	8 months ago
environment.yml	new environment file	11 months ago
requirements.txt	Update requirements.txt	10 months ago

README.md

Project ATHENA

This is the course project for CSCE585. Students will build their machine learning systems based on the provided infrastructure --- Athena.

Overview

This project assignment is a group assignment. Each group of students will design and build an adversarial machine learning system on top of the provided framework (ATHENA) then evaluate their work accordingly. The project will be evaluated on a benchmark dataset MNIST. This project will focus on supervised machine learning tasks, in which all the training data are labeled. Moreover, we consider only evasion attacks in this project, which happens at the test phase (i.e., the targeted model has been trained and deployed).

Each team should finish three tasks independently --- two core adversarial machine learning tasks and a competition task.

About

This is the course project for CSCE585: ML Systems. Students will build their machine learning systems based on the provided infrastructure --- Athena.

adversarial-machine-learning
adversarial-example
adversarial-attacks
machine-learning-systems
adversarial-defense

Readme

MIT License

Contributors 3

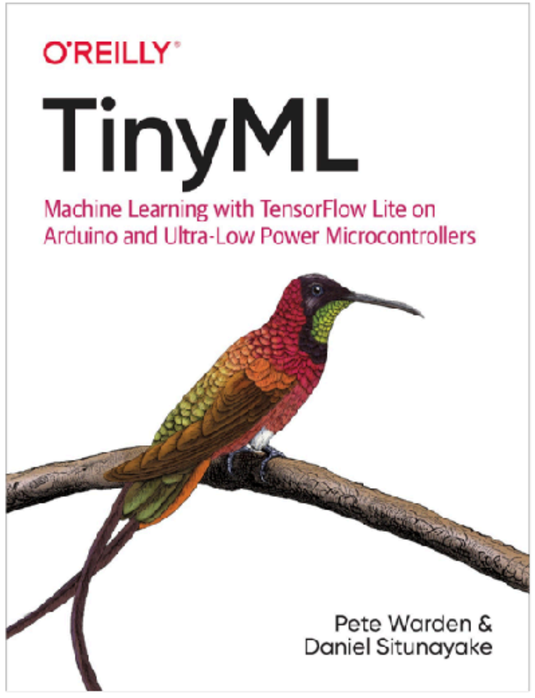
MENG2010 MENG

pooyanjamshidi Pooyan Jamshidi

Kronemeyer

Languages

- **Python** 61.6%
- **Jupyter Notebook** 38.4%



Reference Book

We recommend Pete's TinyML book as a reference for the projects and programming assignments. The book is a good primer for anyone new to embedded devices and machine learning. It serves as a good starting point for understanding the machine learning workflow, starting from data collection to training a model that is good enough for deploying on ultra-low power computing devices.

The course builds on top of some concepts covered within this book. We are also preparing an e-book that is a good primer to fill-in material that is supplementary to this book. Stay tuned!

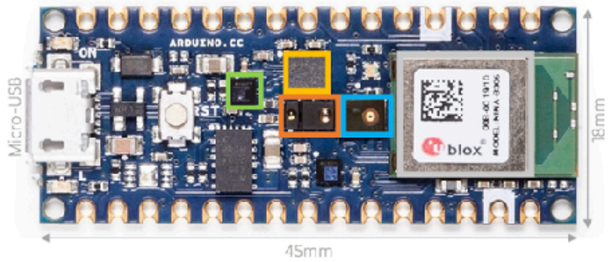
Coding Assignments

To get everyone familiar with coding on embedded systems with ML, we will be using the examples provided in this book as a starting point. Each assignment will build on the examples provided.

Projects

The course will culminate with project demos! You will have an opportunity to showcase what you have learned by incorporating your experience into a hands-on project of your liking. Alternatively, we will provide a list of suggested projects that will allow you to start from the class assignments.

Development Platforms



- Color, brightness, proximity and gesture sensor
- Digital microphone

Cortex-M4 Microcontroller

You will learn to run your ML models on a Nordic nrf52840 processor (256KB RAM, 1 MB Flash, 64 MHz) on the Arduino Nano 33 BLE Sense platform.

TensorFlow Lite

TensorFlow Lite (Micro)

You will use TF Lite (Micro) to deploy your ML models, which is offered free of cost by Google.

Other project ideas

- Focusing on one aspect of ML Systems like testing, deployment, explainability, etc.
- You can work with a company (interview, etc) for documenting their ML practices, then writing a report to be submitted to a conference or a workshop
- Mining software repositories for ML Systems practices (with a central hypothesis)

Learning Materials

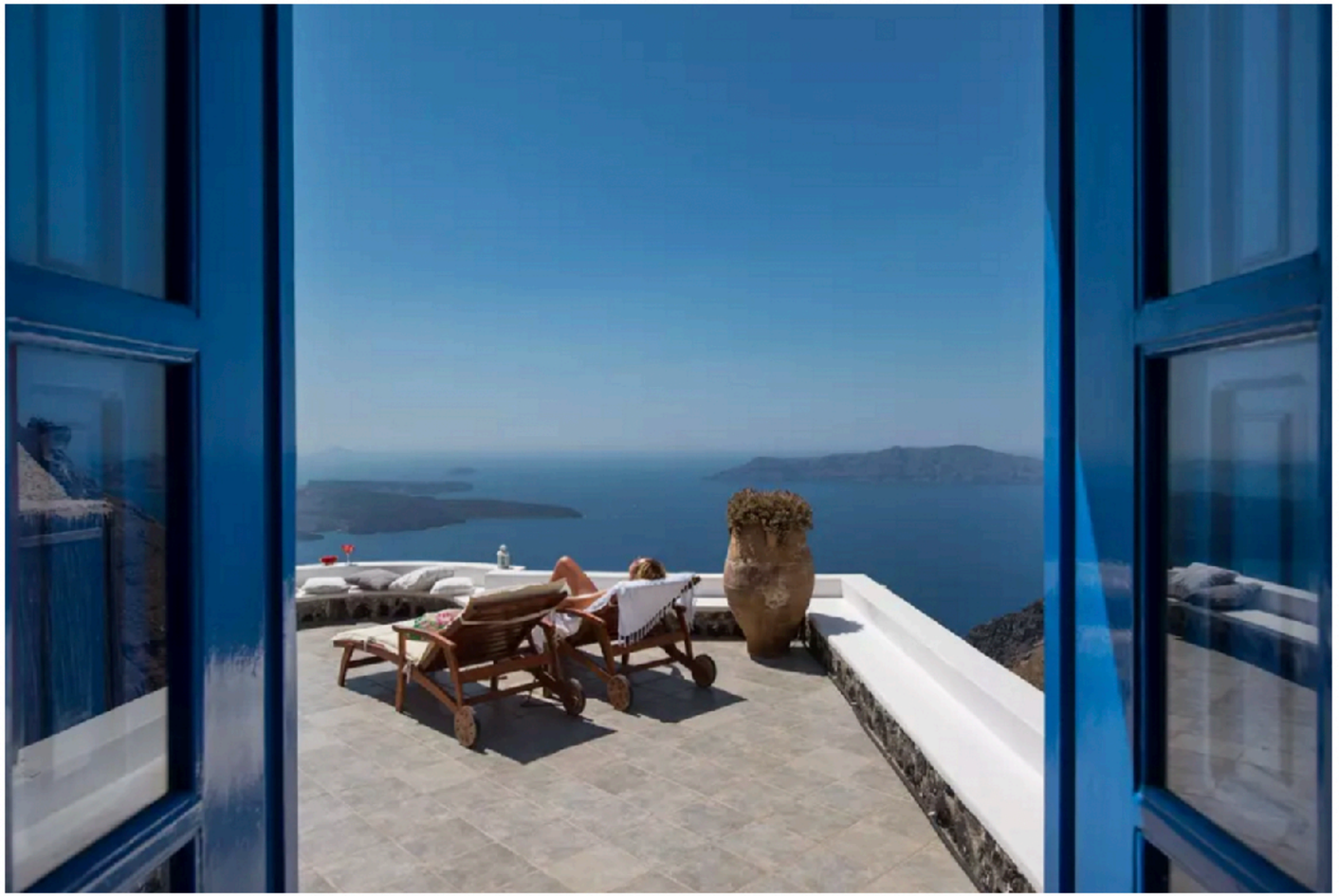
Learning Materials

- Learn one of these frameworks: TensorFlow, PyTorch, JAX
  - There are many good tutorials for each framework on their website
- Try to build some very simple models with available benchmarks
  - Search for the LeNet model and train it with the MNIST dataset
  - Try to feed some input data and get the prediction and print it on the console
  - Then try to measure basic performance metrics such as Accuracy or Inference time

Case I

Using Machine Learning to Predict Value of Homes On Airbnb

by Robert Chang



Amazing view from a Airbnb Home in Imerovigli, Egeo, Greece

Using Machine Learning to Predict Value of Homes On Airbnb

- Airbnb used machine learning to predict a vital business metric: the value of homes on Airbnb.
- It walks you through the entire workflow: feature engineering, model selection, prototyping, and moving prototypes to production.
- It’s completed with lessons learned, tools used, and code snippets too.

Case II

Netflix Technology Blog

Mar 22, 2018 · 7 min read · Listen

Using Machine Learning to Improve Streaming Quality at Netflix

by Chaitanya Ekanadham

One of the common questions we get asked is: “Why do we need machine learning to improve streaming quality?” This is a really important question, especially given the recent hype around machine learning and AI which can lead to instances where we have a “solution in search of a problem.” In this blog post, we describe some of the technical challenges we face for video streaming at Netflix and how statistical models and machine learning techniques can help overcome these challenges.

Using Machine Learning to Improve Streaming Quality at Netflix

- As of 2018, Netflix streams to over 117M members worldwide, half of those living outside the US.
- This blog post describes some of their technical challenges and how they use machine learning to overcome these challenges, including:
  - predicting the network quality,
  - detect device anomaly,
  - and allocate resources for predictive caching.

Case III

150 successful machine learning models: 6 lessons learned at Booking.com

OCTOBER 7, 2019 ~ ADRIAN COLYER

150 successful machine learning models: 6 lessons learned at Booking.com
Bernadi et al., *KDD'19*

Here's a paper that will reward careful study for many organisations. We've previously looked at the deep penetration of machine learning models in the product stacks of leading companies, and also some of the pre-requisites for being successful with it. Today's paper choice is a wonderful summary of lessons learned integrating around 150 successful customer facing applications of machine learning at Booking.com. Oddly enough given the paper title, the six lessons are never explicitly listed or enumerated in the body of the paper, but they can be inferred from the division into sections. My interpretation of them is as follows:

150 Successful Machine Learning Models: 6 Lessons Learned at booking.com

- As of 2019, Booking.com has around 150 machine learning models in production.
  - Predicting users' travel preferences and how many people they travel with
  - Optimizing the background images and reviews to show for each user.
- Lessons Learned:
  - Machine-learned models deliver strong business value.
  - Model performance is not the same as business performance.
  - Be clear about the problem you're trying to solve.
  - Prediction serving latency matters.
  - Get early feedback on model quality.
  - Test the business impact of your models using randomized controlled trials.