IMEN891K / IDSC501 / AXGS-501 · DATA ENGINEERING · FALL 2026

2.

Data Collection

Sangwoo Mo · POSTECH

Industrial & Management Engineering

Where We Are

Collection and preparation shape the data that later stages can use



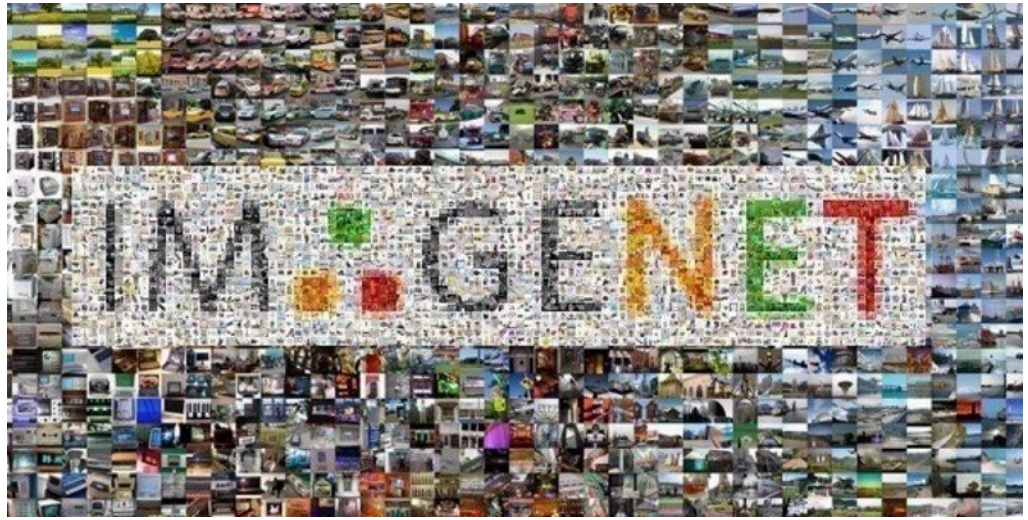
every later stage can only use what you collect here

You have a problem worth solving
You have no data

Now what?

Data for Your Task

A public dataset can help, but it cannot supply your missing observations



Machine inspection

Which parts need inspection?

Inputs: vibration, current, tool

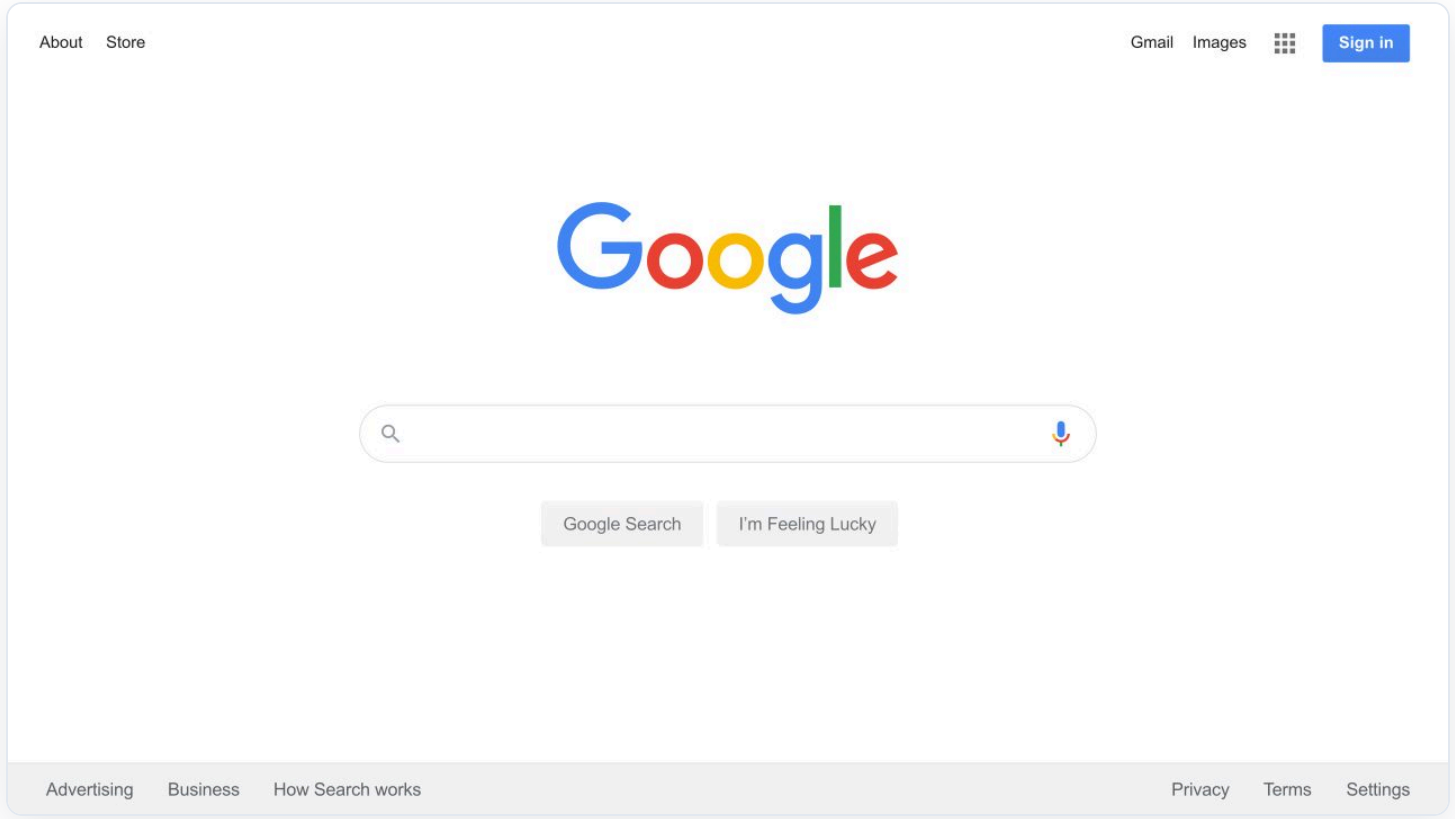
Label: the inspection result

Collect both for the same part

A useful dataset must connect the input to the outcome

The record is the competence

Google: queries and clicks provide feedback that an algorithm alone cannot supply



Search use generates **new queries and feedback**

The record is the moat

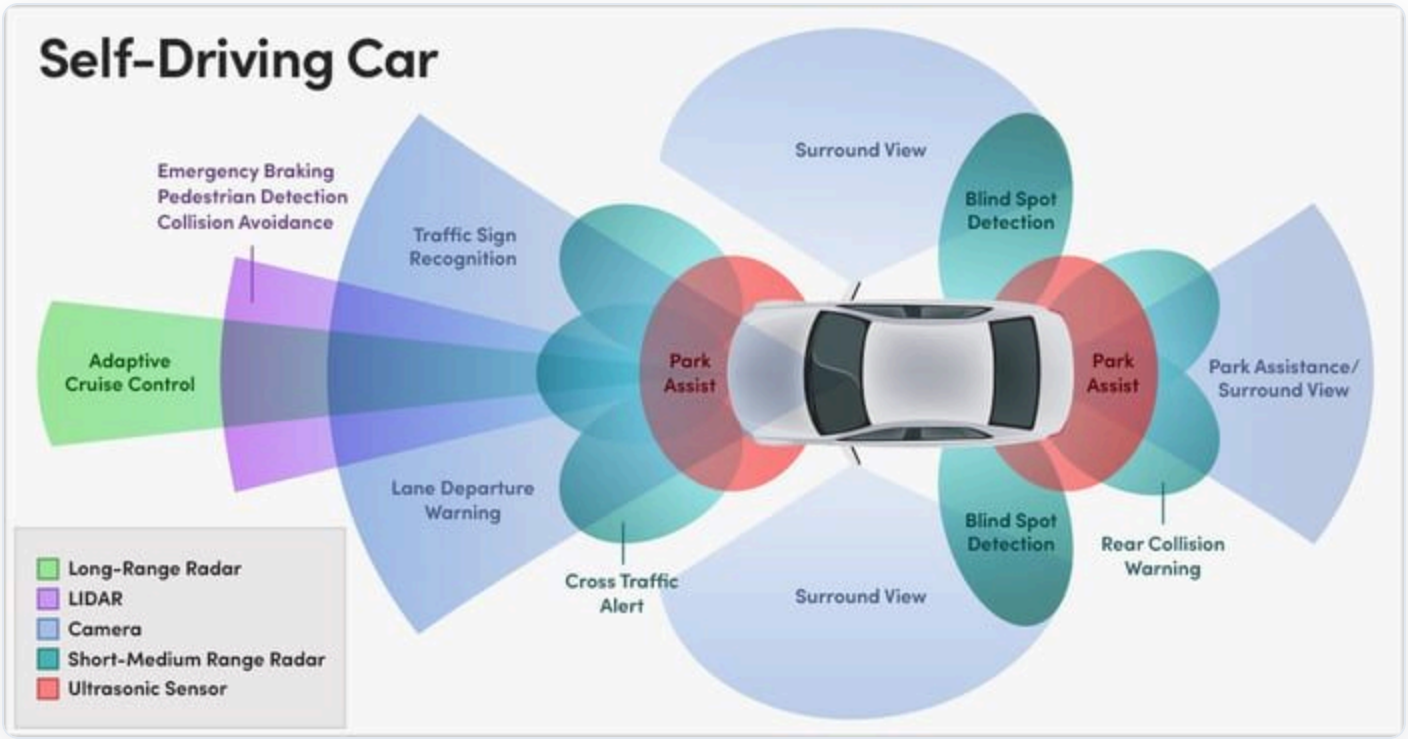
QANDA: product use generates observations that a downloaded dataset does not contain



Each student photo records a moment of difficulty and creates another observation

The record is the strategy

Sensor choices determine which aspects of the world become data



Each sensor captures different information

Case study: Waymo vs Tesla

Different sensor designs produce different training evidence

Collection choice	Waymo	Tesla Vision
Sensors	Cameras + lidar + radar	Camera-based perception
Distance evidence	Lidar range measurements	Depth estimated from images
Data work	Align readings across sensors	Train and validate visual depth

A measurement and a model estimate need different checks

Outline: Data Collection

01

Define the Record

What does one row hold?

`one row = one part`

02

Build the Collection Path

Where does each value come from?

`reuse, observe, generate`

03

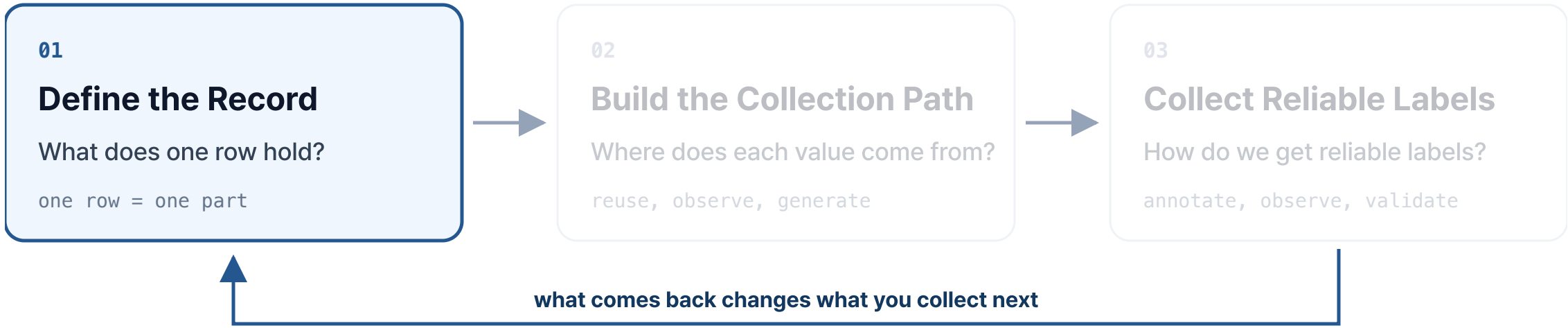
Collect Reliable Labels

How do we get reliable labels?

`annotate, observe, validate`



Define the Record



Record Design

Define the meaning, choose the representation, and set the permitted use

Meaning

Measurement

Definition and measurement procedure

Representation

Record structure

Unit, resolution and metadata

Use

Boundaries

Purpose and identifiability

Label Definitions

Different definitions create different labels for the same parts

Reader	Defect definition	Same 400 parts
QC inspector	Outside drawing tolerance	40
Line supervisor	Customer returned the part	12
Sales	Complaint call received	3

A model learns the label definition used to build its dataset

Constructs and Measures

Operationalization turns a concept into a measurement procedure

Construct

Defect

The property we want to identify

→

Operationalization

Inspection rule v3

Measure against drawing tolerances

→

Measure

Pass / defect

The recorded inspection result

Validity

Does the procedure measure the intended property?

Reliability

Do repeated measurements agree?

Agreement alone does not show that the right property was measured

Unit of Analysis

Choose what one row represents before choosing its columns

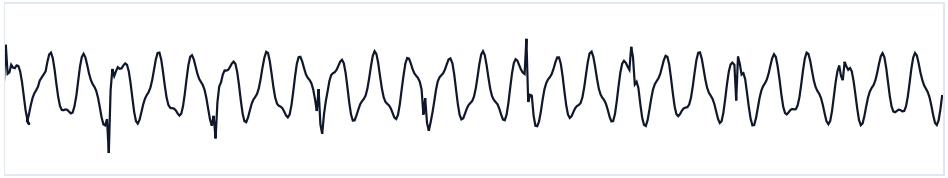
One row represents	Useful for	Can identify a defective part?
One part	**Part inspection**	**Yes, with part_id**
One machine-minute	Machine monitoring	Only with an additional part-level link
One work order	Yield reporting	No individual part result

The task determines the grain of the record

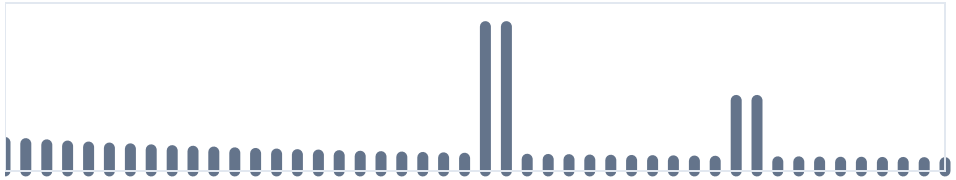
Measurement Resolution

A summary can discard information that a later task needs

20 kHz raw waveform



spectrum



bearing fault visible

RMS

over 1 min

1-minute RMS (root mean square)

spectrum



fault frequency lost

Context Around the Value

Keep the unit, instrument and event time with the value

4.7

4.7 of what? when? measured how?

Metadata makes the measurement interpretable and reusable

Context Around the Value

Keep the unit, instrument and event time with the value



Metadata makes the measurement interpretable and reusable

Dataset Documentation

A dataset needs enough context for someone else to interpret and use it

Dataset: part inspections	Document with the data
Meaning	One row per part; defect defined by rule v3
Collection	Which machines, dates and inspection policy
Use and limits	Inspection purpose; uninspected parts have no label
Maintenance	Named owner; update and guideline versions

A datasheet records choices that column names cannot explain

Demo: Reading a Record

A measurement needs an entity, a time, a unit and a rule

Field	Recorded value
part_id	P-104
measured_at	09:14:02
value	38.6
unit	missing
source	CNC-02
label	defect

1. Which fields are missing?
2. Can two inspectors reproduce this label?

Answer: A Reproducible Record

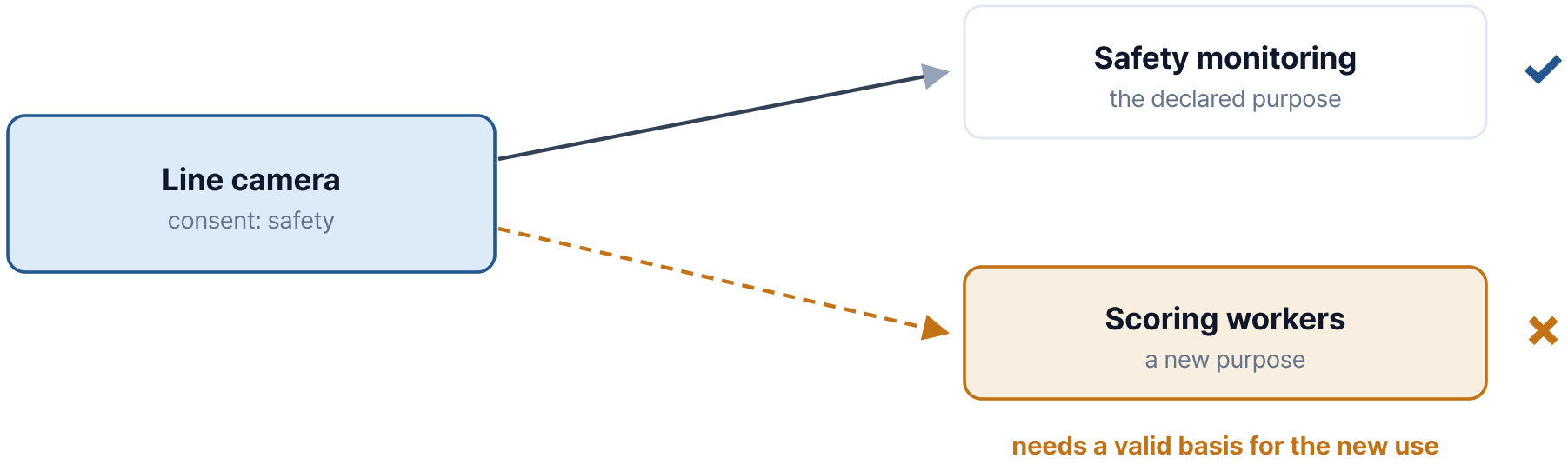
Store the measurement meaning and the labeling rule

Field	Recorded value
part_id / measured_at	P-104 / 09:14:02
measurement / value / unit	cycle_time / 38.6 / s
source	CNC-02 / cycle log
label / rule_version	defect / inspection-v3

The value becomes interpretable when its context is recorded

Privacy: Purpose Limitation

Specify the purpose and legal basis before collecting personal data



Failure: safety cameras on the line, later used to score workers. Consent covered safety, not scoring

Caveat: Anonymization

Removing names does not prevent identification through other fields

Released dataset · name removed

ZIP	Birth date	Sex	Diagnosis
37673	1965-07-31	F	Hypertension
37673	1971-02-14	M	Asthma
37675	1965-03-02	F	Diabetes

Caveat: Anonymization

Removing names does not prevent identification through other fields

Released dataset · name removed

ZIP	Birth date	Sex	Diagnosis
37673	1965-07-31	F	Hypertension
37673	1971-02-14	M	Asthma
37675	1965-03-02	F	Diabetes



join

Public list · has names

Name	ZIP	Birth date	Sex
J. Kim	37673	1965-07-31	F

J. Kim has hypertension

Caveat: Anonymization

Removing names does not prevent identification through other fields

Released dataset · name removed

ZIP	Birth date	Sex	Diagnosis
37673	1965-07-31	F	Hypertension
37673	1971-02-14	M	Asthma
37675	1965-03-02	F	Diabetes



join

Public list · has names

Name	ZIP	Birth date	Sex
J. Kim	37673	1965-07-31	F

J. Kim has hypertension

k-anonymity

Every (ZIP, birth date, sex) combination must appear in at least k rows

Generalize to reach it: ZIP 376**, birth year 1960s. Each step costs precision

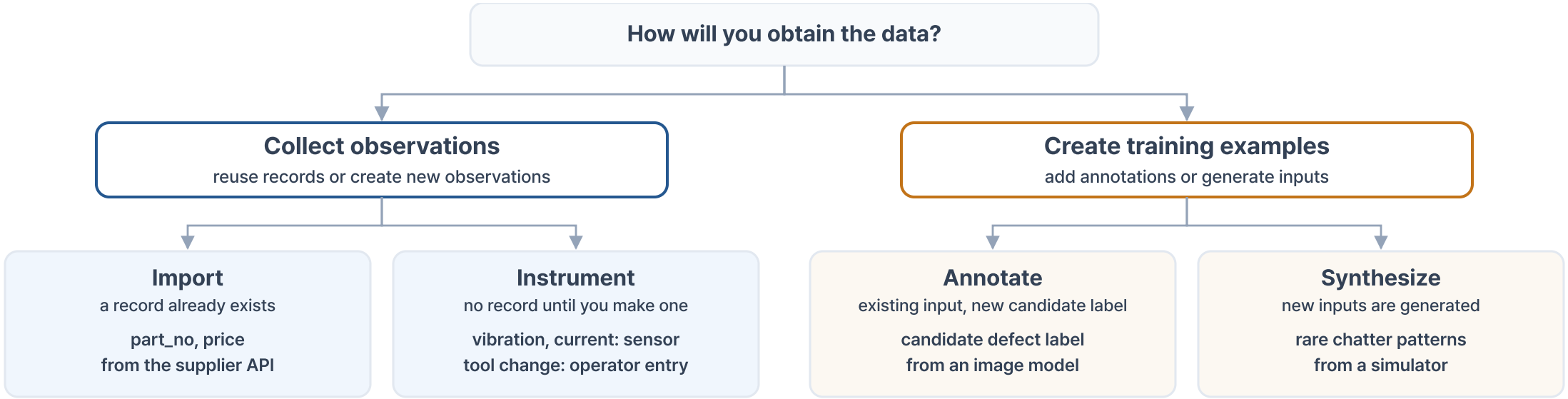
Failure: drop the name column and call it anonymous. A public list with matching fields can join it back

Build the Collection Path



Data Acquisition

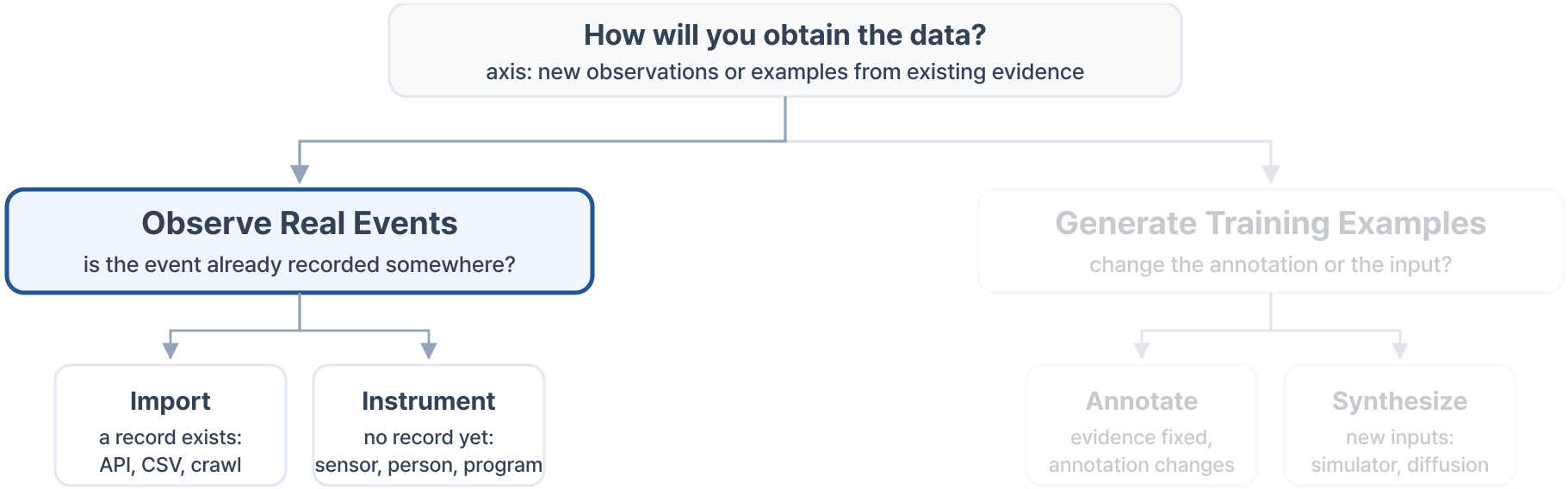
Collect observations, or create training examples from existing evidence



Keep the source; keep the original when adding labels or generating variants

Observe Real Events

02 · BUILD THE COLLECTION PATH



Existing Records

Start with the system that already records the event

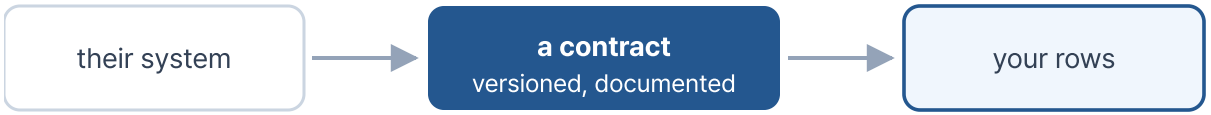
Source	Record to reuse	Access route
Supplier system	Part specification, price	File export or API
Production system	Part, tool, program, operator	Database view or API
Machine controller	Signals and alarm events	Log file or event stream

The source tells you what was recorded; the interface tells you how to retrieve it

API or Crawler

Prefer an explicit interface; page extraction also depends on the page layout

API



An explicit interface can be versioned and checked

CRAWLER



A redesign breaks extraction without an HTTP error

Demo: Crawling a Supplier Portal

Extract prices, change the page layout, then repair the selector

1. Does a successful request guarantee extracted rows? 2. Repair the selector after redesign

```
# extraction from one fetched page
r    = requests.get("https://supplier-portal.example.com/products?page=1")
rows = BeautifulSoup(r.text).select(".product .price")
```

Selector `.product .price` Run extraction robots.txt rate limit Site redesign

```
GET /products?page=1

<div class="listing" data-page="1">
  <div class="product">
    <span class="name">Bearing 6205-2RS</span>
    <span class="price">12,400</span>
    <span class="rating">4.6</span>
  </div>
  <!-- 5 more products -->
</div>
Preview: first product of page 1. Run extraction for all 3 pages.
```

PAGE	CLASS	EXTRACTED TEXT
1	price	12,400
1	price	268,000
1	price	431,000
1	price	8,900
1	price	154,000

Answer: Detect a Broken Extraction

A successful request does not guarantee usable records

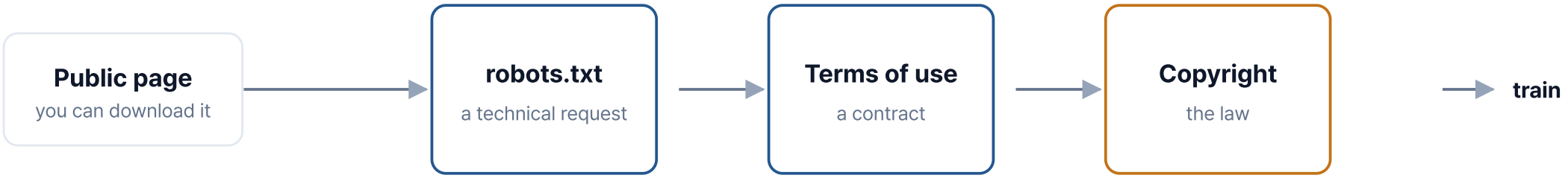
Run	Response	Extracted prices	Action
Before redesign	200 OK	18	Validate fields
After redesign	200 OK	0	Investigate the drop
Selector repaired	200 OK	18	Compare with the last good run

`.product .price` → inspect HTML → `.product .pdp-price__amount`

Check counts, required fields and types before accepting the run

Copyright: Three Gates

Access, site terms and rights to reuse are separate checks



Passing one gate does not pass the others

Training rights depend on the data, use and jurisdiction

Failure: a crawled dataset with no documented basis for use. Record the license or the exception you rely on

Creating New Observations

An event becomes data only when something records it

Software

Alarm acknowledged

Instrument the button press

Sensor

Vibration measured

Mount and calibrate the sensor

Person

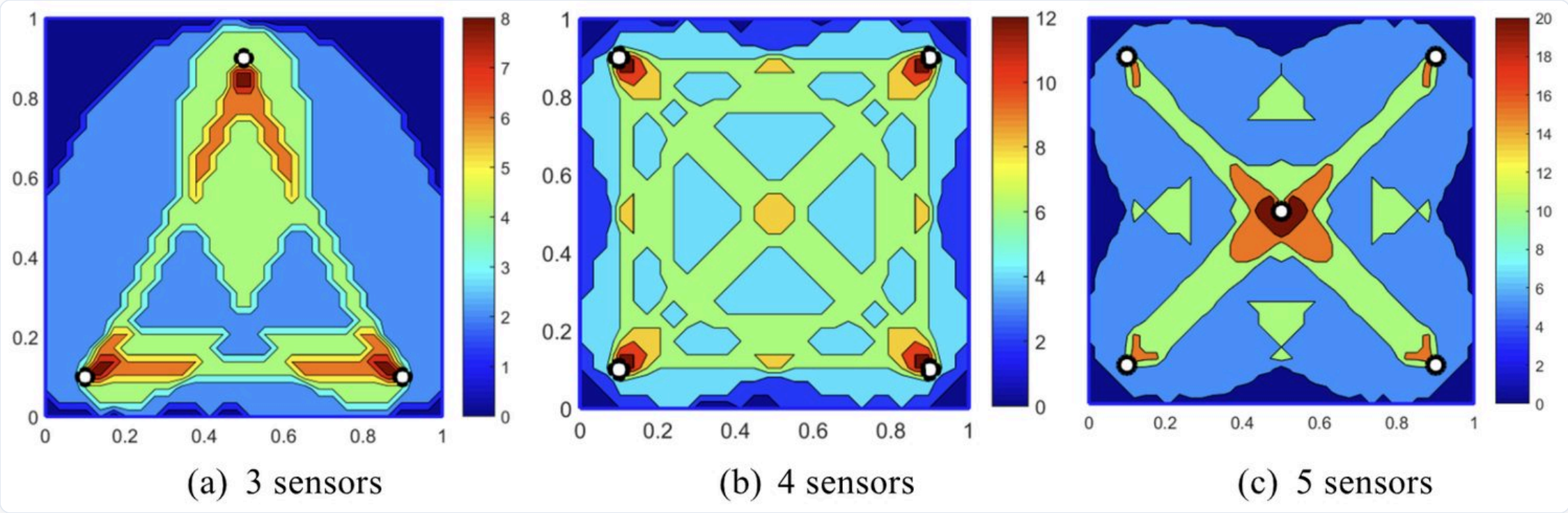
Defect judged

Apply the inspection guideline

Logging needs instrumentation; physical measurements and judgments need it too

Sensor Placement

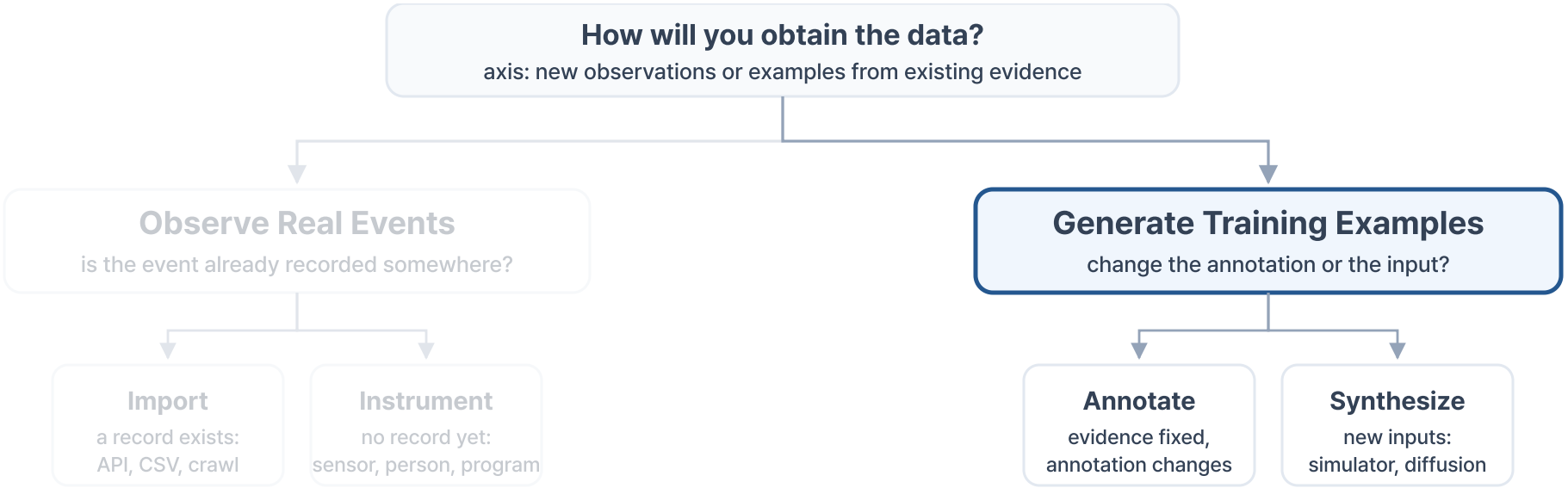
Sensor coverage depends on placement as well as the number of sensors



Going from 3 to 5 sensors changes **what is detectable**, not just the accuracy

Generate Training Examples

02 · BUILD THE COLLECTION PATH



Weak Supervision

Labeling functions provide noisy votes that can agree, conflict or abstain

Part	Vibration rule	Return report	Image model	Votes
4471	defect	—	defect	agree
4472	—	—	pass	one source
4473	defect	—	pass	**conflict**
4474	—	—	—	**uncovered**

```
Labeling functions → label model → candidate training labels
```

Coverage: 3 of 4 parts receive a vote. Check conflicts on reviewed examples

Synthesizing the Rare Class

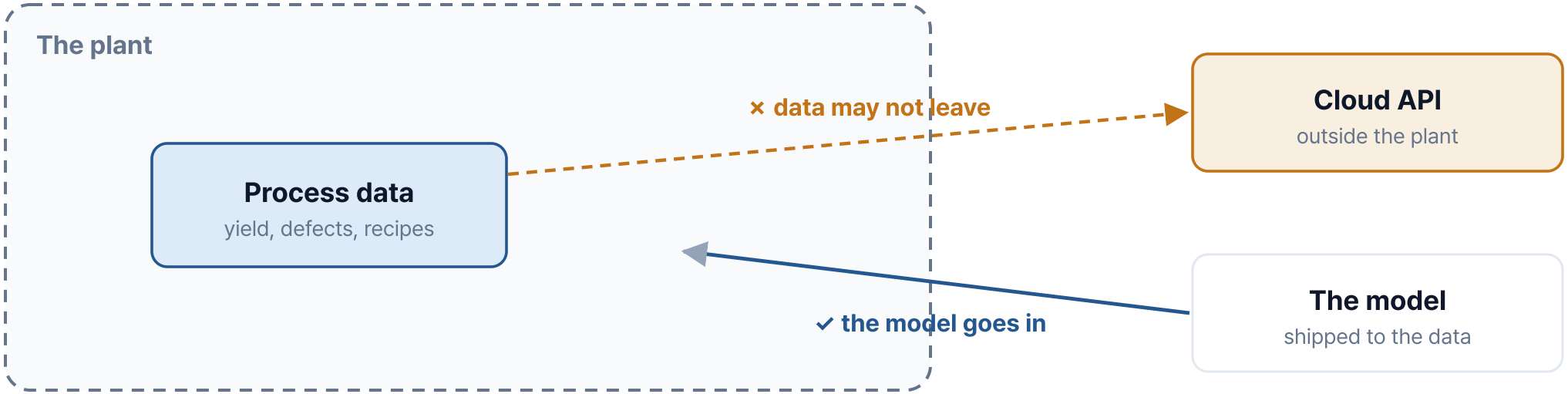
Augmentation changes an input while preserving its label; generation proposes new examples



Variations **supplement scarce real examples**; their realism and value still need testing

Confidentiality: Trade Secrets

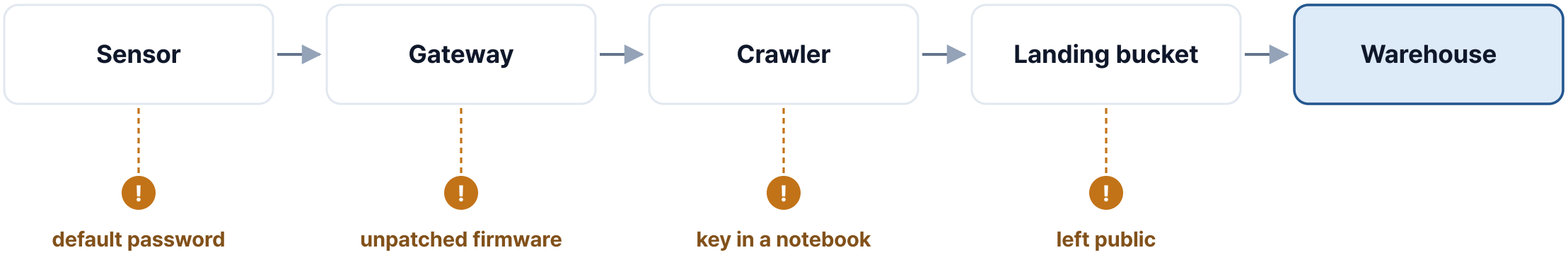
Check whether the data may leave the organization before choosing a cloud service



Failure: pipeline built on a cloud service, then the confidentiality agreement forbids upload. Rebuild inside, or drop that route

Security: Attack Surface

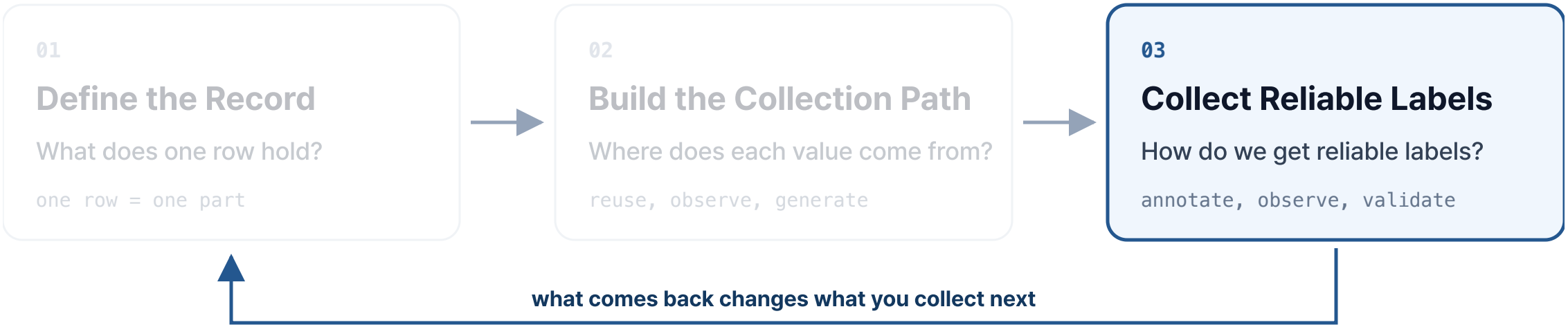
Limit access at every point where data is collected, transferred or stored



Each one needs encryption, least privilege and a named owner

Failure: a bucket left public for a demo, indexed by a search engine before anyone notices

Collect Reliable Labels



Sources of Labels

A label can come from deliberate annotation or from an observed outcome

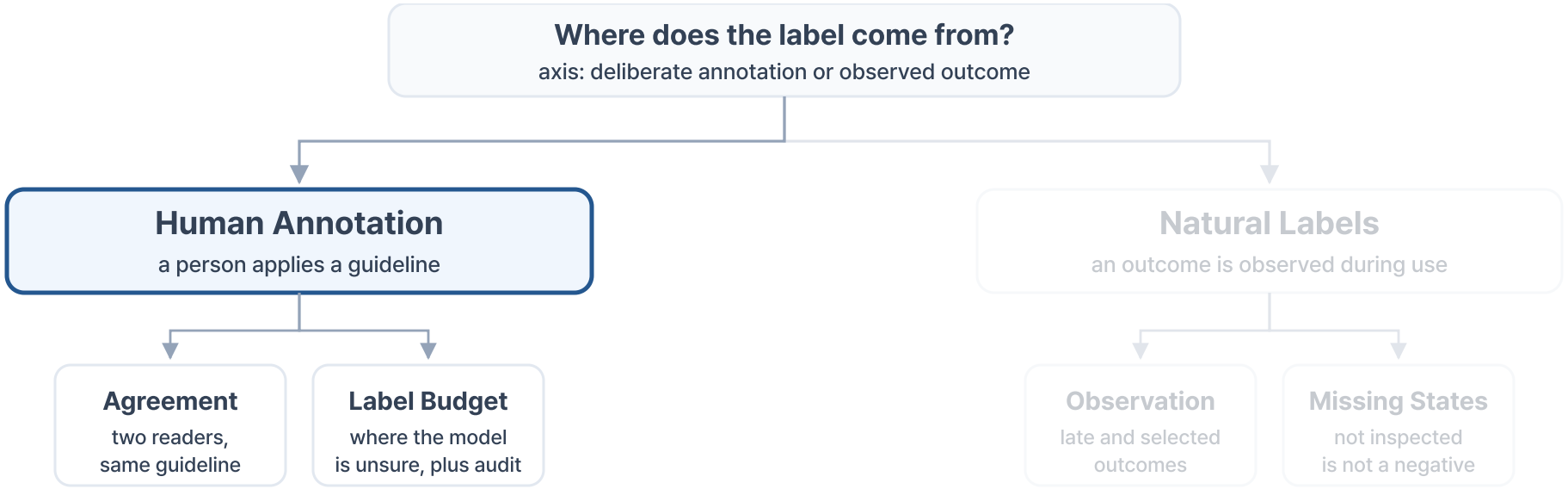
Source	Example	Main collection problem
Human annotation	An inspector judges a part	Make the guideline reproducible
Natural label	A delivered shipment has an arrival time	Record and link the outcome

Model-generated labels remain candidates until checked

Both sources need a definition and a quality check

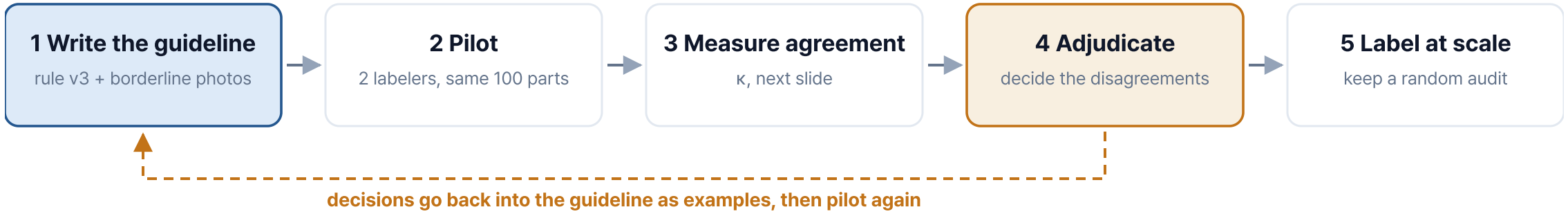
Human Annotation

03 · COLLECT RELIABLE LABELS



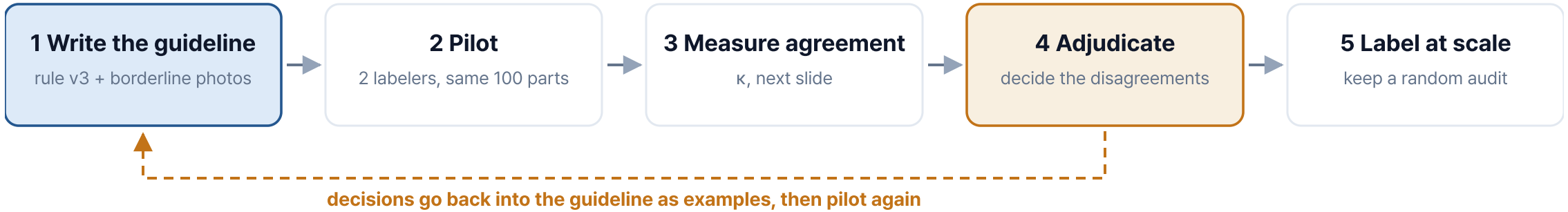
Running a Labeling Project

Pilot the guideline, resolve disagreements, then expand the labeling effort



Running a Labeling Project

Pilot the guideline, resolve disagreements, then expand the labeling effort



IN-HOUSE EXPERTS

Use domain expertise for ambiguous cases and audits

VENDOR OR CROWD

Use when the task can be taught with clear examples

TOOLING

Label Studio / CVAT: configure labeler, time and guideline tracking

Demo: Inspector Agreement

Two inspectors each mark five defects among the same 100 parts

	B: defect	B: pass
A: defect	1	4
A: pass	4	91

1. What is the overall agreement?
2. How many defects did both inspectors identify?

Answer: Inspector Agreement

Cohen’s kappa corrects agreement for chance matches

	B: DEFECT	B: PASS
A: defect	**1**	4
A: pass	4	**91**

$$p_o = (1 + 91) / 100 = 0.92$$ observed

Two inspectors, 100 parts. Each calls 5 defects

Answer: Inspector Agreement

Cohen’s kappa corrects agreement for chance matches

	B: DEFECT	B: PASS
A: defect	**1**	4
A: pass	4	**91**

Two inspectors, 100 parts. Each calls 5 defects

$p_o = (1 + 91) / 100 = 0.92$ observed

$p_e = 0.05 \cdot 0.05 + 0.95 \cdot 0.95 = 0.905$ by chance

Answer: Inspector Agreement

Cohen’s kappa corrects agreement for chance matches

	B: DEFECT	B: PASS
A: defect	**1**	4
A: pass	4	**91**

Two inspectors, 100 parts. Each calls 5 defects

$$\kappa = (p_o - p_e) / (1 - p_e)$$

$p_o = (1 + 91) / 100 = 0.92$ observed

$p_e = 0.05 \cdot 0.05 + 0.95 \cdot 0.95 = 0.905$ by chance

$\kappa = 0.015 / 0.095 = \mathbf{0.16}$

One shared defect. **Adjudicate, revise the rule, measure again**

Active Learning: Uncertainty Sampling

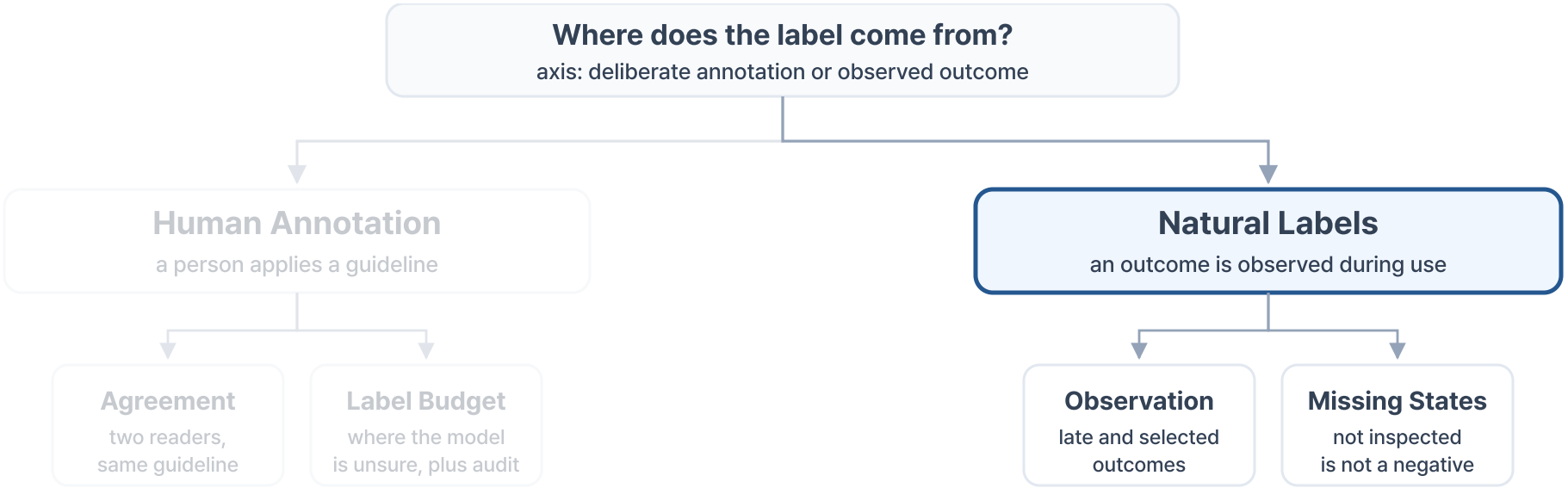
Target uncertain cases for learning; audit independently to find blind spots

Part	p(defect)	What a new label tells us
4471	0.99	Model is confident; it can still be wrong
4472	0.50	Model is uncertain; request a label
4473	0.80	Less uncertain under this model
4474	0.02	A random audit can check confident predictions

Model-guided labeling and random auditing answer different questions

Natural Labels

03 · COLLECT RELIABLE LABELS



Predictions Without Outcomes

Prediction logs alone cannot tell you whether the predictions were useful

Inspection recommendations, 12 months

date	part_id	score	recommendation	outcome
2025-10-03	3471	0.91	**inspect**	—
2025-10-03	4480	0.88	**inspect**	—
2025-11-12	5102	0.93	**inspect**	—
2026-01-20	6210	0.86	**inspect**	—
2026-03-04	7004	0.90	**inspect**	—
2026-06-17	8831	0.95	**inspect**	—
2026-09-30	9916	0.89	**inspect**	—

... 2,140 rows

The operator screen

ALARM part 9916 score 0.89

recommend: inspect

No button to record what happened next

The operator inspects, or does not

The part is fine, or it is not

Nothing writes that back

A year later, this log cannot say if the model works

Feedback Loop Length

An outcome can arrive after the prediction record has already been stored

Decision recorded

part_id: P-104

09:14 · inspect

→

Outcome pending

Inspection not finished

No label yet

→

Outcome linked

09:20 · pass

label = 0

Other examples	Possible feedback delay
Recommendation → click	Minutes
Inspection → result	Days
Fraud → dispute	Weeks

A pending outcome is not a negative outcome

Selected Outcomes

The inspection policy controls which outcomes become labels

100 produced parts

20 selected for inspection

Labels will be observed

→

80 shipped without inspection

Outcomes unknown

They are not recorded as pass

→

Random audit of 5

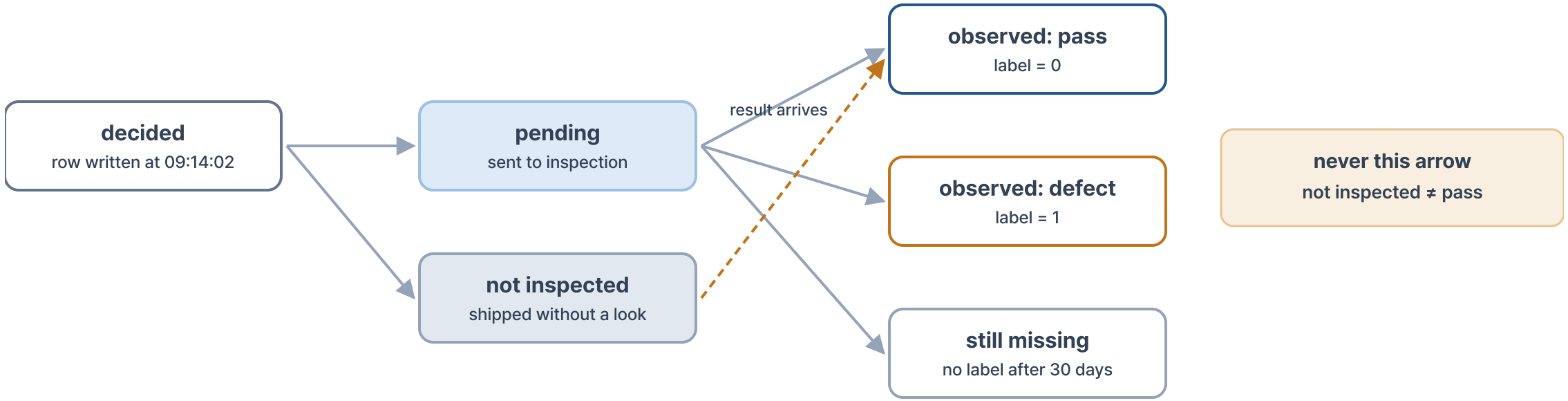
25 labels in total

75 outcomes remain unobserved

Audit outside the model-selected sample to observe what it missed

Missing Outcomes

Keep pending, unobserved and negative outcomes distinct



A model trained on "label or 0" learns that skipped parts are good

Linking by Identity and Time

The same identifier is necessary; the event sequence still matters

Earlier observation

```
P-104 · 08:30
defect

Not this decision's
outcome
```

→

Decision

```
P-104 · 09:14
inspect
```

→

Later outcome

```
P-104 · 09:20
pass

Match identity and
time
```

A prior inspection cannot be the outcome of a later decision

Demo: Linking Inspection Outcomes

Use identifiers and event times instead of row order

Predictions

part_id	decided_at	decision
P-104	09:14	inspect
P-105	09:15	ship
P-106	09:16	inspect

Inspection log

part_id	inspected_at	result
P-106	09:22	pending
P-104	09:20	pass
P-104	08:30	defect

1. Which inspection belongs to P-104's decision?
2. Does P-105 have a negative label?

Answer: Observed and Missing Labels

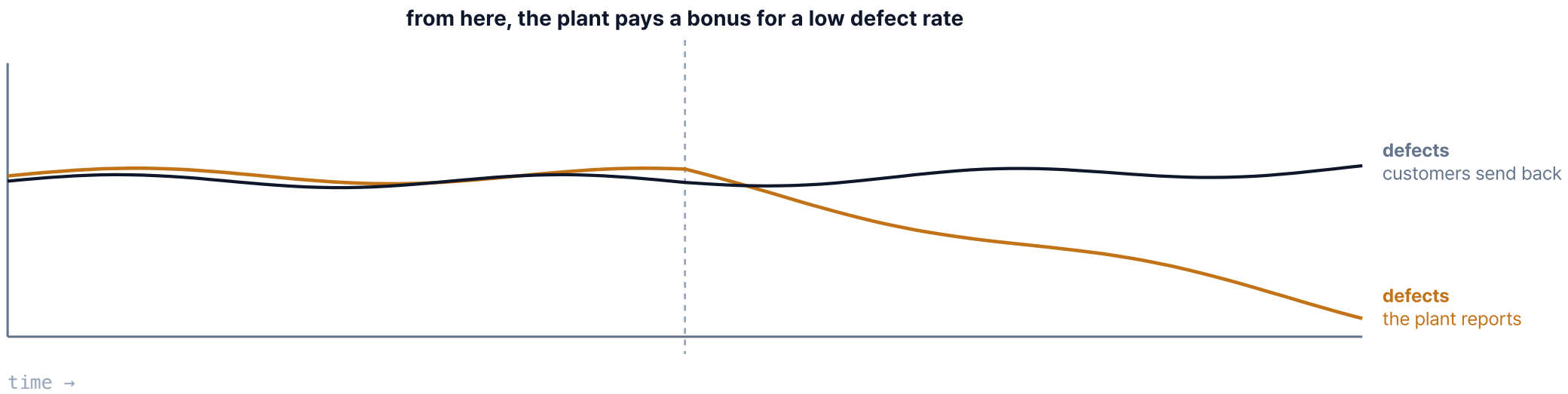
Policy: link the first inspection episode after each decision

part_id	linked inspection	state	label
P-104	09:20	observed pass	0
P-105	none	not inspected	NULL
P-106	09:22	pending	NULL

Missing feedback is a state, not a negative label

Goodhart’s Law

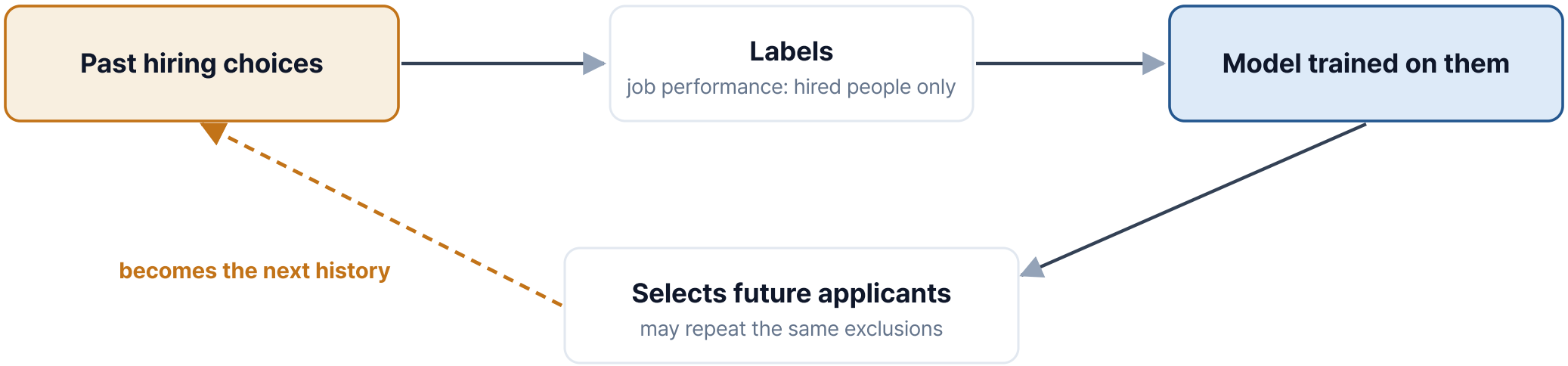
A reward tied to a reported metric can change how people report it



Reported defects fall while returns stay flat; audit whether reporting behavior changed

Fairness: Historical Selection

Labels reflect past access and decisions, not only the property we want to predict



Job performance is unobserved for rejected applicants; rejection is not evidence of poor performance

Closing the Loop

Use verified outcomes and audits to improve the next collection cycle



Takeaway: Data Collection

Define what to observe, obtain the evidence, and check the labels

01	**Define the Record**	Define the unit, measurement and permitted use
02	**Build the Collection Path**	Reuse records, instrument events, or generate examples
03	**Collect Reliable Labels**	Check agreement, delayed outcomes and selection bias