MODULE 1 · SESSION 2

Establish Synthetic Data Trust

How to trace where data came from, and keep synthetic evidence clearly separated from evidence about the real world.

Theory session

Beginner

Running case: LumaForge Robotics

ORIENTATION

Where this session sits

1.1 set the architecture; 1.2 asks what the data flowing through it is worth

MODULE 1
AI System Foundations

MODULE 2
Controlled AI Operations

MODULE 3
Strategic AI Operating Choices

SESSIONS IN THIS MODULE

1.1 Architect a Multi-Model Product
How specialised models cooperate inside one product

1.2 **Establish Synthetic Data Trust** — YOU ARE HERE
Traceable data, and knowing which evidence is real

1.3 Route Models Across Workloads
Choosing a model by task, quality, latency and cost

SESSION OUTCOMES

What this session explains

Knowing where data came from, and what it can prove

01 **Provenance answers four questions**

Where did this come from, what was done to it, who may use it, and what does it prove?

02 **Synthetic data is a tool, not a substitute**

It can fill gaps in training. It cannot be evidence about the real world.

03 **Silent errors are the expensive ones**

A mismatched name or a wrong address looks configured and is not.

04 **One description beats many local fixes**

A single source of truth is what made the warehouse deployment reliable.

DEFINITION

What provenance means

Borrowed from art dealing, and used the same way

WORKING DEFINITION

Provenance is the recorded history of a piece of data: where it originated, every transformation applied to it, who is permitted to use it, and therefore what claims it can legitimately support.

1

Origin

A real sensor reading, a human label, a simulation, or another model's output.

2

Transformation

Every filter, join, relabel and augmentation, in order — because each one can change what the data means.

3

Permission

What it may be used for. Consent and licence travel with the data or they are lost.

Moreau, L. & Groth, P. (2013). Provenance: An Introduction to PROV. San Rafael: Morgan & Claypool.

CONCEPT 1 · EVIDENCE

Not all data is equally good evidence

A simple ladder for deciding what a dataset can prove

REAL, FROM THE OPERATING ENVIRONMENT

Collected where the product actually runs. The only thing that proves real-world performance.

REAL, FROM A DIFFERENT SETTING

Genuine data, but from another site or period. Useful, and needs a stated assumption.

AUGMENTED REAL DATA

Real examples with controlled variation added. Stretches coverage; inherits the original's limits.

FULLY SYNTHETIC

Generated. Excellent for rare cases and privacy; proves nothing about reality on its own.

THE RULE

Never evaluate on synthetic alone

Train on synthetic if you must. Judge performance on real data from the environment you will operate in.

THE WAREHOUSE EXAMPLE

Real on-site labels

The vision model in Session 1.1 was trained on actual operational data labelled with grasping order and placement.

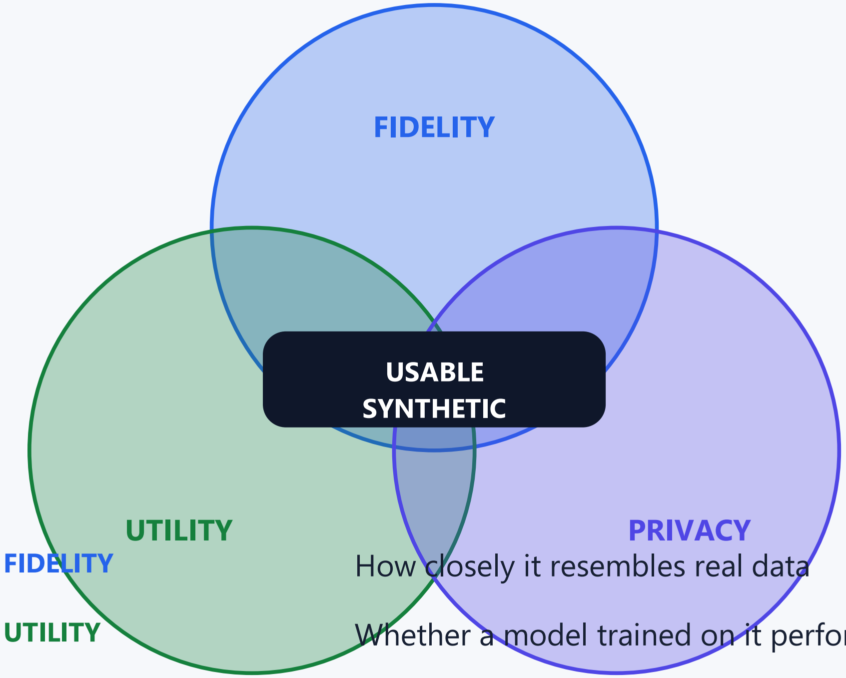
THE PRACTICAL HABIT

Tag every row with its rung

If a dataset cannot tell you which rung each row sits on, you cannot defend any claim made from it.

Three things you want, and can rarely maximise together

Fidelity, usefulness and privacy pull against each other



- **FIDELITY** How closely it resembles real data
- **UTILITY** Whether a model trained on it performs
- **PRIVACY** That no individual can be recovered from it

THE TRADE-OFF

Perfect fidelity is a privacy risk

Data that reproduces the original closely enough can leak the individuals it was built from.

WHAT TO MEASURE

Utility, not resemblance

Does a model trained on it perform on real held-out data? That is the only question that matters.

WHERE IT SHINES

Rare and dangerous cases

The arrangements that almost never occur, and the ones too costly to stage in a real warehouse.

Four things to check before trusting a dataset

Ask these of every source, synthetic or real

Dimension	The question	How it fails quietly
Accuracy	Do the labels say what actually happened?	Labels drift as the labelling guide is reinterpreted over months.
Completeness	Which situations are missing entirely?	Night shifts, new packaging, or an unusual shelf layout never appear.
Timeliness	Does it still describe the current environment?	The warehouse was reconfigured; the data was not refreshed.
Consistency	Does the same thing mean the same thing everywhere?	Two cameras name the same object differently — the mismatched-topic problem.

WHY SILENCE IS THE DANGER

Three of these four fail without any error message. That is why provenance has to be a structure you build, not a check you remember to run.

What a traceable data supply looks like

Every step recorded, so any output can be traced back

THE PROVENANCE CHAIN

Capture: source, time, sensor, consent	Transform: every step, in order	Label: who, when, by which guide	Use: which model, which version

ONE QUESTION THAT TESTS THE WHOLE CHAIN

The test: pick any prediction the system made last month and trace it back to the exact rows that trained it and where each row came from. If you cannot, you do not have provenance — you have a folder.

Manual coordination or a single source of truth

The difference the warehouse study actually measured

MANUAL

Each step configured by hand

HOW IT WORKS

People configure devices, names and addresses for each deployment.

WHAT GOES WRONG

Mismatched topic names and incorrect IP assignments — silent, and found later.

THE HIDDEN COST

Debugging time, and no dependable description of how the system is meant to fit together.

MODEL-DRIVEN

One description generates the rest

HOW IT WORKS

A semantic model describes devices, streams and software; configuration is generated from it.

WHAT IT PREVENTS

Mismatches cannot be introduced by hand because the configuration is derived, not typed.

THE EXTRA BENEFIT

Schema changes and data migration are managed in one place, keeping legacy data usable.

WHAT THE STUDY FOUND

The finding was that abstracting the complexity improved consistency and sped up scaling — the reverse of the usual worry that an extra layer slows everything down.

Provenance has a sensible amount

Both too little and too much stop people using it

TOO LITTLE

A dataset with a filename and no history. Nothing can be reproduced or defended.

PROPORTIONATE

Source, transformations, labelling rules, permissions and version — captured automatically.

TOO MUCH

Every keystroke logged into a store nobody queries. Cost without recall.

WHAT TO CAPTURE AUTOMATICALLY

Record	Because
Source and timestamp.	Without these, nothing else can be checked.
Transformation and code version.	The same input and a different version produce a different dataset.
Real or synthetic, per row.	This is the single flag that prevents the most expensive mistake.

Three ways data trust is lost

Each one is quiet until it is very loud

Synthetic data in the test set

Looks like Excellent evaluation numbers.

Why it fails You measured how well the model learned the generator, not the world.

Fix Hold out real data from the operating environment, and keep it separate.

Provenance added afterwards

Looks like A documentation task at the end of the project.

Why it fails Nobody can reconstruct what happened months earlier.

Fix Capture it automatically at the moment of each step, or it will not exist.

Schema change with no migration path

Looks like A field renamed in the new version.

Why it fails Every historical record silently stops matching, and old results become unreproducible.

Fix Migration in both directions, managed in the same model as the schema.

CASE STUDY

Warehouse robotics orchestration study

A semantic model used as a single source of truth for a changing robot deployment.

A single source of truth for a changing deployment

Configuration-driven workflows in place of manual orchestration

THE PROBLEM

Configuration drifted silently

THE SETTING

A real warehouse section of 600 m^2 fitted with 25 cameras and AgileX and Clearpath robots, studied in June 2026.

THE FAULTS

Manual deployment produced frequent silent bugs — mismatched topic names and incorrect IP assignments.

WHY THEY HID

A deployment could look configured while the relationships between devices, streams and software were not actually aligned.

THE FIX

One description of how it fits together

THE MECHANISM

A semantic model as a single source of truth, unifying schema evolution and data migration.

THE WORKFLOW

Configuration-driven deployment replaced manual orchestration, reducing setup errors and debugging time.

CONTINUITY

Bidirectional schema migration kept legacy data compatible as the software evolved.

READING THE CASE

Provenance is not documentation written after the fact. It is the structure that lets meaning survive while the system around it keeps changing.

Drawn from the session case pack.

APPLICATION

Bringing the theory to LumaForge Robotics

Building a data supply you can defend

Step	What you put in place	LumaForge example
Tag origin on every row	Real, augmented or synthetic — always.	A single required field, set at capture time and never editable afterwards.
Hold out real data	Evaluation uses operating-environment data only.	One week of real picks per month, reserved and never trained on.
Record transformations automatically	In the pipeline, not in a document.	Each dataset version stores the code version that produced it.
Describe the system once	Devices, streams and software in one model.	Camera ids, topics and addresses generated from a single description.
Plan migration both ways	Before the first schema change.	Old pick logs remain readable after the label format changes.
Use synthetic for rare cases	And say so when you report results.	Unusual shelf arrangements generated; performance still reported on real picks.

WHAT COMES NEXT

Session 1.3 uses this foundation to decide which model should handle each request.

SUMMARY

What to carry out of this session

1. **Provenance answers what a dataset can prove.**
   Origin, transformations, permissions — recorded as you go, not written afterwards.

2. **Synthetic data trains; real data judges.**
   Evaluating on synthetic data measures the generator, not the world.

3. **Fidelity, utility and privacy trade against each other.**
   Measure utility on real held-out data rather than resemblance.

4. **Silent errors are the expensive ones.**
   Mismatched names and wrong addresses look configured and are not.

5. **One description beats many local fixes.**
   Generated configuration cannot contain a typo nobody typed.

FURTHER READING

Sources for this session

Primary references behind each concept

PROVENANCE	Moreau, L. & Groth, P. (2013). *Provenance: An Introduction to PROV*. San Rafael: Morgan & Claypool.
STANDARD	W3C (2013). *PROV-DM: The PROV Data Model.* W3C Recommendation, 30 April 2013.
DATA QUALITY	Wang, R. & Strong, D. (1996). Beyond accuracy. *Journal of Management Information Systems*, 12(4), 5–33.
DOCUMENTATION	Gebru, T. et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
DATA CASCADES	Sambasivan, N. et al. (2021). “Everyone wants to do the model work, not the data work.” *CHI 2021.*
SYNTHETIC DATA	Jordon, J. et al. (2022). Synthetic data — what, why and how? *Royal Society and Alan Turing Institute.*
PRACTICE	Kleppmann, M. (2017). *Designing Data-Intensive Applications.* Sebastopol: O'Reilly.