

Exception Intelligence & Margin Leakage

Build EI-4716D1. Every figure below comes from that run of the published artifacts. If an evidence package in front of you names a different build, it is a different set of numbers.

The aggregate that looked fine

Halbrook Industrial Supply finished the year at **\$38,399,999** of revenue and a **26.4%** gross margin, down from **28.0%** two years earlier. That is 1.59 points across twenty-four months. Nothing in a monthly package makes a movement that size look like an event, and the package reported one blended margin in absolute dollars, so nothing in it said which accounts, which products or which branches had moved.

The company is not short of the data. 278,169 invoice lines across 142 accounts and 36 months are enough to isolate where the movement sits and which questions management has to resolve. They do not answer all of them: the ledger carries a cost, not a cost breakdown, so it can show that cost rose and price did not and cannot say the increase was freight. The question is whether anyone can be pointed at the few lines that matter without reading all of them.

What was underneath it

A blended margin is an average of two businesses. Distribution sells parts at 22.1%; Installed Service sends technicians out at 41.8% — nearly double. Both moved a little over the window: Distribution gave up 1.30 points and Installed Service gained 0.60. Weighted by what each was worth, that is 0.81 points of the fall.

THE DECLINE, DECOMPOSED	POINTS
Net effect of within-line margin changes	-0.81
Mix shifting toward Distribution	-0.79
Total decline	-1.59

The first row is a net: Distribution deteriorated and Installed Service improved, and Distribution is much the larger line. Calling it "line margins deteriorating" said both lines got worse, which the two lines above it disprove. The second row holds the closing line margins constant and changes only the revenue mix - a counterfactual, not a claim that either margin stood still.

The other 0.79 points is the higher-margin line shrinking from 26% of revenue to 22%: holding the closing line margins fixed, the worse revenue blend explains that on its own. That is a different problem from either line getting worse and it needs a different decision. Pricing changes alone would not reverse the mix component; holding segment margins constant, a higher share of Installed Service revenue would mechanically improve the blend, on a longer timescale than a pricing decision. A report showing only the blend cannot tell the two apart, and this one showed only the blend.

The decline is concentrated. Contribution analysis attributes the movement in blended margin to each account exactly, and the contributions sum to the movement. **2 of 142 accounts** — Meridian Mechanical Contractors at -0.554 points and Cedar Valley Foods at -0.327 — account for half of it. That is an attribution, not a ranking of how unusual anything looks: an account is on the list because of how much of the decline it explains.

And one finding is not a margin problem at all. 180 invoice lines were billed per-each at per-case pricing, carrying negative gross profit on their face. Nobody needs to negotiate anything; the records are wrong and get corrected. A record belongs to the issue defined by the full set of rules that caught it, so that no line's money is counted twice — which puts the 179 lines both record rules caught at the top of the queue, worth \$14,373 a year, and the one that only the below-cost rule caught in a row of its own further down it.

What the engine produced

629 detections across six ways of cutting the book, aggregated to **122 issues**, 97 of them with a size estimate. Every issue carries the baseline it departed from, by how much, what it is estimated to cost, and who should act — because what went wrong is not the same as what someone must do.

THE TOP OF THE QUEUE	GRAIN	MEASURE	EST. IMPACT	ACTION
179 lines	record	gross profit	-\$14,373/yr	correct
Omaha	branch	gross margin	-\$210,033/yr	investigate
Electrical	product family	cost price spread	-\$292,030/yr	reprice
Pumps	product family	cost price spread	-\$101,445/yr	reprice
Trellis Works	customer id	discount rate	-\$44,881/yr	renegotiate
Kestrel Fabrication	customer id	discount rate	-\$37,248/yr	renegotiate

These estimates overlap and must not be summed. An account sits inside a branch and a product family inside a line of business, so the same economics is reported at every cut a manager might ask about. The sized rows come to \$7,531,753 against \$612,096 of actual margin deterioration — 12 times over. Every row carries `impact_additive: false` for that reason. The exact figure for how much moved is the margin bridge above; the queue's job is to say where to look first, and its estimates size individual findings rather than adding up to anything.

Rules run before statistics, because a rule management already understands produces an explanation rather than a flag. A line billed below what it cost is an economic exception on its face, even where the business later decides it was intentional; an account discounting past a configured threshold is a conversation someone can have. Statistical detection then covers what no rule anticipated — a segment sitting away from its peers, or a series that changed level on a date.

A per-unit measure is only computed where a unit is one thing. The book bills in eaches, feet, boxes, cases and hours. A product family is one of them and Installed Service is all hours, so a cost per unit there is a real cost per unit; a customer, a branch or the Distribution line is four at once, and a "unit cost" over the result moves with the product mix while no cost has changed. Twenty of twenty-nine per-unit rows in an earlier version of this queue were that artifact, including the largest estimate in it.

How well it performed

Measured against a ground-truth register, not asserted. The dataset is generated with its problems deliberately seeded and recorded, so what the engine found is checked against what was put there.

DETECTION PERFORMANCE

Seeded problems found	7 of 7
Individual segments identified	14 of 14
Mis-billed lines found	180 of 180
Lead action right, per condition	6 of 6
Expected action surfaced somewhere	6 of 6
Management issues promoted	20 of 122 rows
Issue precision, promoted tier	100%
Issue precision, top ten rows	100%
Issue precision, whole queue	19%
Impact error, median over 5 registered magnitudes	2.08x
Impact error, worst	3.33x on cost that rose on six accounts and was not passed through
Routes never firing on a control account	4 of 7 measured

Every seeded segment was identified at the grain the register records it at. The estimate runs high because a comparison against peers attributes all of a segment's difference to the problem when some of it is the segment being genuinely different; publishing what it is worth beats calling it accurate.

That impact figure has a denominator worth stating. Only five of the seeded conditions carry a registered financial magnitude an estimate can be scored against, and the median is over those. The 97 sized queue rows have no ground-truth amount to be compared with, so nothing here claims their estimates are accurate to any figure - and "median error 2.08x" would be read as covering all of them if it were left unqualified.

Two of those rows say different things and the difference matters. The lead action is what a manager reads first on the most severe row for a condition, and it is right 6 times of 6. Whether the expected action appeared anywhere among the rows for that condition is a weaker question, and the answer to it is 6 of 6. The weaker figure was the one published, under a heading that read as accuracy. It is both here, because the strict one is the claim worth making and the gap between them is worth seeing.

Two precision figures, because landing on the right population is not the same as being right. A row scores on the seeded-population hit rate if its segment carries a registered condition at that row's grain. Issue precision is the question a manager actually has - if I open this row, is the stated problem the problem? - and it is measured against what the register says each condition **is**: the measure whose movement defines it, a detector that can establish it, the grain the evidence appears at, and the action it asks for. All four have to agree. On the whole queue the two readings are 32% and 19%, and the gap is rows that found the right account and described a different thing about it.

That distinction is not academic. Installed Service carries the mix shift, and a trend break fires on that line's revenue, gross profit, realized unit price, order value and revenue share. Scoring on the segment alone credited all five, published a trend-break precision of 71% at line-of-business grain, and used it to promote a realized-price movement above the mix shift itself. Precision is measured per detector, grain **and measure** now, which is the same lesson as the grain one level down, and `detector_measures.json` ships each cell with the count behind it.

A row that is real movement on a seeded population without being the registered finding is reported as corroborating evidence - 9% of the queue - rather than as either a finding or noise.

So the queue is tiered, and only one tier is presented as work. 20 rows reach the management tier - on decisive evidence, or because the route that found them has measured precision of at least 70% at that grain and on that measure - and they run at 100% issue precision **in sample**. The other 102 are published as review signals: explained, checkable, and not put in front of anyone as something to act on. Calling all 122 of them issues was a promise the same artifacts disproved.

In sample, and the qualification is not a formality. A route is promoted on precision measured against the same register the tier is then scored against, so the promotion threshold is calibrated on this demonstration's own answer key. `promotion_calibration.json` says how much evidence stands behind it: held out one condition at a time, 5 of the 6 promoting routes make no detections anywhere except on the condition they promote. A precision of 1.00 for the discount rule means the eight accounts it flagged are the eight the register seeds, with nothing else in the data for it to be wrong about.

Splitting this population into calibration and evaluation sets would not fix that, which is worth saying rather than implying: one synthetic company carries one instance of each condition, so holding the instance out

leaves nothing to evaluate. What the tier demonstrates is the workflow - measure a route's quality, promote the reliable ones, leave the weak ones as review signals - and not a generalisation estimate. In an engagement, promotion is calibrated from adjudicated history and monitored prospectively against outcomes.

The precision is bought with recall, and here is the price. The whole queue carries a row for 7 of the 7 seeded conditions; the promoted tier carries 6. The one it leaves behind is the silent volume attrition, whose route measures 27% precision - the three real accounts rank first, second and third, and eight other accounts shrank too, so no threshold separates them. Those rows are in the review tier with their evidence attached rather than promoted on a route the measurement does not support.

A cohort of 28 accounts — 19.7% of the book — carries no seeded condition and absorbs none of the compensation. These are **unseeded account-level controls**, which is a narrower claim than it was. Seven of them sit in Omaha, which is the branch the register seeds a divergence on, and five belong to the account manager whose book carries the discount condition - so an account-grain flag on one is a false positive at account grain and not proof that nothing about the account's surroundings moved. What the cohort does give is the only denominator that means anything, and the routes are reported as they measure. 4 of the 7 measurable ones never fire on a control account. Statistical detection at account grain does: peer departure sits at a lift of 0.18 and trend break at 0.73, where 1.0 would mean the route is indifferent to whether an account is clean. Attribution is the route that works at that grain, because it never asks whether an account is unusual — only how much of the movement it accounts for, which has an exact answer.

One problem was deliberately left out of the data. Concentration growth was expected to be part of the story and is not: the largest account falls from 9.3% of revenue to 7.8%, the top ten from 42.2% to 40.3%, and a Herfindahl index computed across the whole distribution falls too. The engine reported it as absent. A demonstration that only ever confirms what it looks for has not shown it can tell the difference.

Three capabilities this demonstrates

A finding is reported at the grain it exists at. Cost rose on six accounts and their prices did not follow, a cost step on two product families, discounting on one account manager's book, divergence on one branch, and mis-billing on individual lines. A problem is credited only to a detection at the grain the register records it at, so a customer-level flag does not take credit for a branch-level condition.

Attribution alongside detection, because a movement is only actionable once it is attributed. The exception queue says what looks wrong; contribution analysis says how much of the decline each account explains, exactly, summing to the movement with no residual. The two answer different questions and the page publishes both.

Every claim on the page is recomputable from data shipped with it. The register is the evidence: every accuracy, every rate, every impact ratio is computed against it rather than stated. The register is withheld from the engine and used only after the outputs exist, which is the independence that matters - it does

ship beside those outputs, because a published accuracy figure whose evidence is absent cannot be checked at all. The size estimate's error is published for the same reason: a claim about financial impact that cannot be checked is an assertion about validation rather than validation.

Where the work would apply

Margin and profitability analysis, discount and pricing governance, customer and product mix reporting, month-end exception review, and any recurring management pack whose aggregates are hiding the movements underneath them.

See the full project: materialsignal.co/work/exception-intelligence

Demonstration project using synthetic data designed to simulate a realistic distribution business with an installed-service arm. No client data appears anywhere in this portfolio.

Demonstration build RC-698FA2 · exception EI-4716D1 · case digest HB-9EF73D · generated 2026-09-14