

---

# Table of Contents

<b>Preface.....</b>	<b>ix</b>
<b>1. Overview of Machine Learning Systems.....</b>	<b>1</b>
When to Use Machine Learning	3
Machine Learning Use Cases	9
Understanding Machine Learning Systems	12
Machine Learning in Research Versus in Production	12
Machine Learning Systems Versus Traditional Software	22
Summary	23
<b>2. Introduction to Machine Learning Systems Design.....</b>	<b>25</b>
Business and ML Objectives	26
Requirements for ML Systems	29
Reliability	29
Scalability	30
Maintainability	31
Adaptability	31
Iterative Process	32
Framing ML Problems	35
Types of ML Tasks	36
Objective Functions	40
Mind Versus Data	43
Summary	46
<b>3. Data Engineering Fundamentals.....</b>	<b>49</b>
Data Sources	50
Data Formats	53
JSON	54

Row-Major Versus Column-Major Format	54
Text Versus Binary Format	57
Data Models	58
Relational Model	59
NoSQL	63
Structured Versus Unstructured Data	66
Data Storage Engines and Processing	67
Transactional and Analytical Processing	67
ETL: Extract, Transform, and Load	70
Modes of Dataflow	72
Data Passing Through Databases	72
Data Passing Through Services	73
Data Passing Through Real-Time Transport	74
Batch Processing Versus Stream Processing	78
Summary	79
<b>4. Training Data.....</b>	<b>81</b>
Sampling	82
Nonprobability Sampling	83
Simple Random Sampling	84
Stratified Sampling	84
Weighted Sampling	85
Reservoir Sampling	86
Importance Sampling	87
Labeling	88
Hand Labels	88
Natural Labels	91
Handling the Lack of Labels	94
Class Imbalance	102
Challenges of Class Imbalance	103
Handling Class Imbalance	105
Data Augmentation	113
Simple Label-Preserving Transformations	114
Perturbation	114
Data Synthesis	116
Summary	118
<b>5. Feature Engineering.....</b>	<b>119</b>
Learned Features Versus Engineered Features	120
Common Feature Engineering Operations	123
Handling Missing Values	123
Scaling	126

Discretization	128
Encoding Categorical Features	129
Feature Crossing	132
Discrete and Continuous Positional Embeddings	133
Data Leakage	135
Common Causes for Data Leakage	137
Detecting Data Leakage	140
Engineering Good Features	141
Feature Importance	142
Feature Generalization	144
Summary	146
<b>6. Model Development and Offline Evaluation.....</b>	<b>149</b>
Model Development and Training	150
Evaluating ML Models	150
Ensembles	156
Experiment Tracking and Versioning	162
Distributed Training	168
AutoML	172
Model Offline Evaluation	178
Baselines	179
Evaluation Methods	181
Summary	188
<b>7. Model Deployment and Prediction Service.....</b>	<b>191</b>
Machine Learning Deployment Myths	194
Myth 1: You Only Deploy One or Two ML Models at a Time	194
Myth 2: If We Don't Do Anything, Model Performance Remains the Same	195
Myth 3: You Won't Need to Update Your Models as Much	196
Myth 4: Most ML Engineers Don't Need to Worry About Scale	196
Batch Prediction Versus Online Prediction	197
From Batch Prediction to Online Prediction	201
Unifying Batch Pipeline and Streaming Pipeline	203
Model Compression	206
Low-Rank Factorization	206
Knowledge Distillation	208
Pruning	208
Quantization	209
ML on the Cloud and on the Edge	212
Compiling and Optimizing Models for Edge Devices	214
ML in Browsers	222
Summary	223

<b>8. Data Distribution Shifts and Monitoring.....</b>	<b>225</b>
Causes of ML System Failures	226
Software System Failures	227
ML-Specific Failures	229
Data Distribution Shifts	237
Types of Data Distribution Shifts	237
General Data Distribution Shifts	241
Detecting Data Distribution Shifts	242
Addressing Data Distribution Shifts	248
Monitoring and Observability	250
ML-Specific Metrics	251
Monitoring Toolbox	256
Observability	259
Summary	261
<b>9. Continual Learning and Test in Production.....</b>	<b>263</b>
Continual Learning	264
Stateless Retraining Versus Stateful Training	265
Why Continual Learning?	268
Continual Learning Challenges	270
Four Stages of Continual Learning	274
How Often to Update Your Models	279
Test in Production	281
Shadow Deployment	282
A/B Testing	283
Canary Release	285
Interleaving Experiments	285
Bandits	287
Summary	291
<b>10. Infrastructure and Tooling for MLOps.....</b>	<b>293</b>
Storage and Compute	297
Public Cloud Versus Private Data Centers	300
Development Environment	302
Dev Environment Setup	303
Standardizing Dev Environments	306
From Dev to Prod: Containers	308
Resource Management	311
Cron, Schedulers, and Orchestrators	311
Data Science Workflow Management	314
ML Platform	319
Model Deployment	320

Model Store	321
Feature Store	325
Build Versus Buy	327
Summary	329
<b>11. The Human Side of Machine Learning</b>	<b>331</b>
User Experience	331
Ensuring User Experience Consistency	332
Combatting “Mostly Correct” Predictions	332
Smooth Failing	334
Team Structure	334
Cross-functional Teams Collaboration	335
End-to-End Data Scientists	335
Responsible AI	339
Irresponsible AI: Case Studies	341
A Framework for Responsible AI	347
Summary	353
<b>Epilogue</b>	<b>355</b>
<b>Index</b>	<b>357</b>

