[WIP] Real-Time Machine Learning: A Practitioner's Guide

With Examples from Fintech

rtmlbook.queirozf.com

Felipe Q. B. Almeida

August 2026

20f7c61ff5c47f9833d0ef0eb10d65428f693d1d

Contents

Hello! **3**
About This Book 3

1 Real-Time ML: What and Why **5**
1.1 What Is Real-Time ML 5
1.2 Anatomy of a Real-Time ML System: Credit Card Fraud Detection . . . 9
1.3 What Real-Time ML Makes Possible 9
1.4 How Real-Time ML Differs from Traditional Systems 10
1.5 The Game Changes After the First Few Models in Production 10
1.6 Summary 10

Hello!

Welcome to Real-Time Machine Learning: A Practitioner's Guide. This book provides practical guidance for building and maintaining real-time machine learning systems in production as **an engineering discipline**.

About This Book

This book covers the entire lifecycle of real-time ML systems, from initial project planning to steady-state operation of a single model to scaling to multiple models across an organization.

The scope of this book is ML systems in *reasonable-scale* organizations.

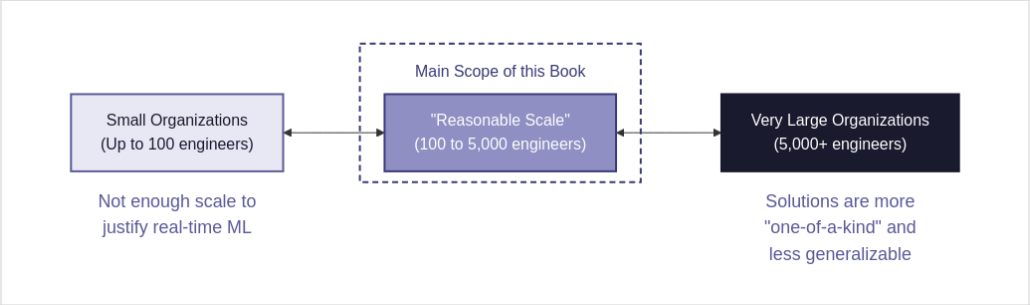

Figure 1: "Reasonable scale" organizations: not too small, not too large.

"Reasonable-scale" organizations are the sweet spot for this book, because **(i)** they have reached the level where there's data and real use-cases for real-time ML and **(ii)** the challenges they face are similar to those faced by other companies, so here is where an example-led approach works best.

Examples from Fintech

I firmly believe that the best way to teach is by leading with simple examples to quickly convey the big picture and then add an explanation, following up with more advanced examples.

Financial organizations such as banks and insurance companies have been using applied ML for a long time. There are interesting and complex problems, which allow us to explore the many dimensions and trade-offs involved in real-time ML. These include: adversarial incentives, regulatory constraints, market pressure from competition, cost concerns, high cost of failure, explainability requirements, among many more.

Fintech is just the modern iteration of traditional financial organizations so it's a natural place to draw examples from.

Patterns and Lessons Learned

In addition to examples, **patterns** and **lessons learned** are two ways I found to convey experience from operating real-time ML systems. These come from my own learnings and from the literature at large.

Patterns: things that are self-contained and common enough to merit a short "name". Examples: shadow-mode deployment, pre-mortem, walking skeleton.

Lessons Learned: These are short sentences with statements of fact, suggestions or warnings about a particular aspect of real-time ML.

What this book is not

- This is not an introductory book to ML. It assumes the reader already understands what ML is and how it's used.
- This book is not aimed at teaching you how to use a particular tool or framework. It is focused on high-level concepts and ideas instead.

1 Real-Time ML: What and Why

This chapter covers

- Lorem ipsum dolor sit amet, consectetur adipiscing elit
- Ut enim ad minim veniam, quis nostrud exercitation
- Duis aute irure dolor in reprehenderit in voluptate
- Excepteur sint occaecat cupidatat non proident, sunt culpa
- Quis nostrud exercitation ullamco laboris nisi ut aliquip

The internet has caused and explosion of digital organizations and data. This means that most interactions between customers and organizations are now mediated by systems and, being digital, collect a ton of data.

IMG data generated over time

this means that the opportunities to use ML have likewise exploded. many decisions flows in business and organizations can be otpimized if not downright replaced by ML systems.

> **Machine Learning**
>
> Machine learning (ML) is a collection of techniques to learn statistical patterns from data. This book assumes basic knowledge of ML.

in this chapter we will present a complete introduction to real-time ML, which means performing inference with ML models as part of real-time flows in a business or organization.

We'll give a clear picture of what it looks like and why you should care about it. We'll also introduce two key areas that'll be explored later in the book, namely the ways in which real-time ML systems are different from *traditional* systems and the patterns that emerge when one starts applying ML to several areas in an organization.

1.1 What Is Real-Time ML

Real-time ML refers to cases where one performs inference with an ML model and uses the predictions immediately as part of some decision process.

Real-Time ML VS Online ML

these terms are sometimes conflated but

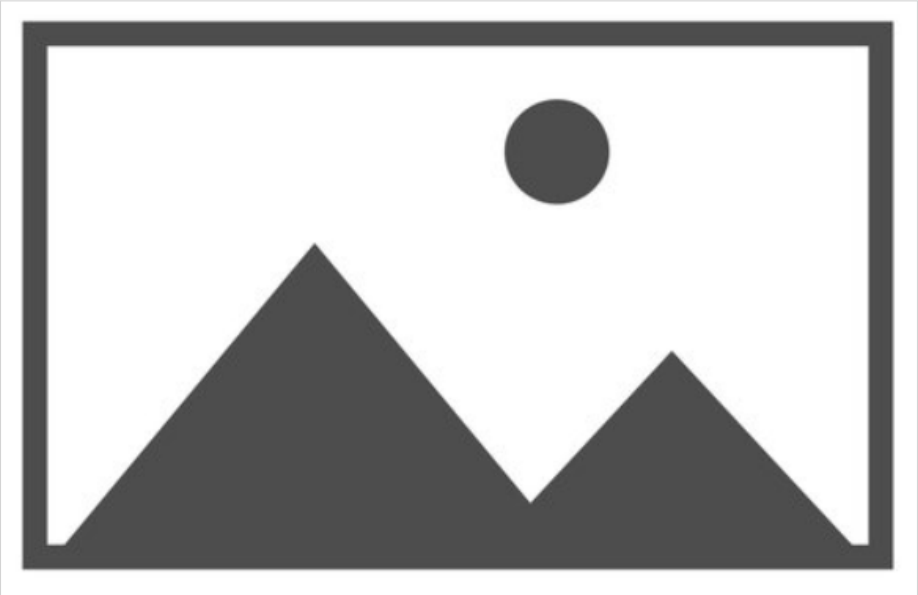

Figure 1.1: legend-here

Let's jump straight into examples:

- **Credit Card Fraud Detection**

  Detecting whether a given credit card operation is a fraud incident (e.g. leaked number, stolen card) is a classic use case of real-time ML.

- **Loan Origination**

  The main component in deciding whether to extend a loan to a customer is the likelihood the loan will be paid back. The default risk is often calculated by an ML model, in real-time.

- **Customer Service Chat Routing**

  Although many use Generative AI for support, customers still often need to speak to humans. Routing the call or chat to the correct team reduces the waiting time and increases customer satisfaction.

The 3 examples above display the basic pattern common to all ML-enabled real-time systems: a payload is built and fed to a previously trained model, which outputs a prediction immediately (in the order of dozens or hundreds of milliseconds), which is then used to make some kind of decision.

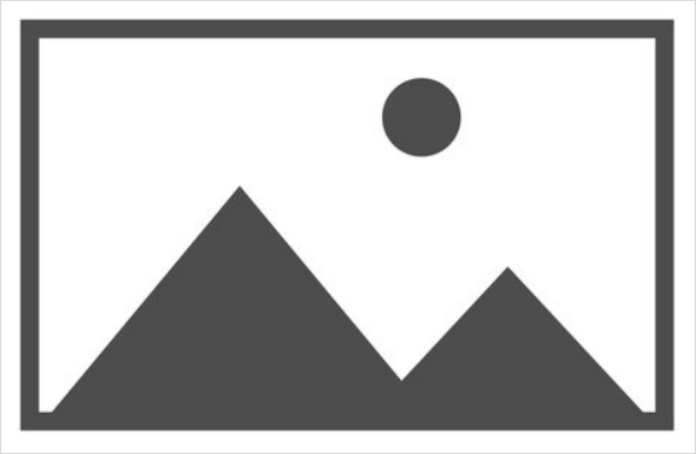

Figure 1.2: legend-here

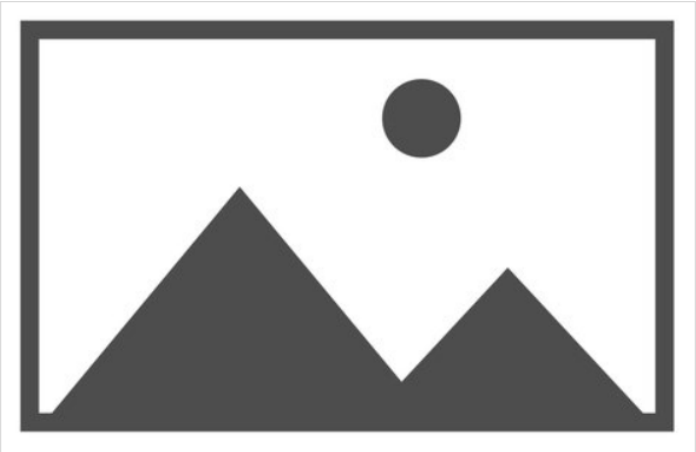

Figure 1.3: legend-here

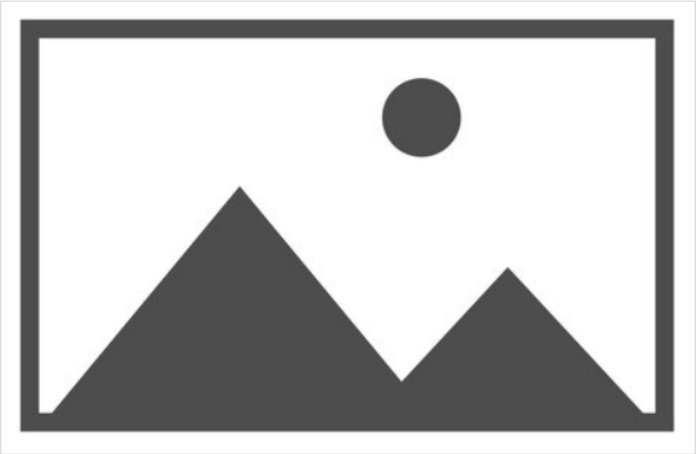

Figure 1.4: legend-here

But wait, aren't *all* ML systems like this? Not really.

ML models have been in use for a long time in banks and financial organizations, usually to predict probabilities (credit default, fraud, insurance claim) even before real-time systems were a thing. They were executed in *batch* (weekly or monthly), not in real-time. And even today, many ML use-cases can be accomplished with batch execution, which is simpler and cheaper.

Table 1.1: Batch ML vs Real-time ML

	What is Scored	**Latency**
Batch ML	A dataset containing many instances (rows)	Minutes or Hours
Real-Time ML	A single instance	Sub-second

Only in the last decades, with the spread of retail internet banking, have ML models been fully integrated into live, production systems that make decisions in these organizations. In this way, they have gone from being strictly in the realm of mathematics and statistics to being an *engineering* discipline. [1]

[1] maybe mention that the explosion of data from internet banking is another reason why RTML is so widespread today

1.2 Anatomy of a Real-Time ML System: Credit Card Fraud Detection

Let's look in detail at the same main components of a real-time ML system. Credit card fraud detection is the most commonly-cited example of real-time ML because it's a simple problem that lends itself well to ML modeling and it's clear it needs to run in real-time.

image must mention - human on an e-commerce website, which calls a "cloud" that calls the decision layer - fetch features - decision layer - logs - model service

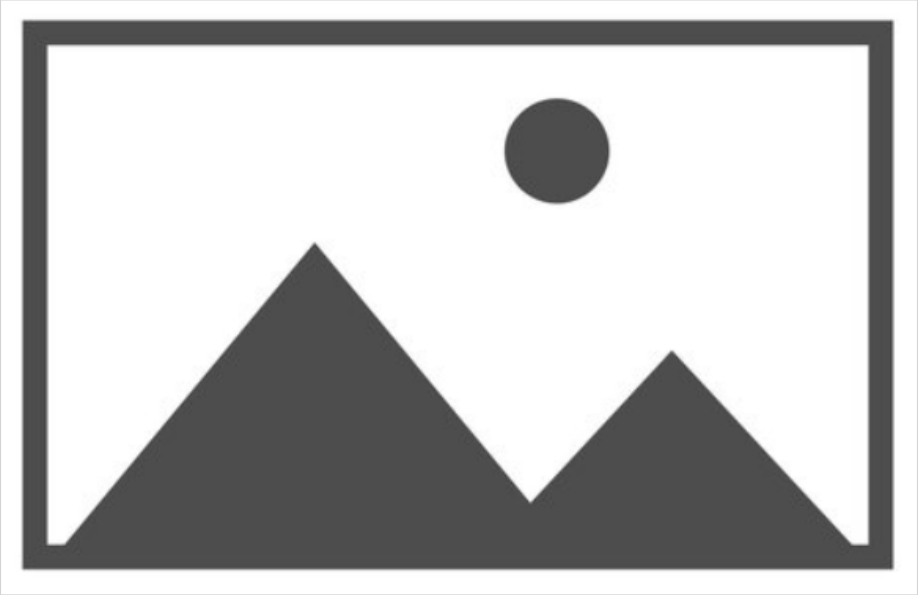

Figure 1.5: legend-here

The image names several important components:

- **Initiator**
- **Decision Layer**
- **Features**
- **Model Service**
- **Logs**

1.3 What Real-Time ML Makes Possible

Modern organizations are mostly or even fully digital enterprises; there's often no brick-and-mortar interface, only an web or a website.

This means that systems mediate every interaction and collect data from it. So one can apply RTML to almost all business flows.

1.4 How Real-Time ML Differs from Traditional Systems

1.5 The Game Changes After the First Few Models in Production

1.6 Summary