MODULE 3 · SESSION 1

Building an Experiment Portfolio

How to allocate experiments across growth uncertainties, balancing evidence quality, resource limits and decision deadlines.

Theory session

Intermediate

Running case: Northstar Habitat

ORIENTATION

Where this session sits

Module 3 decides where to point the engine Modules 1 and 2 built

MODULE 1
Growth Architecture and Market Entry

MODULE 2
Commercial Engines and Expansion

MODULE 3
Strategic Scaling Choices

SESSIONS IN THIS MODULE

3.1 Building an Experiment Portfolio — YOU ARE HERE
Allocating tests across uncertainty, evidence and deadlines

3.2 Expanding Markets Through Partners
Partner roles, incentives, governance and local credibility

3.3 Designing the Scaling Operating Model
Decision rights, phases and resource allocation

SESSION OUTCOMES

What this session explains

Running experiments as a portfolio rather than a queue

01

A portfolio balances exploration and exploitation

All-exploitation stagnates; all-exploration never compounds.

02

Evidence has a price

Stronger evidence costs more and takes longer. Buy only what the decision requires.

03

Deadlines change the design

An experiment that reports after the decision has no value, whatever its rigour.

04

Measurement is part of the intervention

Stone Living Lab's six monitoring methods were not overhead; they were the output.

DEFINITION

What makes a set of tests a portfolio

Three properties that a list of experiments does not have

WORKING DEFINITION

An experiment portfolio is a deliberately allocated set of tests spanning different uncertainties, evidence strengths and time horizons — sized so that the whole set informs decisions the company must actually make.

1

It is allocated, not accumulated

Someone decided what share of capacity goes to each kind of uncertainty, rather than running whatever was proposed.

2

It spans horizons

Some tests resolve this month; some establish conditions for next year. Both are needed.

3

It has a reporting deadline

Each test names the decision it feeds and when that decision happens.

March, J. (1991). Exploration and exploitation in organizational learning. Organization Science, 2(1), 71–87.

CONCEPT 1 · ALLOCATION

Exploration against exploitation

March's trade-off, applied to a growth team's capacity

EXPLOIT

Optimising what already works — conversion, messaging, pricing. Reliable, incremental, and eventually exhausted.

EXPLORE

Testing new channels, segments and models. Mostly fails, and is the only source of the next S-curve.

TRANSFORM

A small share on tests that would change what the business is. Rarely pays this year.

A WORKABLE DEFAULT SPLIT

Horizon	Share of capacity	What it looks like
Exploit	~70%	Onboarding, pricing page, message tests, handoff improvements.
Explore	~20%	A new channel, a new segment, a different packaging model.
Transform	~10%	A second product line, a partner-led motion, a new geography.

CONCEPT 2 · EVIDENCE

What each level of evidence costs

Buy the weakest evidence that would actually change the decision

OPINION AND ANALOGY

Expert judgement, competitor behaviour, case studies. Free, and it settles nothing.

OBSERVATIONAL

Existing data, cohort analysis, correlations. Cheap, confounded, often sufficient.

QUASI-EXPERIMENT

Before-and-after, matched markets, staged rollout. Good when a randomised test is impossible.

CONTROLLED EXPERIMENT

Randomised, powered, pre-registered. Strongest, slowest, and not always available.

THE RULE

Match to reversibility

An easily reversed decision needs weak evidence. An irreversible one justifies the expensive test.

STONE LIVING LAB

Monitoring as the product

Six methods captured physical and biological response. The evidence, not the structure, was the deliverable.

THE COMMON ERROR

Over-buying evidence

Running a rigorous test on a decision you would reverse next month spends the portfolio's capacity on certainty nobody needed.

CONCEPT 3 · SELECTION

Which uncertainties deserve a test

Cost of being wrong against how much you already know

COST IF WRONG ▲

TEST PROPERLY Expensive to get wrong, genuinely unknown. This is what the portfolio is for.	**MONITOR** Important and well evidenced. Watch for drift; do not re-test.
DECIDE AND MOVE Unknown but cheap to reverse. Pick one, ship it, look at the result.	**LEAVE ALONE** Low stakes, well understood. Testing it is capacity spent on comfort.

EVIDENCE ALREADY HELD →

THE DISCIPLINE

Most items are bottom-left

Teams over-test reversible decisions because tests feel rigorous. Shipping and watching is usually faster.

CAPACITY

Four to six live at once

Beyond that, nothing gets analysed properly and results arrive after the decision.

THE FORGOTTEN AXIS

Reversibility

Ask how expensive it would be to undo before asking how sure you are.

CONCEPT 4 · DESIGN

What each experiment has to declare before it starts

Six fields, written down, or the result will be argued about afterwards

Field	Why it must be fixed in advance
The decision it feeds	If no decision depends on the answer, the test is curiosity, not portfolio capacity.
The hypothesis	A directional claim, so the result can contradict it.
The primary metric	One. Choosing afterwards from several guarantees a positive result somewhere.
The threshold	What value changes the decision — set before data arrives.
Duration and sample	Stopping when the result looks good is the most common way teams fool themselves.
The deadline	When the decision happens regardless of whether the test has concluded.

THE TWO THAT GET SKIPPED

The threshold and the deadline are the two most often omitted, and they are the two that make a result actionable.

CONCEPT 5 · POWER

Small effects need large samples

Most growth tests are underpowered and report noise as a finding

A test sized to detect a 20% lift cannot speak to a 3% one. Running it anyway produces a number that means nothing.

Kohavi, R., Tang, D. & Xu, Y. (2020). Trustworthy Online Controlled Experiments. Cambridge University Press.

BEFORE YOU START

What lift would matter?

Decide the smallest effect worth acting on, then check whether your traffic can detect it.

IF IT CANNOT

Change the question

Test a bolder variant, pick a coarser metric, or use a quasi-experiment instead.

THE TRAP

Peeking

Checking daily and stopping at significance inflates false positives severalfold.

CONCEPT 6 · BEYOND THE VERDICT

What a pilot tells you besides whether it worked

Stone Living Lab's most useful output was the second column

PERFORMANCE

Did the intervention work?

WHAT IS MEASURED

The effect itself — erosion reduced, surge buffered, biodiversity enhanced.

WHAT IT SUPPORTS

A decision about this intervention, in this setting.

ITS LIMIT

Site-specific. It says nothing about whether the result travels.

CONDITIONS

What would adoption require?

WHAT IS OBSERVED

What had to be true for it to work — inputs, partners, standards, local circumstances.

WHAT IT SUPPORTS

A decision about scaling, which is a different question entirely.

WHY IT WAS THE GAP

Stone Living Lab produced site-specific evidence with no mechanism to mainstream it.

THE DESIGN INSTRUCTION

Design pilots to answer both. A portfolio that only measures performance will keep producing successful demonstrations that never scale.

CONCEPT 7 · FAILURE MODES

Three portfolios that generate activity and no learning

Each runs plenty of tests

All exploitation

Looks like Twenty button and copy tests a quarter.

Why it fails Incremental gains on a curve that is flattening; no next S-curve is being found.

Fix Ring-fence capacity for exploration.

No stated threshold

Looks like Results discussed after the fact.

Why it fails Whatever the number, someone argues it supports their prior.

Fix Write the decision-changing value before the test starts.

Demonstrations that cannot scale

Looks like A successful pilot nobody can repeat.

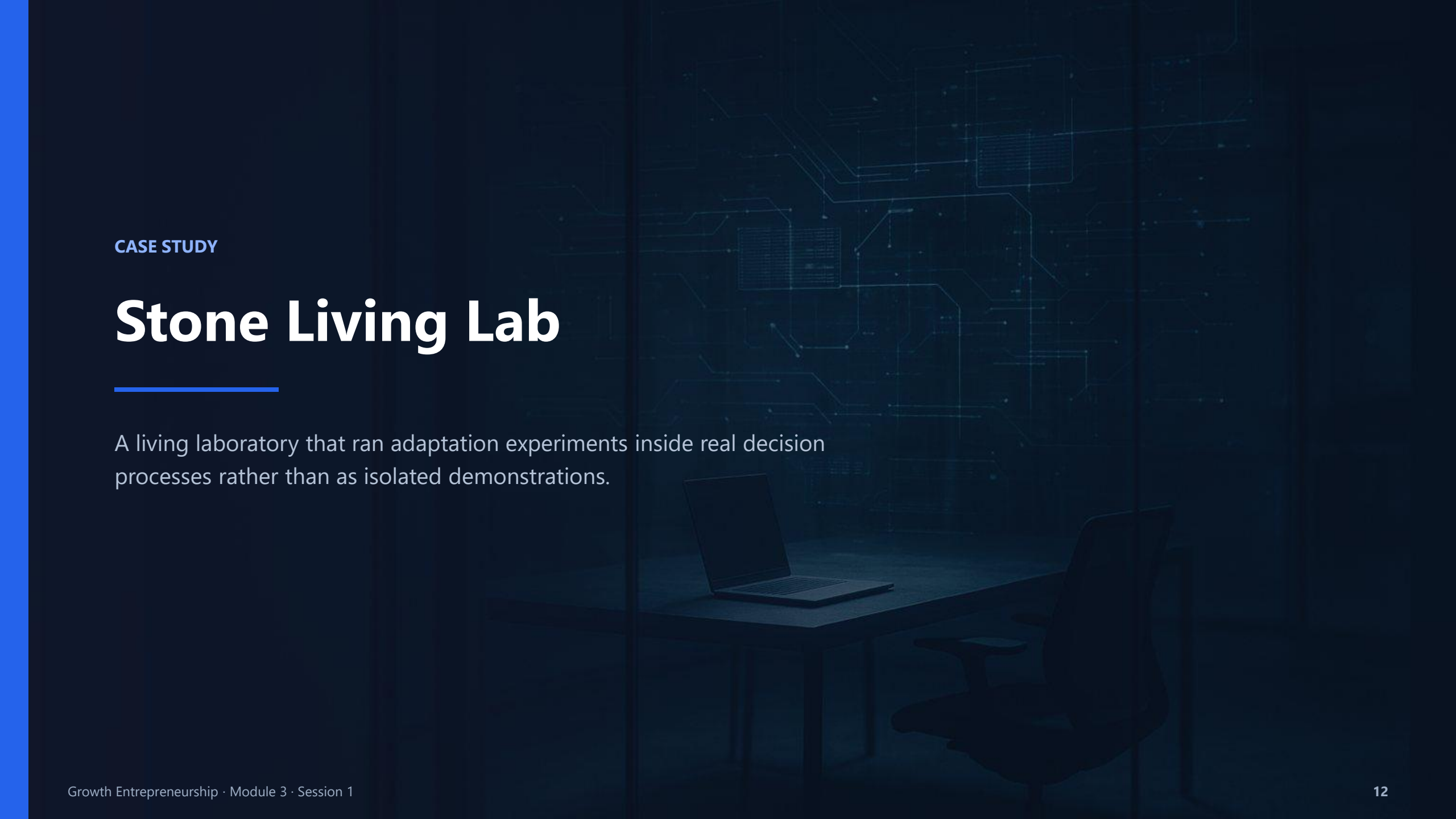
Why it fails Only performance was measured; the conditions for adoption were not.

Fix Record what had to be true, alongside what happened.

CASE STUDY

Stone Living Lab

A living laboratory that ran adaptation experiments inside real decision processes rather than as isolated demonstrations.



CASE STUDY · STONE LIVING LAB

Experiments inside real decisions

Monitoring was part of the intervention, not overhead

THE PORTFOLIO

Varied pilots, embedded measurement

THE PERIOD

Established as a living laboratory from 2018 to 2024, testing in real coastal conditions.

TWO NAMED APPROACHES

Cobble berms for erosion mitigation; living seawalls to enhance biodiversity.

MEASUREMENT AS PART OF THE INTERVENTION

Six monitoring methods — buoys, tide gauges, laser scanning, sonar, drones and biological surveys.

WHAT IT PRODUCED

More than a yes or no

PERFORMANCE EVIDENCE

Site-specific evidence on erosion reduction, storm surge buffering and biodiversity.

THE DESIGN GAP IT EXPOSED

Evidence in one setting, with no mechanism to carry it into wider practice.

THE SHIFT IN THINKING

Results moved stakeholders toward flexible, multifunctional solutions rather than single-purpose ones.

READING THE CASE

A pilot produces more than a verdict. It reveals both how an intervention performs and what conditions its wider adoption would require.

Drawn from the session case pack.

APPLICATION

Bringing the theory to Northstar Habitat

A quarter's portfolio, allocated deliberately

Allocation	The uncertainty	Test and threshold
Exploit (70%)	Does the assessment report drive a second engagement?	Two report formats, matched clients. Threshold: 15 point difference in follow-on rate.
Explore (20%)	Will municipalities buy the same service as corporates?	Four paid pilots at list price. Threshold: two convert without discount.
Transform (10%)	Could monitoring be sold as a subscription?	Concierge delivery for three clients. Threshold: one renews unprompted.
Every test	What conditions would adoption require?	Recorded alongside the result — inputs, partners, standards.
The deadline	When the decision happens regardless.	Budget round in eleven weeks; all three report by week nine.

WHAT COMES NEXT

Session 3.2 takes whichever expansion direction survives and asks who should carry it.

SUMMARY

What to carry out of this session

1. **Allocate capacity across horizons deliberately.**
   All exploitation optimises a flattening curve; all exploration never compounds.

2. **Buy the weakest evidence that would change the decision.**
   Match evidence strength to how reversible the decision is.

3. **Fix the threshold and the deadline before starting.**
   They are the two most often omitted and the two that make a result actionable.

4. **Check that your test can detect the effect you care about.**
   An underpowered test reports noise with a confident number attached.

5. **Measure adoption conditions, not just performance.**
   Otherwise you produce successful pilots that never scale.

FURTHER READING

Sources for this session

Primary references behind each concept

EXPLORATION	March, J. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 71–87.
EXPERIMENTATION	Kohavi, R., Tang, D. & Xu, Y. (2020). *Trustworthy Online Controlled Experiments.* Cambridge: Cambridge University Press.
BUSINESS EXPERIMENTS	Thomke, S. (2020). *Experimentation Works.* Boston: Harvard Business Review Press.
TESTING PRACTICE	Bland, D. & Osterwalder, A. (2019). *Testing Business Ideas.* Hoboken: Wiley.
REAL OPTIONS	McGrath, R. (1999). Falling forward: real options reasoning and entrepreneurial failure. *Academy of Management Review*, 24(1), 13–30.
AMBIDEXTERITY	O'Reilly, C. & Tushman, M. (2004). The ambidextrous organization. *Harvard Business Review*, 82(4), 74–81.
GROWTH PRACTICE	Ellis, S. & Brown, M. (2017). *Hacking Growth.* New York: Currency.