MODULE 2 · SESSION 1

Measure Reliability Before Scale

How to build an evaluation loop that keeps detecting quality problems as the work, the data and the users change.

Theory session

Beginner

Running case: HarborPilot Health

ORIENTATION

Where this session sits

Module 1 built the system; Module 2 keeps it under control

MODULE 1
AI System Foundations

MODULE 2
Controlled AI Operations

MODULE 3
Strategic AI Operating Choices

SESSIONS IN THIS MODULE

2.1 Measure Reliability Before Scale
An evaluation loop that detects quality drift
YOU ARE HERE

2.2 Bound Autonomy With Human Escalation
What the system may decide alone, and what it must escalate

2.3 Defend and Recover AI Services
Detect an attack, contain it, and restore trust

SESSION OUTCOMES

What this session explains

Knowing whether your system is still working

01

A model's quality is not fixed

The world moves; a system that was accurate in March can be wrong in September without changing at all.

02

Evaluation is a loop, not an event

One launch report tells you nothing about next month.

03

Different audiences need different rhythms

Weekly for the people who can act; quarterly for the people who allocate.

04

Measure before you scale

Scaling an unmeasured system multiplies a problem you cannot see.

DEFINITION

What continuous evaluation means

Three parts, all of which have to be present

WORKING DEFINITION

Continuous evaluation is a standing process that measures live system quality on a fixed rhythm, compares it against a defined expectation, and routes any gap to someone who can act on it.

1

A measurement

Something you can compute from real traffic, not a one-off benchmark run at launch.

2

An expectation

A stated level below which the result counts as a problem. Without it, a number is just a number.

3

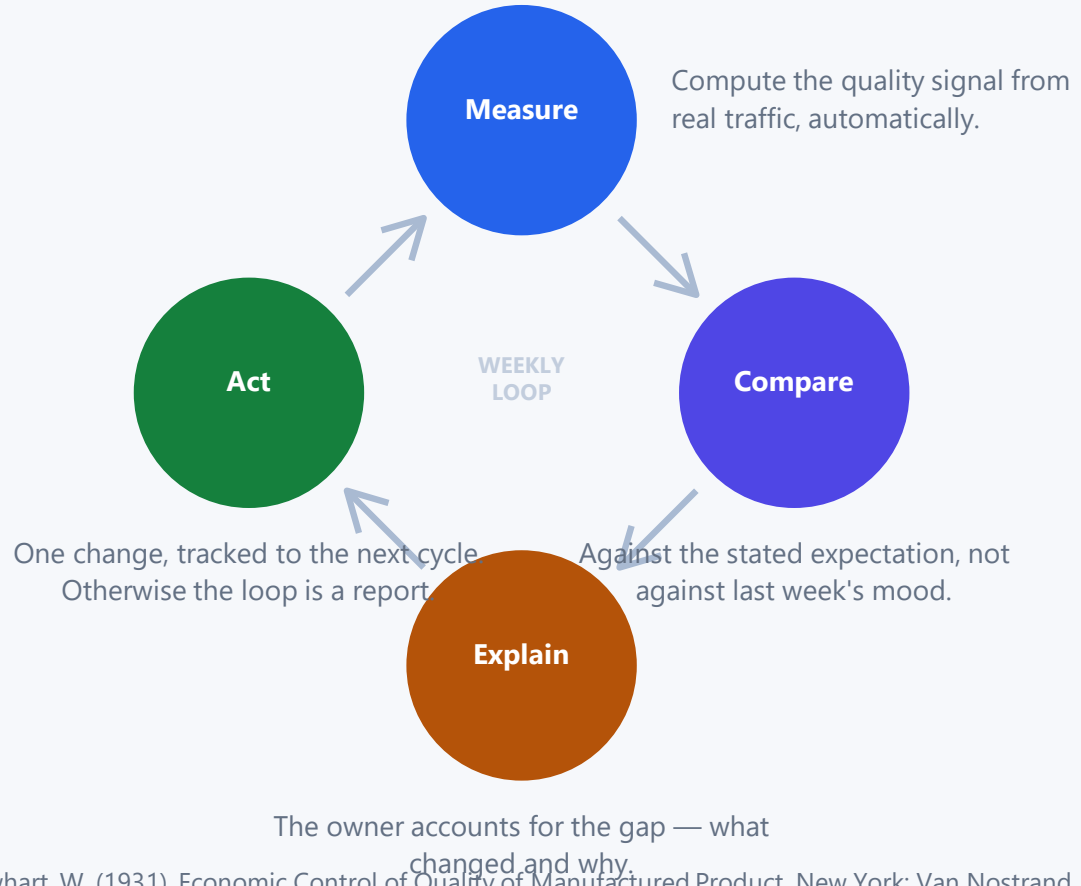
An owner

A named person who receives the gap and can change something. Otherwise the loop does not close.

CONCEPT 1 · THE LOOP

What a working evaluation loop looks like

Four steps, repeating on a fixed schedule



Shewhart, W. (1931). Economic Control of Quality of Manufactured Product. New York: Van Nostrand.

WHAT MADE IT WORK

Two audiences, two rhythms

Weekly metrics to the teams doing the work; quarterly progress to leadership. Each has a defined place in the loop.

THE COMMON FAILURE

A loop with no action step

Dashboards that are read and not acted on train everyone to stop reading them.

THE UNDERLYING IDEA

Measure, compare, act — repeat

This is statistical process control, unchanged since the 1930s and still the right shape.

CONCEPT 2 · DRIFT

Three ways quality decays without anyone changing anything

Knowing which one you have determines the fix

DATA DRIFT	The inputs change. New packaging, new suppliers, a different mix of requests.
CONCEPT DRIFT	The right answer changes. What counted as urgent last year no longer does.
UPSTREAM CHANGE	Something feeding the system changed — a renamed field, a new sensor, a failed job.
FEEDBACK EFFECTS	The system's own outputs change the behaviour it later learns from.

HOW TO TELL THEM APART

Watch inputs and outcomes separately

Inputs shifting with stable accuracy is data drift. Stable inputs with falling accuracy is concept drift.

THE FASTEST ONE TO FIND

Upstream change

Usually a sharp step rather than a slope, and usually fixable in an afternoon — if you are looking.

THE SNEAKIEST

Feedback effects

Suppressing an item from search means you never learn whether it was available.

Gama, J. et al. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1–37.

CONCEPT 3 · WHAT TO MEASURE

Four layers, each catching different problems

Most teams measure only the third

Layer	Example	What it catches
Inputs	Volume, missing fields, distribution of request types.	Upstream breakage and data drift, usually first.
System	Latency, error rate, cost per request.	Problems your users feel before any accuracy metric moves.
Model output	Accuracy on a labelled sample, confidence distribution.	Quality decay — but only where you have labels.
Outcome	Waste, expiration rate, owned inventory.	Whether the system is doing the job it exists for.

WHERE THE CASE'S NUMBERS SIT

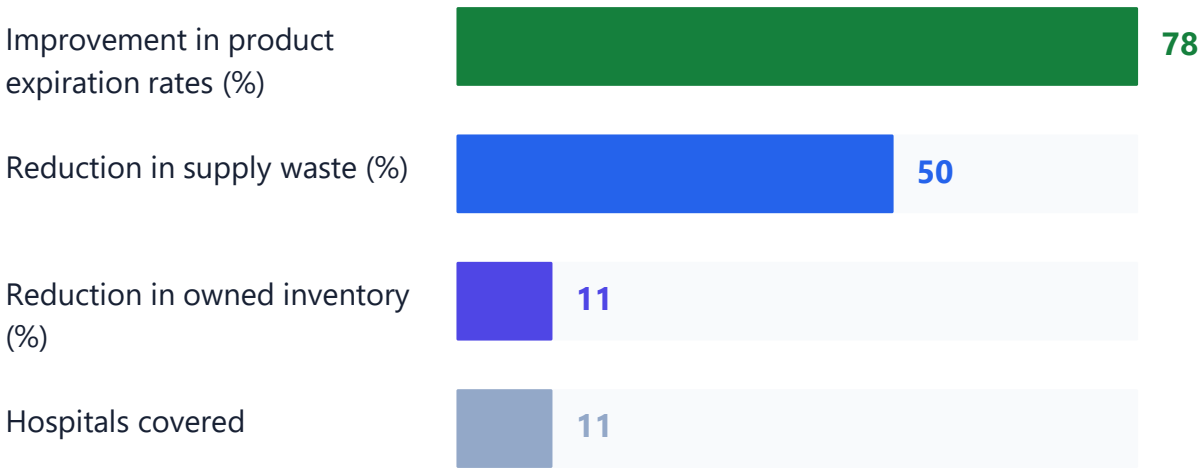
The hospital system's headline numbers were all outcome measures: 78% better expiration rates, 50% less waste, 11% less owned inventory. Those are what the organisation actually wanted.

CONCEPT 4 · EVIDENCE

What the measurement rhythm produced

Eleven hospitals, one loop

REPORTED CHANGE AFTER IMPLEMENTATION



The technology alone was not the change. The weekly and quarterly rhythm around it was.

WHY A SHARED SIGNAL MATTERED

Two functions, one picture

Logistics and clinical teams had been seeing the same environment through different needs.

THE UNMEASURED BENEFIT

Stronger working relationships

A shared number to argue about turns out to be a cooperation mechanism.

THE ORDER OF EVENTS

Visibility first

Nothing improved until performance became visible to the people who could change it.

CONCEPT 5 · TWO KINDS OF SIGNAL

Leading and lagging indicators

You need both, and they answer different questions

LEADING

Tells you early

EXAMPLES

Input distribution shift, confidence falling, escalation rate rising.

STRENGTH

Moves before customers are affected, so you can act in time.

WEAKNESS

Noisy. Some movements mean nothing, so thresholds matter.

LAGGING

Tells you truly

EXAMPLES

Waste, expiration rates, customer complaints, returns.

STRENGTH

Unambiguous — this is the outcome you actually care about.

WEAKNESS

Arrives after the damage, sometimes by months.

HOW TO COMBINE THEM

Run leading indicators weekly and lagging indicators quarterly — which is exactly the cadence the hospital system adopted, for exactly this reason.

CONCEPT 6 · CADENCE

How often to look

Too often and too rarely both stop the loop working

TOO RARELY

Quarterly only. Problems run for months and the cause is untraceable by the time you notice.

MATCHED

Weekly for the people who can act; quarterly for the people who allocate resources.

TOO OFTEN

Daily reviews of a noisy signal. People react to random variation and make things worse.

CHOOSING THE INTERVAL

Ask	And set the rhythm to match
How fast can this actually change?	Look slightly more often than the thing moves, and no more.
How long would a fix take?	There is no value in detecting faster than you can respond.
Is the signal noisy?	Noisy signals need a longer window or a control limit, not more meetings.

CONCEPT 7 · FAILURE MODES

Three evaluation habits that feel rigorous and are not

All three produce numbers nobody acts on

The launch benchmark

Looks like A thorough evaluation before release, and nothing after.

Why it fails It measures a world that stops existing the day you ship.

Fix A standing weekly measurement on live traffic, however small.

Averages that hide segments

Looks like One overall accuracy number, holding steady.

Why it fails A badly failing segment is invisible inside a large healthy average.

Fix Report by segment: site, request type, customer tier.

A dashboard with no owner

Looks like Comprehensive charts nobody is accountable for.

Why it fails Measurement without an owner produces awareness, not change.

Fix Each metric names one person who explains it and one action per cycle.

CASE STUDY

An eleven-hospital system

Clinical supply technology whose value came from the measurement rhythm around it.

CASE STUDY · AN ELEVEN-HOSPITAL SYSTEM

Weekly metrics, quarterly progress

A recurring loop between measurement and action

BEFORE

Inconsistent practice, local views

THE SITUATION

Inventory practices were inconsistent across eleven hospitals.

THE MEASUREMENT PROBLEM

Logistics and clinical teams saw the same supply environment through different operational needs.

THE CONSEQUENCE

Discussions stayed local, and leadership saw only an incomplete picture of performance.

AFTER

One loop, three audiences

THE CADENCE

Weekly metrics to both logistics and clinical organisations; quarterly progress to leadership.

THE RESULTS

Product expiration rates improved 78%, supply waste fell 50%, and owned inventory fell 11%.

THE SIDE EFFECT

Relationships between clinical and logistics teams became stronger, not just the numbers.

READING THE CASE

The system did not rely on a single report or a one-time review. A recurring loop between measurement and action is what produced the change.

Drawn from the session case pack.

APPLICATION

Bringing the theory to HarborPilot Health

Building the evaluation loop before scaling

Step	What you put in place	HarborPilot example
Pick one outcome measure	The thing the product exists to change.	Late or missed clinical deliveries per thousand requests.
Add leading indicators	Signals that move before the outcome does.	Escalation rate, confidence distribution, input volume by site.
State the expectation	The level below which it is a problem.	Written down per metric, agreed with the clinical team.
Name an owner per metric	One person who explains the gap.	Ops lead for delivery outcomes; ML lead for confidence and escalation.
Set two rhythms	Weekly to act, quarterly to allocate.	Weekly review with both functions; quarterly summary to leadership.
Segment everything	Never report a single average.	By hospital site, by request type, by shift.

WHAT COMES NEXT

Session 2.2 uses this measurement to decide what the system may do on its own.

SUMMARY

What to carry out of this session

1. **Quality decays without anyone changing anything.**
   Data drift, concept drift, upstream breakage and feedback effects.

2. **A loop needs a measurement, an expectation and an owner.**
   Missing any one of the three and it is a report, not a loop.

3. **Measure four layers, not one.**
   Inputs, system, model output and outcome each catch different failures.

4. **Match the rhythm to the audience.**
   Weekly for those who can act; quarterly for those who allocate.

5. **Never report a single average.**
   A failing segment disappears inside a healthy overall number.

FURTHER READING

Sources for this session

Primary references behind each concept

PROCESS CONTROL	Shewhart, W. (1931). *Economic Control of Quality of Manufactured Product.* New York: Van Nostrand.
MANAGEMENT METHOD	Deming, W. E. (1986). *Out of the Crisis.* Cambridge, MA: MIT Press.
CONCEPT DRIFT	Gama, J. et al. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 1–37.
ML TECHNICAL DEBT	Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. *NeurIPS 28*.
MONITORING PRACTICE	Breck, E. et al. (2017). The ML test score. *IEEE Big Data 2017*.
SERVICE OBJECTIVES	Beyer, B. et al. (2016). *Site Reliability Engineering.* Sebastopol: O'Reilly.
EVALUATION PRACTICE	Huyen, C. (2022). *Designing Machine Learning Systems.* Sebastopol: O'Reilly.