Managing NLP Experiments

Intro lecture

17.04.2025

Table of contents

01

Course Goal

and some small goals

The main goal of this course is to help you to get a nice job.

Some other goals

Helping with obligatory classes

As a side effect, this course should help with other classes (mainly SNLP 2).

Helping with the projects for other classes

Treating ML projects from an end-to-end point of view will help you to make projects for other classes.

Helps you to create a portfolio of projects

02

How will we ~~try to~~ achieve our goals?

Our plan

01 Orientation on end-to-end ML

02 Versatility

03 Interview-driven approach

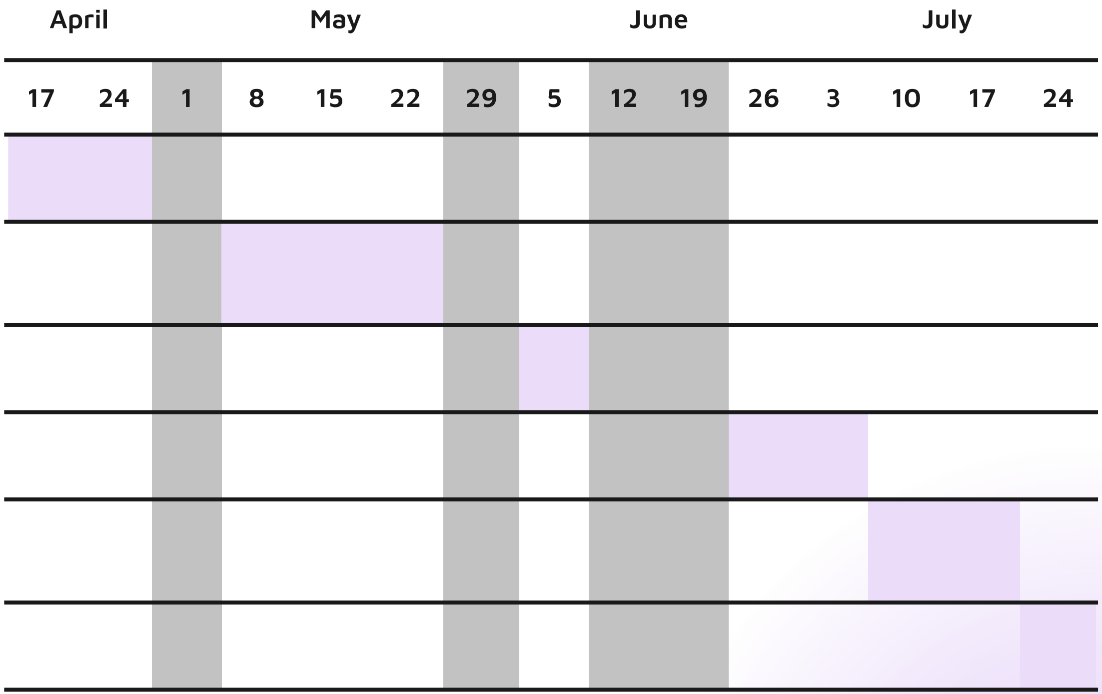
04 Work towards strong project portfolio

05 Course readjustment and targeting

03

Course Structure and Topics

Course Plan



Course Plan

		April		May					June				July			
Module	**Topics**	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24
Introduction	Course Overview, ML Project Design															

Module 0. Introduction

Course Overview

- Overview of the entire course structure.
- Grading.
- Course workflow.
- Comparison of Machine Learning in research versus in production.
- Machine learning use cases.

ML Project Life Cycle

- Requirements for ML Systems.
- Machine learning life cycle and its stages.
- Different tasks in the cycle.
- Skills and knowledge required for solving those tasks.
- Explore various machine learning roles, including ML engineer, applied scientist, data scientist, and other positions.

Course Plan

		April		May					June				July			
Module	**Topics**	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24
Introduction	Course Overview, ML Project Design															
Training Data	Sampling, Labeling, Class Imbalance															

Module I. Training Data

Sampling

- Nonprobability Sampling.
- Simple Random Sampling.
- Weighted Sampling.
- Reservoir Sampling.
- Importance Sampling.

Labeling

- Hand Labels (label multiplicity, data lineage).
- Natural Labels.
- Handling the Lack of Labels (weak supervision, semi-supervision, transfer learning, active learning)

Class Imbalance

- Challenges of Class Imbalance
- Handling Class Imbalance (using right evaluation metrics, resampling, algorithm-level methods)

Course Plan

Module	Topics	April		May					June				July			
		17	24	1	8	15	22	29	5	12	19	26	3	10	17	24
Introduction	Course Overview, ML Project Design															
Training Data	Sampling, Labeling, Class Imbalance															
Data Analysis, Data Visualization	Exploratory Data Analysis, Data Visualization															

Module II. Data Analysis

Exploratory Data Analysis

- Understanding different data types
- Descriptive statistics
- Identifying patterns and relationships (outliers, trends, dependencies)
- Data profiling tools

Data Visualization

- Basics of data visualization
- Most common techniques (histograms, box plots, scatter plots, correlation heatmaps)

Course Plan

		April		May					June				July			
Module	**Topics**	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24
Introduction	Course Overview, ML Project Design															
Training Data	Sampling, Labeling, Class Imbalance															
Data Analysis, Data Visualization	Exploratory Data Analysis, Data Visualization															
Feature Engineering	Common Feature Engineering Operations, Data Leakage, Engineering Good Features															

Module III. Feature Engineering

Common Feature Engineering Operations

- Handling Missing Values (deletion, imputation).
- Scaling.
- Discretization.
- Encoding Categorical Features.
- Feature Crossing.
- Discrete and Continuous Positional Embeddings.

Data Leakage

- Common Causes for Data Leakage (time-correlated data, scaling, filling missing values, group leakage, data generation)
- Detecting Data Leakage

Engineering Good Features

- Feature Importance
- Feature Generalization

Course Plan

		April		May					June				July			
Module	**Topics**	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24
Introduction	Course Overview, ML Project Design															
Training Data	Sampling, Labeling, Class Imbalance															
Data Analysis, Data Visualization	Exploratory Data Analysis, Data Visualization															
Feature Engineering	Common Feature Engineering Operations, Data Leakage, Engineering Good Features															
Model Development, Offline Evaluation	Model Development, Offline Evaluation															
Exam																

Module IV

Model Development

- Evaluating ML Models
- Ensembles
- Experiment Tracking and Versioning
- Distributed Training
- AutoML

Offline Evaluation

- Baselines
- Evaluation Methods

Course Plan

		April		May					June				July			
Module	**Topics**	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24
Introduction	Course Overview, ML Project Design															
Training Data	Sampling, Labeling, Class Imbalance															
Data Analysis, Data Visualization	Exploratory Data Analysis, Data Visualization															
Feature Engineering	Common Feature Engineering Operations, Data Leakage, Engineering Good Features															
Model Development, Offline Evaluation	Model Development, Offline Evaluation															
Exam																

Grading

Exam

- Exam is obligatory.
- The exam will be based on theoretical questions which you will be getting for every individual topic.
- You get 3 CP.
- You entire grade for the course is based on the exam.

Project

- You can get extra 3 CP.
- It is a programing project based on the skills which you learn in this course.
- You will have to write a short paper about it (4-5 pages) in latex.
- You grade for the course will be based both on the exam (1/2) and on the project (1/2).

04
Course Workflow

Course Workflow

Materials

- Questions about the topic.
- Topic cheat sheet.
- Notes about the lecture in Markdown (strongly recommended to use Obsidian)
- Lecture slides.
- Practical examples.

Lecture

- Topic presentation.
- Practical showcases.

05

ML in Research Versus in Production

	Research	Production
Requirements	State-of-the-art model performance on benchmark datasets	Different stakeholders have different requirements
Computational priority	Fast training	Fast inference, low latency
Data	Static	Constantly shifting
Fairness	Often not a focus	Must be considered
Interpretability	Often not a focus	Must be considered

Case

Mobile app that recommends restaurants to users:

- Makes money by charging restaurants a 10% service fee.
- The project involves ML engineers, salespeople, product managers, infrastructure engineers, and a manager.

Different stakeholders

ML engineers

- Want a model that recommends restaurants that users will most likely order from.
- Believe they can do so by using a more complex model with more data.

Different stakeholders

ML engineers

- Want a model that recommends restaurants that users will most likely order from.
- Believe they can do so by using a more complex model with more data.

Sales team

- Wants a model that recommends the more expensive restaurants since these restaurants bring in more service fees.

Different stakeholders

ML engineers

- Want a model that recommends restaurants that users will most likely order from.
- Believe they can do so by using a more complex model with more data.

Product team

- Notices that every increase in latency leads to a drop in orders through the service.
- They want a model that can return the recommended restaurants in less than 100 milliseconds.

Sales team

- Wants a model that recommends the more expensive restaurants since these restaurants bring in more service fees.

Different stakeholders

ML engineers

- Want a model that recommends restaurants that users will most likely order from.
- Believe they can do so by using a more complex model with more data.

Sales team

- Wants a model that recommends the more expensive restaurants since these restaurants bring in more service fees.

Product team

- Notices that every increase in latency leads to a drop in orders through the service.
- They want a model that can return the recommended restaurants in less than 100 milliseconds.

ML platform team

- As the traffic grows, this team has been woken up in the middle of the night because of problems with scaling their existing system.
- They want to hold off on model updates to prioritize improving the ML platform.

Different stakeholders

ML engineers

- Want a model that recommends restaurants that users will most likely order from.
- Believe they can do so by using a more complex model with more data.

Sales team

- Wants a model that recommends the more expensive restaurants since these restaurants bring in more service fees.

Product team

- Notices that every increase in latency leads to a drop in orders through the service.
- They want a model that can return the recommended restaurants in less than 100 milliseconds.

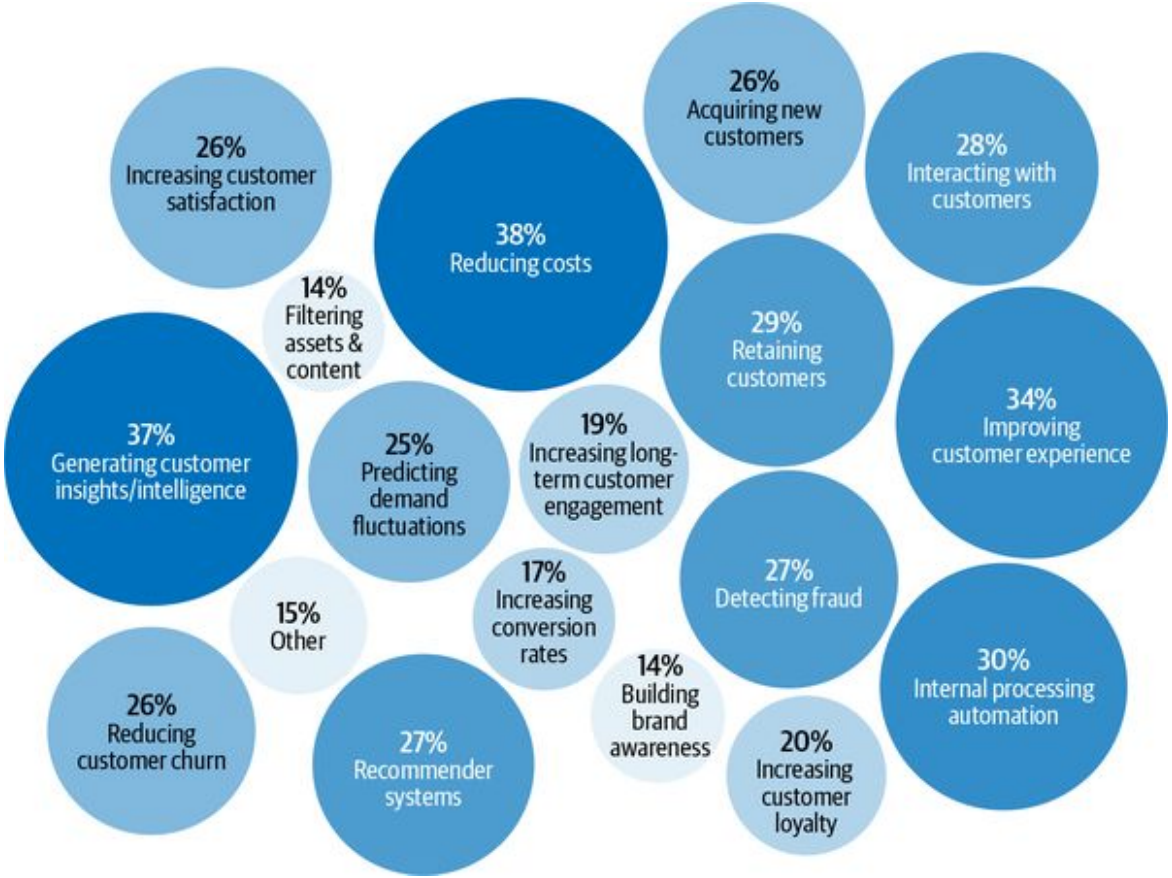
Manager

- Wants to maximize the margin, and one way to achieve this might be to let go of the ML team.

ML platform team

- As the traffic grows, this team has been woken up in the middle of the night because of problems with scaling their existing system.
- They want to hold off on model updates to prioritize improving the ML platform.

ML Use Cases in the enterprise

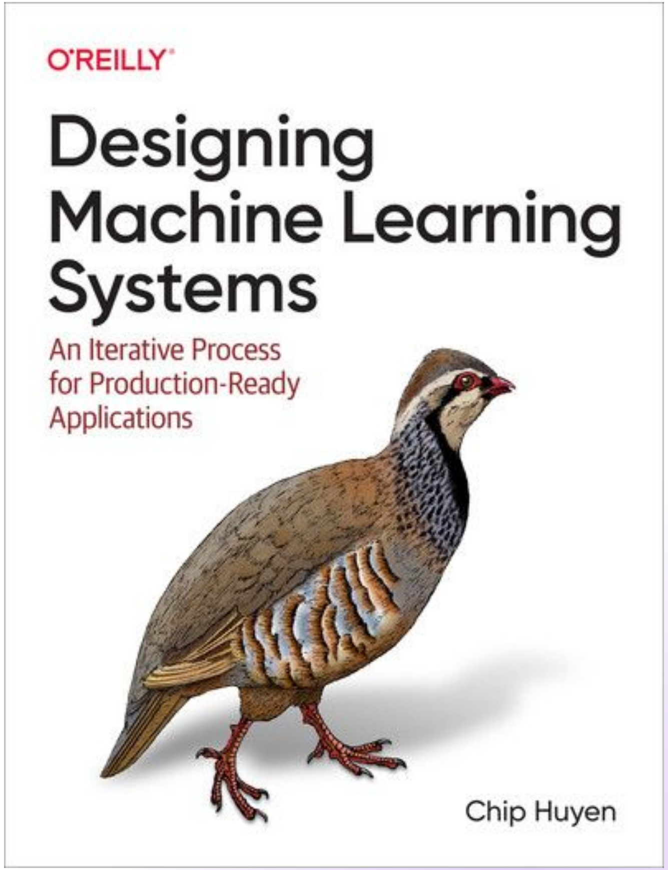


2020 state of enterprise machine learning. Source: Adapted from an image by Algorithmia

06
Literature

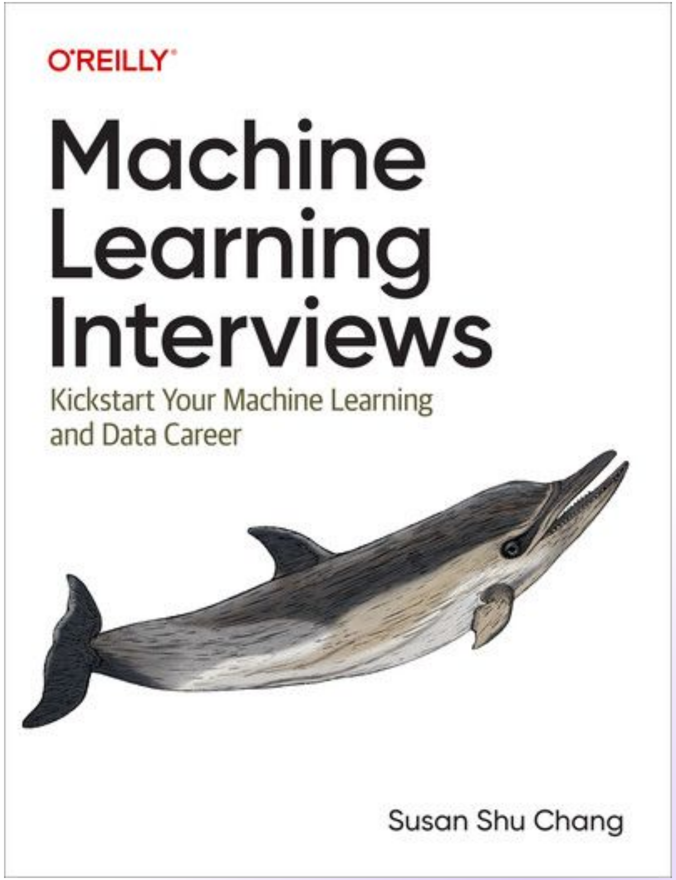
Topics

- Data Engineering Fundamentals
- Training Data
- Feature Engineering
- Model Development And Offline Evaluation
- Model Deployment And Prediction Service
- Data Distribution Shifts and Monitoring
- Continual Learning And Test in Production
- Infrastructure and tooling for ML Ops



Topics

- Machine Learning Roles And The Interview Process
- Machine Learning Job Application And Resume
- Technical Interview: Machine Learning Algorithms
- Technical Interview: Model Training And Evaluation
- Technical Interview: Coding
- Technical Interview: Model Deployment And End-To-End ML
- Behavioral Interviews
- Tying It All Together: Your Interview Roadmap
- Post-Interview And Follow-Up



Thanks!