Design Approaches to ML System Architecture

General ML Architectures

1. Train by batch, predict on the fly, serve via REST API
2. Train by batch, predict by batch, serve through a shared database
3. Train, predict by streaming
4. Train by batch, predict on mobile (or other client)

Architecture Comparison

	Pattern 1 (REST API)	Pattern 2 (Shared DB)	Pattern 3 (Streaming)	Pattern 4 (Mobile App)
Training	Batch	Batch	Streaming	Streaming
Prediction	On the fly	Batch	Streaming	On the fly
Prediction result delivery	Via REST API	Through the shared DB	Streaming via Message Queue	Via in-process API on mobile
Latency for prediction	So so	High	Very Low	Low
System Management Difficulty	So so	Easy	Very Hard	So so

Focus of this course - best trade-off for most cases.

See you in the next section!