VERIFIABLE AI LAB ANALYSIS

MEASURED. SIGNED. TRUSTED?



A food test.
A product-authenticity marker check.
A forensic analysis for criminal investigation.
A cleaning measurement certificate.

Different results. One question:
can the receiver verify the evidence?

Building on a Spherity Research case study

A result can take many forms

Food testing

A measured concentration, a detected substance or a finding about a tested sample.

Textile and pharma authenticity

An invisible-marker match linked to a product, batch or authorised reference.

Product provenance

A scoped origin or identity finding, supported by traceability records and chain of custody.

Milk-tank cleaning

A measurement certificate linked to the tank, cleaning cycle, method, time and device.

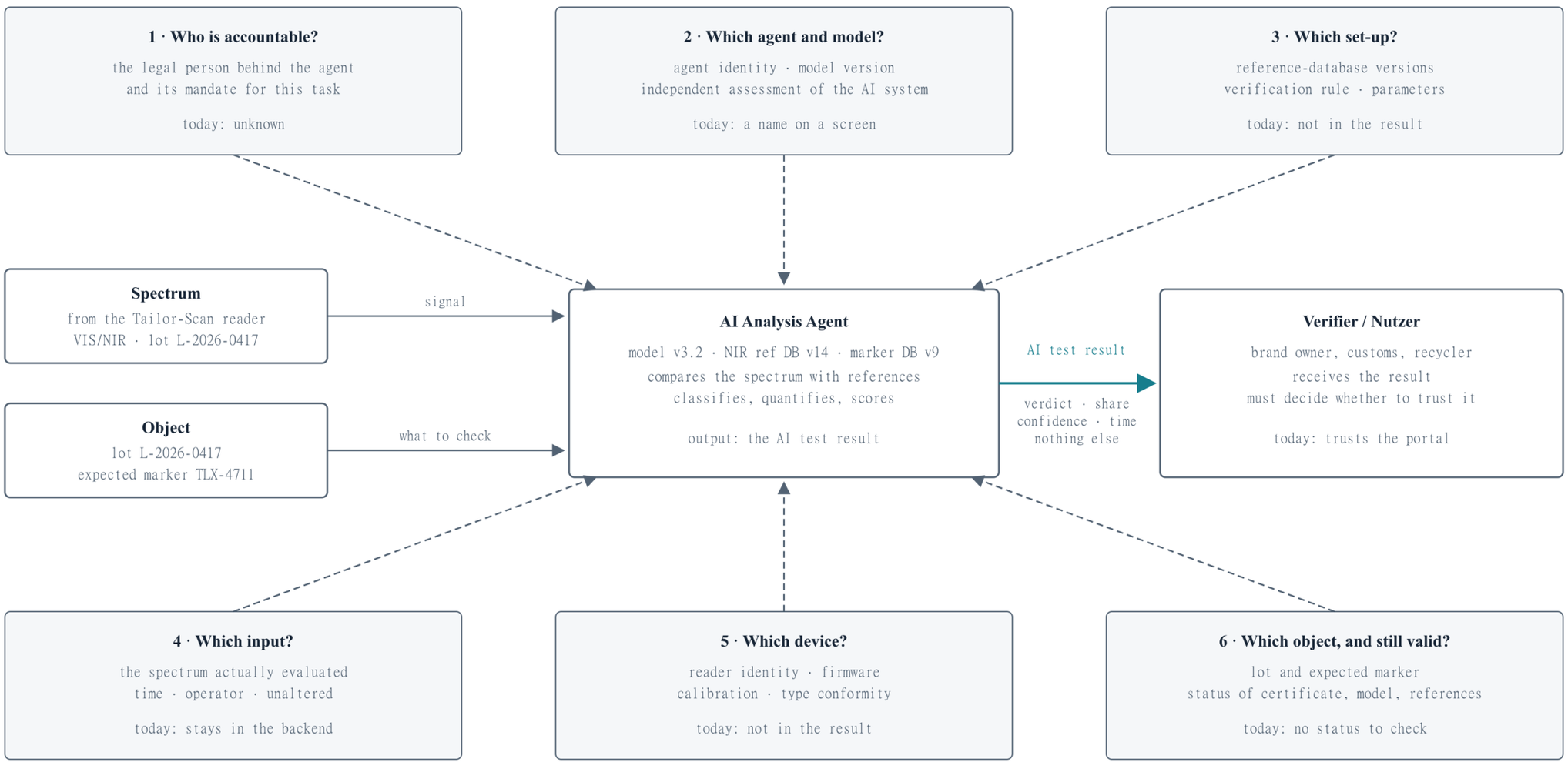
Forensic / criminal investigation

A laboratory finding linked to an evidence item, method, analyst and documented chain of custody.

A pass/fail verdict is one possible interpretation.

Illustrative applications of the evidence pattern.

The result is only the beginning



Conceptual design · identifiers and role assignments are illustrative · no endorsement or certificate issuance claimed

FIGURE 1 / SIX QUESTIONS A RECEIVER SHOULD ASK

01 **Who is accountable?**

02 **Which agent and model?**

03 **Which configuration?**

04 **Which exact input?**

05 **Which device?**

06 **Which object and status?**

Figure depicts a portal-only baseline, not a claim about every lab.

Original Figure 1: Dr. Carsten Stöcker / CC BY 4.0. Tap the figure to inspect.

What did the agent actually receive?

PROPOSED INPUT AND PROVENANCE RECORD

01 Bind the physical sample

Sample, lot or tank; collection time and chain of custody.
Raw signal, units, operator, device and calibration reference.

02 Preserve each transformation

Hash retained raw data and processed input separately.
Pin preprocessing code, parameters and reference snapshots.

03 Capture the task context

Record task instructions, policy and supplied context.
For tool-using agents: retain retrieved data and tool outputs.

04 Bind the run to its output

Link run ID, AI Service Passport, execution log and result hash.
A declared configuration is not proof of actual execution.

Proposed design; provenance concepts: W3C PROV. Hashes do not prove input truth.

Calibrate. Measure. Match. Bind the evidence.

CALIBRATION / TRACEABLE STANDARDS AND REFERENCE MATERIALS

DCR
Calibration request

Device ID + scope
Range + requirements

Calibration laboratory

Standards / CRM IDs
Values + uncertainties

DCC
Calibration certificate

Results + uncertainty
Method + traceability

Check calibration applicability to this setup

Input material / evidence item

Sample or lot ID
Operator + time
Chain of custody

Device setup + measurement

Setup + consumable lots
Raw signal + method
Value + uncertainty

Reference matching

Marker, fingerprint / DNA
DB version + match rule
Finding + limitations

VERIFIABLE RESULT CREDENTIAL

Bind DCR/DCC, device, input, consumable and reference-material IDs.
Link certificates, run controls, raw data, matching and uncertainty.

DCR = Digital Calibration Request; DCC = Digital Calibration Certificate.

Terminology: QI-Digital / PTB. Proposed Spherity evidence workflow.

QI-Digital context: Bundesanstalt für Materialforschung und -prüfung (BAM).

Trust the materials behind the measurement.

CONSUMABLES + REAGENTS + REFERENCE MATERIALS

01 Identity & authenticity

Manufacturer, product, lot / unit
Supplier, receipt and seal evidence
Match item to issuer records

02 Certificate & scope

Verify issuer, lot and revision
CRM: certified value + uncertainty
Check method, matrix and range

03 Condition at use

Expiry, opened / prepared time
Storage and transport evidence
Track dilution and aliquot lineage

04 Release for this test

Check method approval + lot QC
Bind blanks and controls to run
Record accept / hold + reviewer

REAL EXAMPLES / DETAILS AND SOURCE SCOPE IN THE NOTES

Food: immunoaffinity columns; ERM-BD284 milk-powder CRM.

Forensic DNA: DNA-grade swabs; STR reagents and control DNA.

Cleaning: UltraSnap ATP swabs and a positive control kit.

Optical measurements: BAM-FCalKit fluorescence CRMs.

Bind the actual lots used, their evidence and the release decision.

A certificate of analysis alone does not certify fitness for every test.

CRM certificate scope: European Commission / JRC. Proposed evidence design.

Forensic consumable production: ISO 18385:2016. Examples are not endorsements.

Give the agent a Verifiable AI Service Passport

PROPOSED PASSPORT / IDENTITY + SERVICE + VALIDATION

Identity & mandate

Accountable organisation
Agent key and delegation
Task scope and expiry

Verifiable Agent Card

Service identity, skills, endpoints
Provider signature + version
Authentication requirements

Deployed configuration

Model, tools and permissions
Validated reference versions
Operating domain and limits

Verifiable Agent Validation

Based on task-specific profiles
Test reports + confidence intervals
Assessor, scope and status

The passport connects identity to validation evidence.
Each run binds its result to the passport version.

A signed Agent Card alone does not establish capability.

Proposed passport. Agent Card context: A2A specification; supporting notes included.

Validate against a task-specific profile

PROPOSED PROFILES / VERSIONED TESTS + ACCEPTANCE RULES

01 Marker recognition

Positives, negatives, mixtures, near matches and degraded samples.
Report sensitivity, false accepts and confidence intervals.

02 Reference databases + marker matching

Pin validated reference versions, provenance and matching rules.
Test held-out lots, wrong references and out-of-domain inputs.

03 Biological fingerprint / forensic evidence matching

Specify the modality, custody, contamination and comparison tests.
Use method-specific error metrics and statistical interpretation.

Profile version + tested configuration + assessor + report

Specify material grades, certificate scope and lot acceptance.

Bind uncertainty, limitations and status to the validation claim.

A profile defines the test. Evidence shows how the service performed.

Proposed validation design; NIST AI RMF. Forensic statistical context: technical notes.

Test the system you actually deploy

TEVV: Test, Evaluation, Verification and Validation

01 TEST: exercise the failure cases

Use held-out lots, devices and operating conditions.
Challenge wrong lots, degraded reagents and expired calibration.
For tool-using agents: test injection and permission boundaries.

02 EVALUATE: quantify the trade-offs

Measure false accepts, false rejects, calibration and abstention.
Report sample sizes, confidence intervals and subgroup results.
Define acceptance criteria before inspecting the test results.

03 VERIFY: check specified requirements

Check deployed setup, schemas, controls and output bindings.
Verify material identity, certificate scope and release status.
Cryptographic verification is one part of this work.

04 VALIDATE: establish fitness for use

Assess the complete workflow in its intended setting.
Use independent review; monitor drift and failures in operation.
Reassess model, method and material changes.

Proposed test programme, informed by NIST AI RMF 1.0 / MEASURE (2023).

A confidence score is not a confidence interval

MODEL SCORE

A score of 0.97 is not automatically a 97% chance of being right.
Probability claims need calibration evidence in the relevant domain.

CONFIDENCE INTERVAL

An interval for an estimated quantity, such as detection sensitivity.
Its meaning depends on sampling, method and stated assumptions.

MEASUREMENT UNCERTAINTY

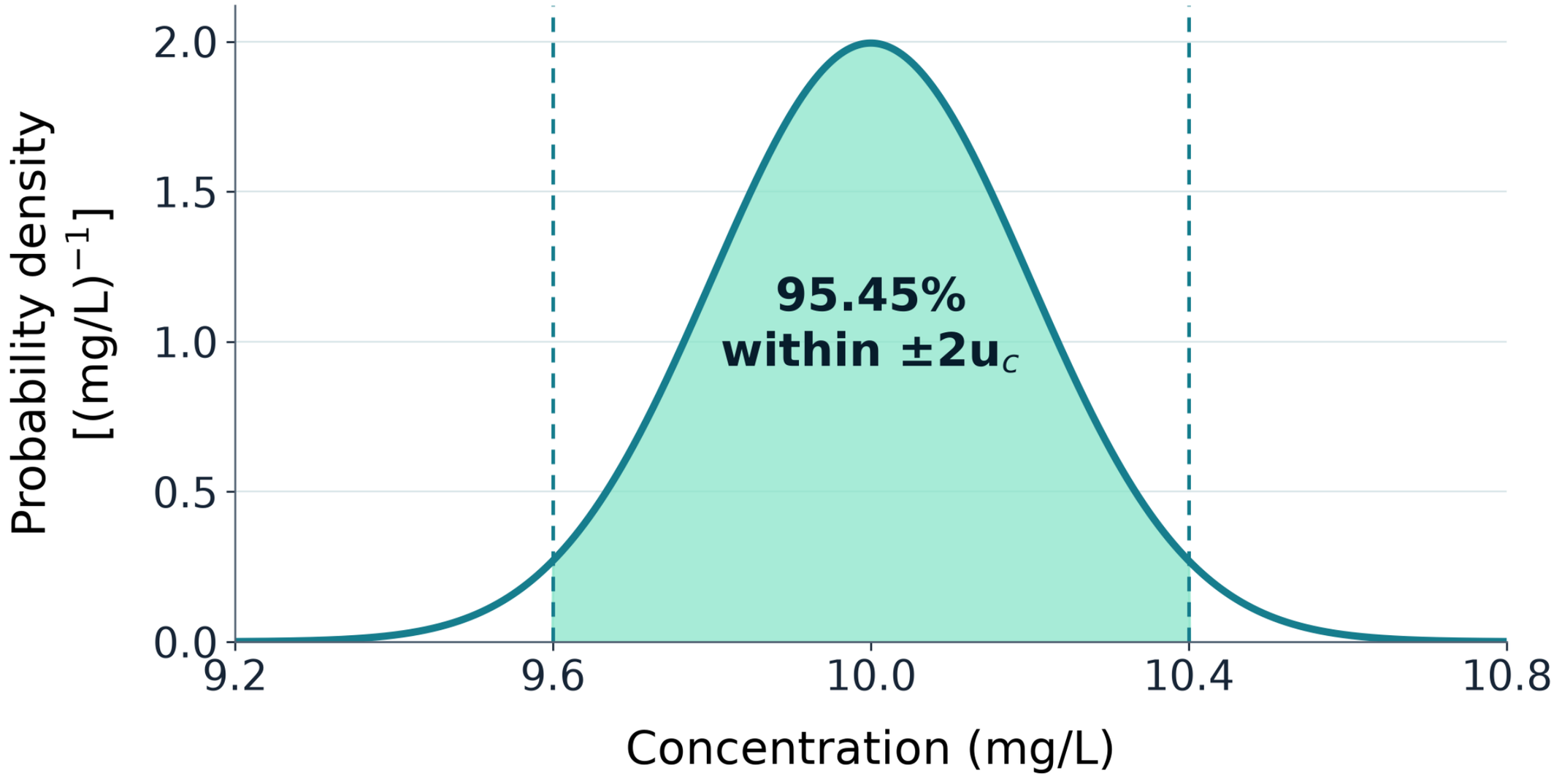
Uncertainty attached to a measured value in its physical units.
State the method; give the coverage factor for expanded uncertainty.

Name the quantity. Name the method. State the limits.

Measurement reporting: NIST TN 1297, section 7. Statistical sources in technical notes.

Show uncertainty. State the assumptions.

SYNTHETIC NORMAL (GAUSSIAN) UNCERTAINTY MODEL



10.00 ± 0.40 mg/L

Combined standard uncertainty: 0.20 mg/L. Coverage factor: k = 2.
9.60-10.40 mg/L: 95.45% coverage under the normal model.

Measurement uncertainty is not a probability of a correct match.

Illustrative calculation, not measured performance. Method: JCGM GUM, section 6.

Same 98%. Different evidence.

SYNTHETIC MARKER-DETECTION EXAMPLE

Known-positive samples correctly detected.
Two-sided 95% Wilson intervals for sensitivity.

98 / 100

93.0% - 99.4%



980 / 1,000

96.9% - 98.7%



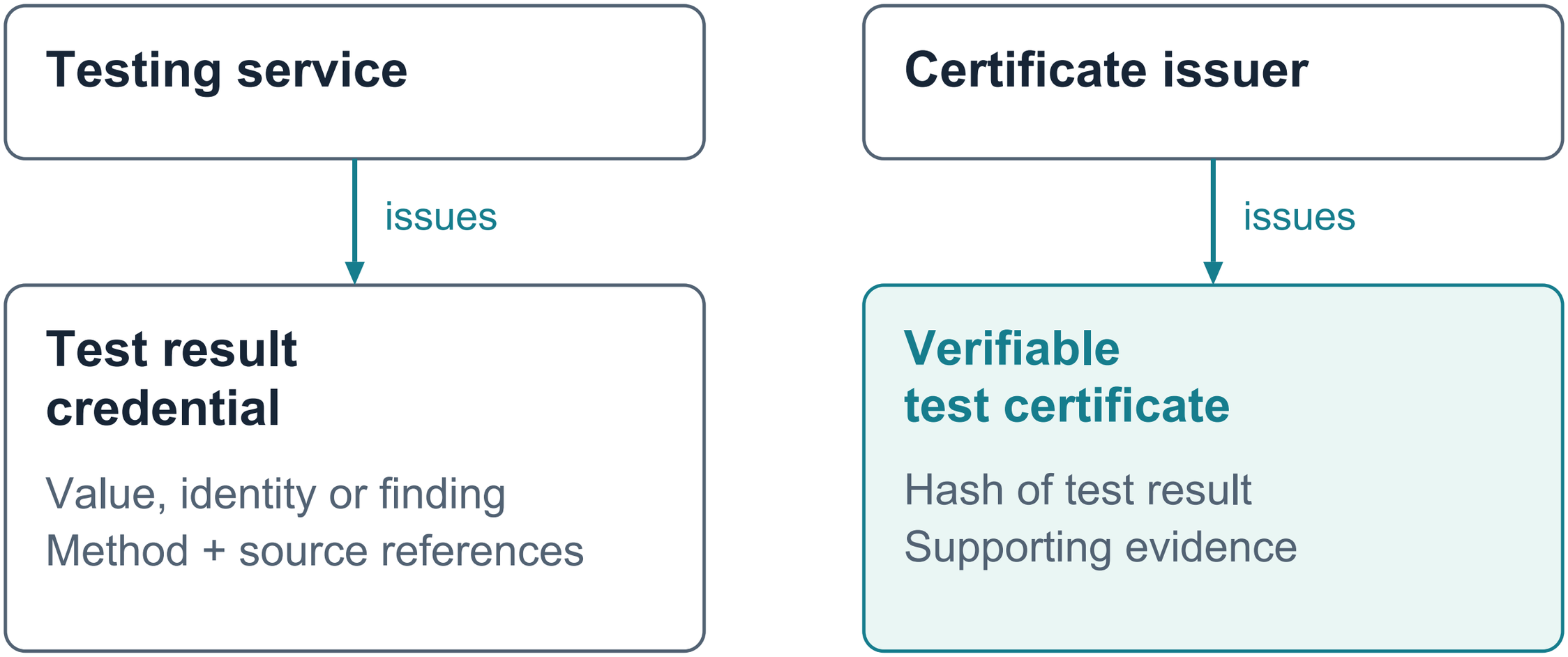
Assumes independent, representative positive samples.
Repeated scans of one lot are not independent lots.
False-positive rates require a separate negative cohort.

Neither interval gives the probability that one product is authentic.

Calculated illustration, not measured Spherity performance. Method: NIST / Wilson.

Keep the result. Explain the decision.

THE RESULT AND ITS SUPPORTING EVIDENCE



01 Preserve the test result

Retain findings, values, units and source evidence.

02 Record the interpretation

Identify the method and any AI model used.

03 Make the verdict optional

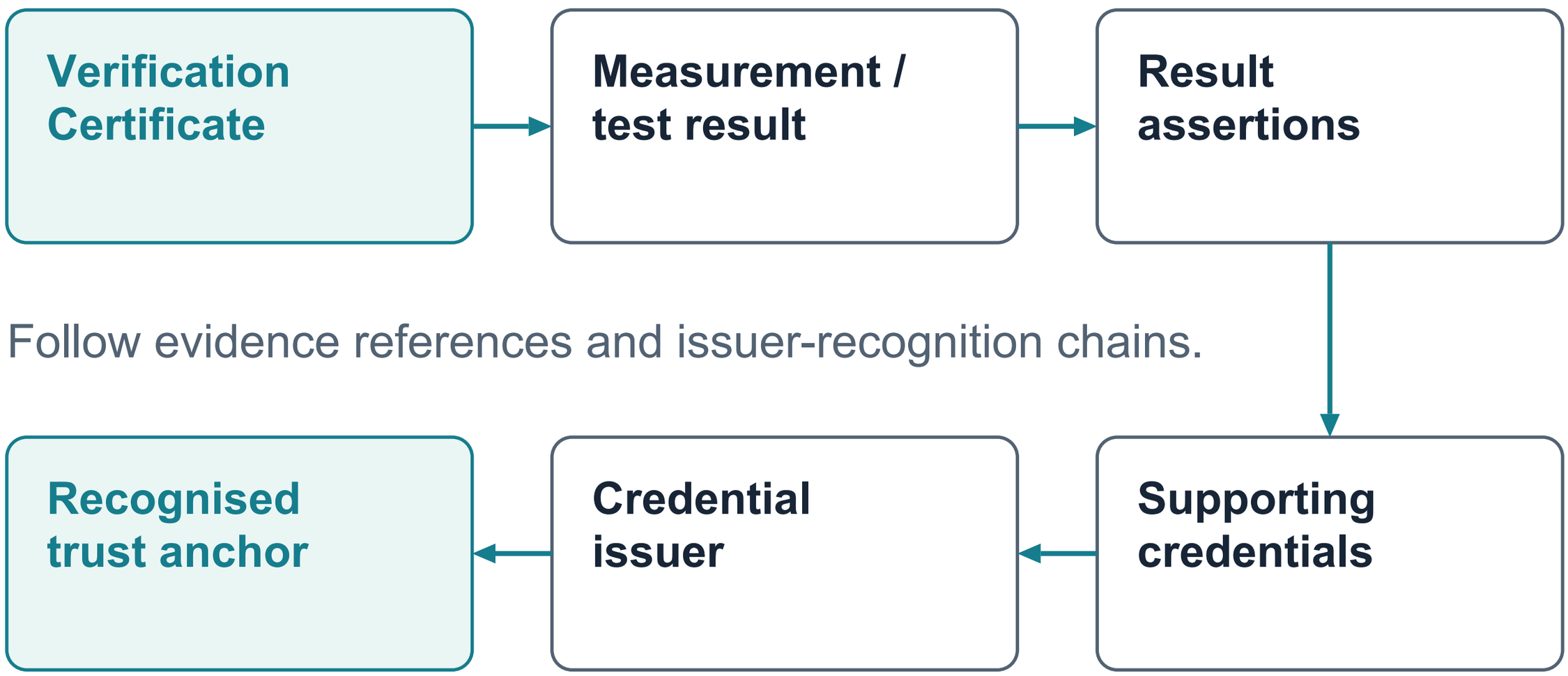
Record thresholds and policy when making a decision.

Adapted from Figure 2. Simplified view. Dr. Carsten Stöcker / CC BY 4.0

Conceptual design. Roles are illustrative. No endorsement or certificate issuance claimed.

Follow every link

Start with the certificate. Follow its evidence links.



01 **Check integrity and attribution**

Verify hashes, proofs and trusted key bindings.

02 **Check authority and scope**

Test issuer recognition, mandates and status.

03 **Keep the scientific decision separate**

A valid chain alone cannot validate the conclusion.

Adapted from Figure 3. Simplified view. Dr. Carsten Stöcker / CC BY 4.0

Conceptual design. Roles are illustrative. No endorsement or certificate issuance claimed.

VALID SIGNATURES

≠

VALID SCIENCE

CRYPTOGRAPHIC AND POLICY CHECKS

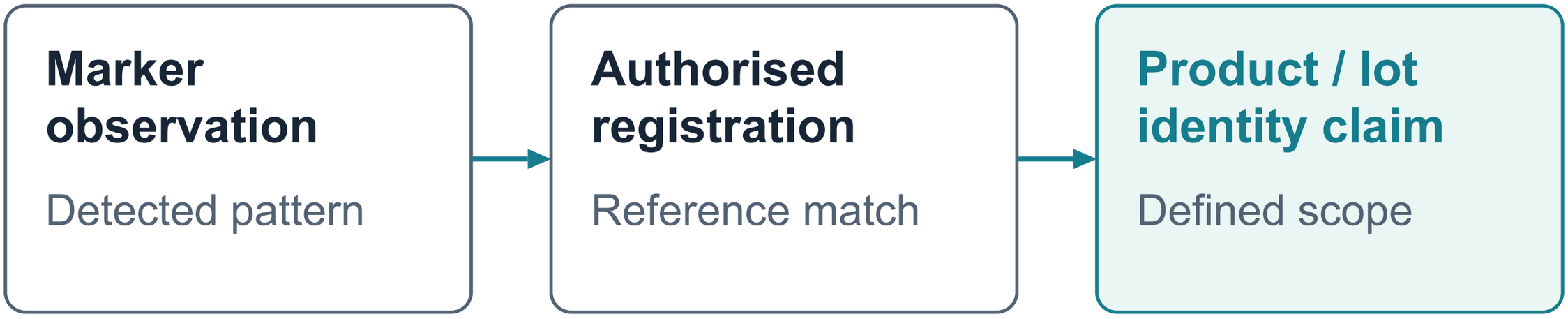
Integrity of signed evidence
Attribution under trusted key bindings
Authority within a recognised scope

SCIENTIFIC CHECKS STILL REQUIRED

Correct sample and object binding
A suitable method and calibrated instrument
Validated interpretation, including AI if used

An invisible marker needs a visible trail

ILLUSTRATIVE EXTENSION: TEXTILES AND PHARMA



Check the links behind the authenticity claim.

A validated marker match supports a scoped identity claim. Origin and custody need their own evidence.

01 **Bind the marker to the product**

Verify the registration and physical-object binding.

02 **Retain the test evidence**

Preserve the signal, device state and method.

03 **Check the provenance claim separately**

A marker match alone does not prove the full history.

A cleaning certificate should show the evidence

ILLUSTRATIVE EXTENSION: MILK-TANK CLEANING

Sensor / test service

issues

Cleaning test result

Cycle, tank and time
Values, units and method

Certificate issuer

issues

Cleaning measurement certificate

Hash of measurement
Device + calibration links

01 Record what was measured

Connect the reading to the tank and cleaning cycle.

02 Keep hygiene assessment separate

A signed reading alone does not establish cleanliness.

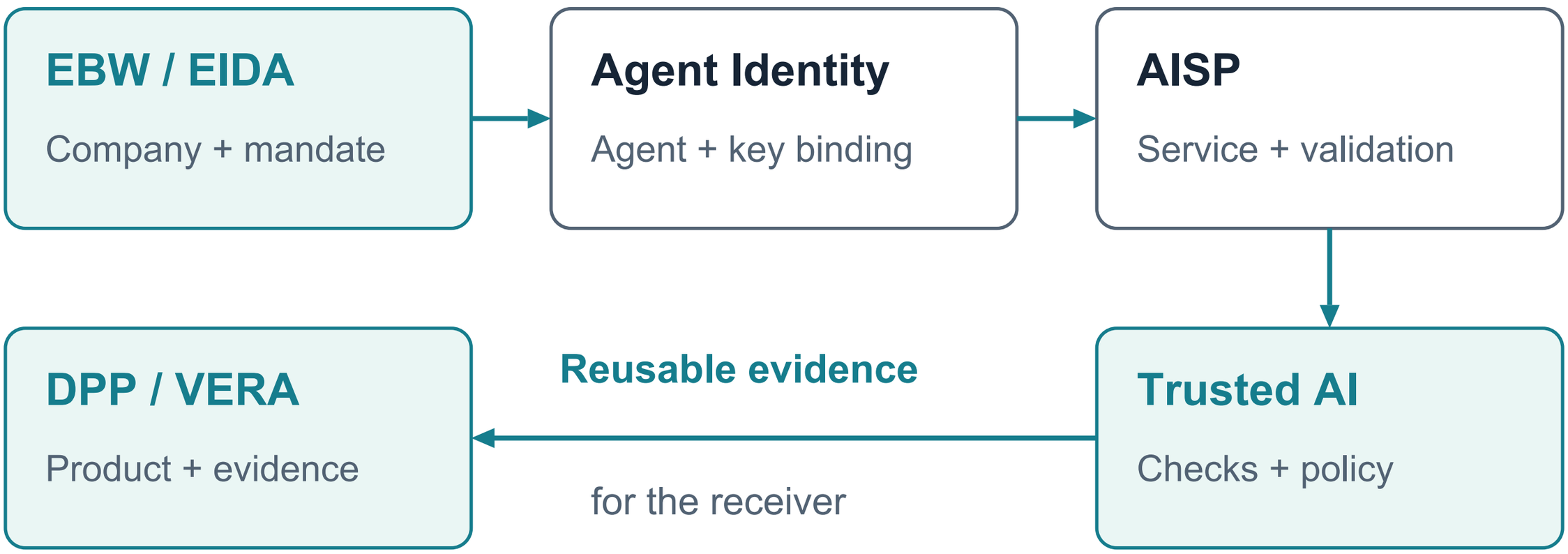
03 Let the receiver evaluate

Apply validated cleaning rules and retained evidence.

How Spherity can help

Connect products, organisations and AI to their evidence.

CONCEPT: AN INTEGRATED EVIDENCE WORKFLOW



PRODUCT / CAPABILITY	HOW IT HELPS THIS CASE
DPP / VERA	Link product and lot identity to test evidence.
EBW / EIDA	Present company identity and delegated authority.
AI Service Passport	Link service versions, profiles and TEVV evidence.
Agent Identity	Bind an agent and its keys to its organisation.
Trusted AI	Apply evidence checks, access policy and audit trails.

All our products are based on open standards.

Active in W3C: Verifiable Credentials and DPP / Business Wallet work.

Labs and assessors supply scientific assurance. Integration is case-specific.

Product context: spherity.com / VERA, EIDA and Agentic Trust

W3C participation: Verifiable Credentials Working Group

Does your test result carry its evidence?

EXPLORE THE WORKED AI CASE

Verifiable AI Lab Results: From Scan to Certificate

Explore the architecture.
Challenge the assumptions.
Test the failure cases.

READ THE CASE STUDY

spherity.github.io/spherity-research/case-studies/

Scan the code or tap this page.

Proposed architecture. No deployment or measured outcomes reported.