

# Lecture 1 — Introduction to Business Data Analytics

## CRISP-DM · KPI Tree · Problem Framing

Lecturer: Assoc. Prof. Nguyen Binh Minh

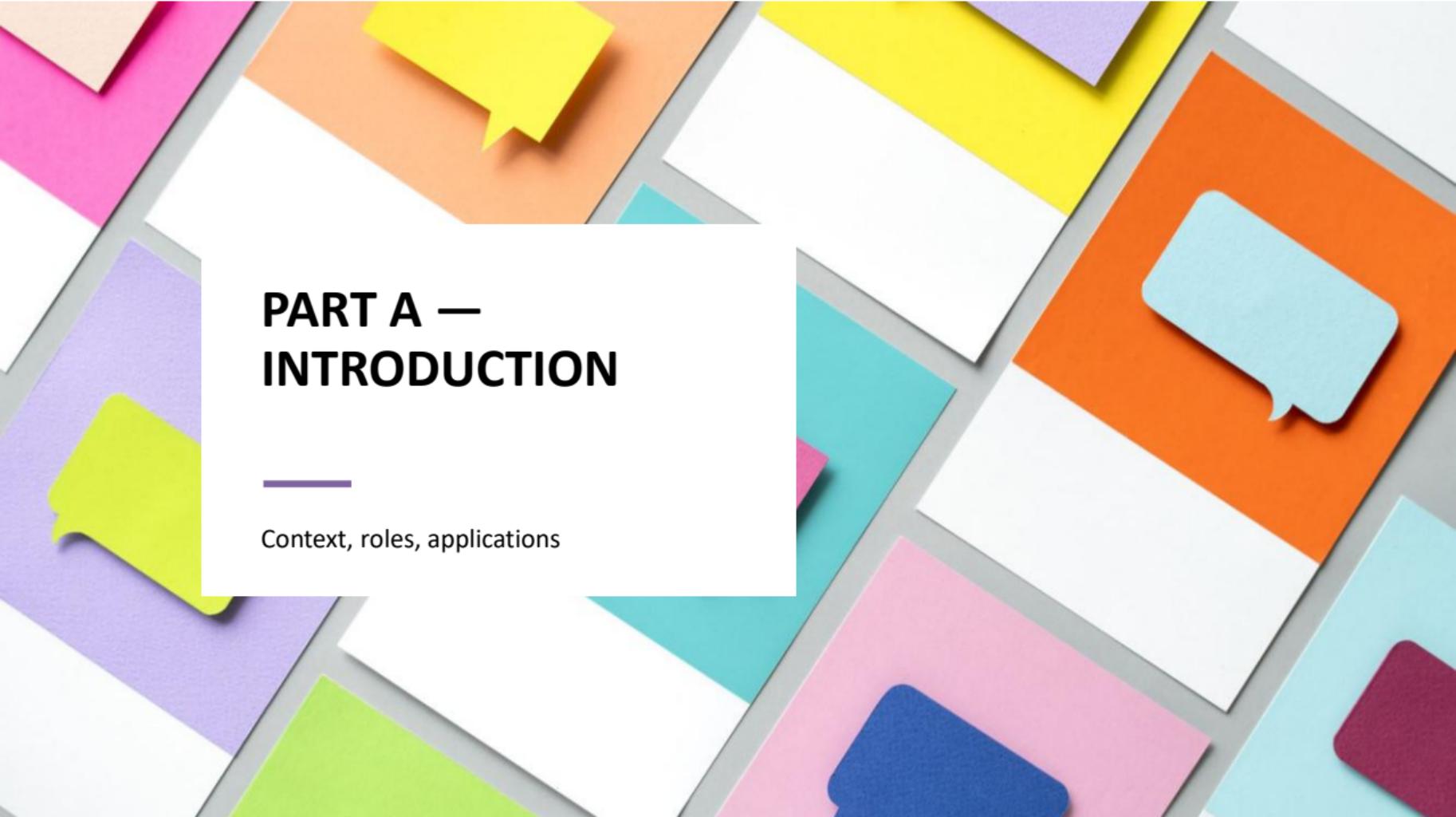
# Agenda (3 Hours)

- 1) Intro to Business Data Analytics (30')
- 2) CRISP-DM: the end-to-end analytics workflow (70')
- 3) KPI Tree: goal-driven metric design (40')
- 4) Problem Framing: decisions, success criteria (30')
- 5) Quick hands-on & homework (10')



# Learning Outcomes

- Understand the analytics ecosystem in enterprises and team roles.
  - Apply CRISP-DM to plan and execute end-to-end analytics projects.
  - Build a KPI Tree for a business goal; choose leading vs lagging indicators.
  - Frame problems around decisions, options, and criteria for success.
  - Practice on an e-commerce case: revenue, conversion, churn.
-



## PART A — INTRODUCTION

---

Context, roles, applications

# What is Business Data Analytics?

A process to turn data into insight and action.

Layers: descriptive, diagnostic, predictive, and prescriptive analytics.

Outputs: reports/KPIs, models, A/B tests, and operational recommendations.

# From BI to DS/AI (quick distinction)

- BI: reporting, dashboards, KPI monitoring (past/present).
- DA/DS: root-cause analysis, forecasting, segmentation, optimization.
- AI/ML: learn from data to automate prediction/recommendation in products.

# Common Applications by Industry

- Retail/E-commerce: demand forecasting, pricing, recommendations, basket analysis.
- Finance/Banking: credit scoring, fraud detection, collections optimization.
- Telecom/SaaS: churn prediction, onboarding optimization, MRR/ARPA improvement.



# Data Value Chain

- Collect → Store → Clean/Transform → Analyze/Model → Deploy → Monitor.
  - Product thinking: lifecycle, quality, reliability, repeatability.
  - Principle: start from the decision to be supported.
-

# Team Roles & Skills

- Product/Data PM, Data Analyst, Data Scientist, Data Engineer, BI Dev, MLOps/Platform.
- Core skills: SQL, stats, EDA, visualization, modeling, data storytelling.
- Supporting: domain knowledge, experiment design, data ethics & governance.

# Ethics, Legal & Data Governance

---



Privacy, security, compliance;  
bias and explainability.



Governance: KPI definitions,  
data quality, lineage & catalog,  
RACI.



Trustworthy AI mindset: fairness,  
safety, transparency, auditability.

# Class Case: Growing E-commerce Revenue

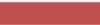
- Goal: +15% revenue next quarter without increasing ad budget.
- Levers: Conversion Rate (CR), Average Order Value (AOV), Retention/Repeat Rate.
- Questions: Where is the bottleneck? Which channel/device? Which user segments respond?



# Available Data for the Case

- Web/app events: sessions, page\_view, add\_to\_cart, purchase (with timestamp, device, channel).
- Orders: value, category, voucher, payment method.
- Campaigns: source/channel, cost; user: cohort, frequency, RFM.





# PART B – CRISP-DM

---

Cross-Industry Standard Process for Data Mining

# CRISP-DM Overview



Six phases: Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment.



Iterative: loop back when new insights or goal changes occur.



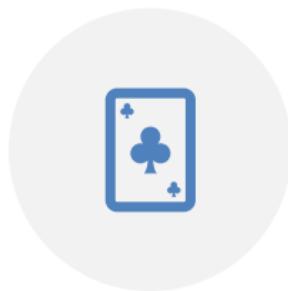
Artifacts: charter, data dictionary, feature list, model card, runbook.

# 1) Business Understanding – goals & questions

---



CLARIFY BUSINESS OBJECTIVES  
AND CONSTRAINTS (SMART).



IDENTIFY DECISIONS THAT  
ANALYTICS WILL INFORM  
(DECISION-CENTRIC).



DEFINE SUCCESS KPIS AND  
GUARDRAILS (E.G., CR, AOV, CAC,  
NPS).

# Business Understanding — e-commerce example

Goal	Goal: +15% Q4 revenue; guardrail: do not reduce overall CR by >1pp.
Increase	Hypotheses H1: increase AOV via bundles/upsell; H2: optimize mobile checkout funnel.
Prioritize	Decision: prioritize checkout improvements before pricing promotions.

# Outputs & Checklist (BU)

- Problem statement, scope, stakeholders & RACI.
- List of analysis questions, hypotheses, KPIs, timeline & risks.
- Data rules: definitions of session, transaction, user, attribution.



## 2) Data Understanding — explore & validate

- Inventory sources: schema, size, freshness, access permissions.
  - EDA: distributions, outliers, missingness, temporal consistency.
  - Reconcile KPI definitions with real data (order date vs ship date, returns).
-

# Data Understanding — quick wins

- Funnel analysis: session → view → add\_to\_cart → checkout → purchase.
- Segment by channel/device: mobile vs desktop; organic vs paid.
- Detect leakage & double-counting of purchases due to retries/timeouts.

# 3) Data Preparation – cleaning & features

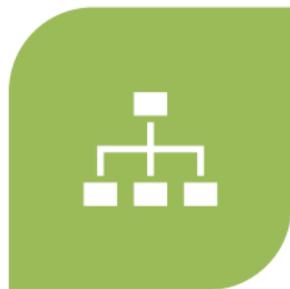
- Handle missing/outliers; normalize units, currency, timezones.
- Feature engineering: recency/frequency/monetary (RFM), price\_sensitivity, device\_speed.
- Train/validation split; prevent data leakage; time-based split.

# Data Prep — quality & Feature Store

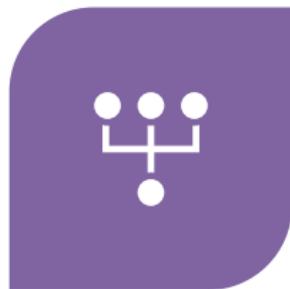
---



DATA TESTS: SCHEMA, RANGES,  
FRESHNESS, UNIQUENESS.



STANDARDIZE TAXONOMY & DATA  
DICTIONARY; EVENT NAMING;  
LATE-ARRIVING DATA.



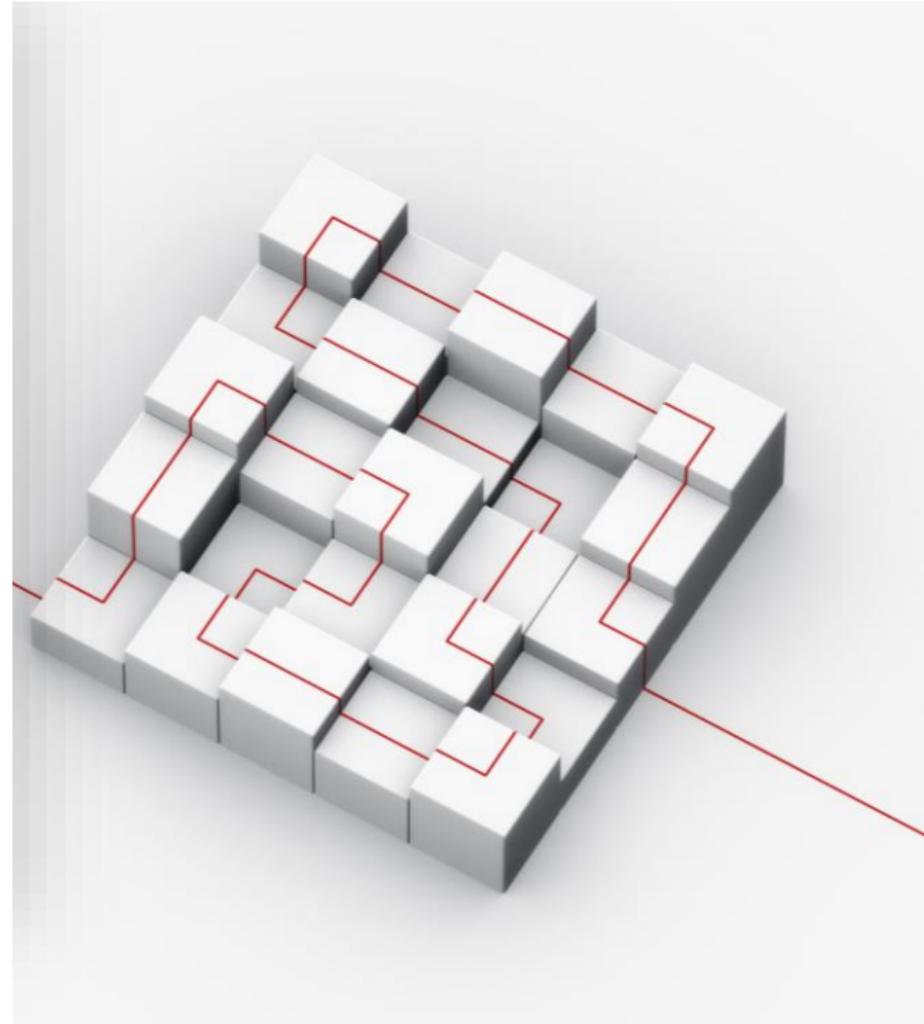
REUSE FEATURES VIA A FEATURE  
STORE; LINEAGE AND VERSIONING.

## 4) Modeling – choose task & baseline

- Classification (churn), regression (AOV), clustering (segments), optimization (pricing).
- Baseline: simple models + rules; compare AUC/MAE/lift.
- Tune threshold to business goals and error costs.

# Modeling — improvement loop

- Time-aware cross-validation; regularization/feature selection.
- Control overfitting; model explainability (e.g., SHAP) and bias checks.
- Packaging: model card, reproducibility, dependency locking.



## 5) Evaluation — technical & business

- Technical: AUC/PR, RMSE/MAE, calibration, cohort stability.
- Business: revenue/profit uplift, KPI & guardrail impact.
- Experimentation: A/B, holdout, careful sequential testing.

# 6) Deployment — operate & monitor



Batch vs real-time scoring; AB switch; canary/blue-green.



Monitoring: data/model drift, SLA/latency, alerting & auto-rollback.



MLOps: CI/CD, registry, feature–model consistency.

# CRISP-DM — handover package

- Project charter, data dictionary, EDA report & quality checklist.
- Model card, evaluation report, rollout plan, monitoring dashboard.
- Operations runbook & continuous improvement plan.

# CRISP-DM — common pitfalls

- Starting from data, not from decisions; unmeasurable KPIs.
- Data leakage; inconsistent KPI definitions across teams.
- Great model, zero business impact — no adoption/experimentation.

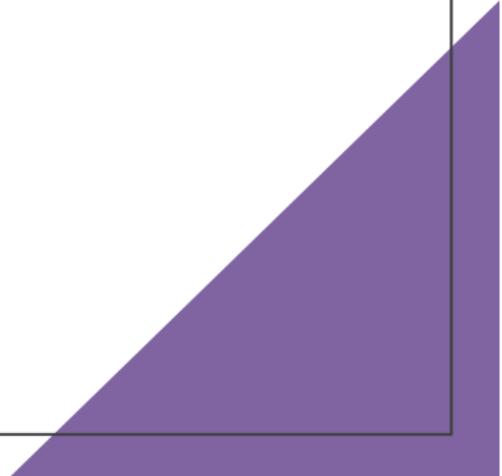


# Quick Exercise 1 (10')

- For the e-commerce case: fill the BU charter (goals, KPIs, guardrails).
  - List 3 key hypotheses and the data needed to test them.
  - Draft RACI: PM/DA/DE/DS/Product Owner.
-

# **PART C – KPI TREE**

From North Star → drivers → actionable levers



# KPI: definitions & types

- Lagging vs Leading; Input vs Output; Health vs Performance.
- Principles: fewer but better; measurable, available, actionable.
- Avoid vanity metrics unless tied to decisions.

# KPI Tree — 5-step process

- 1) Choose the North Star (e.g., Monthly Revenue, MRR, GMV).
- 2) Decompose into drivers you can influence.
- 3) Attach levers and owners to each branch.
- 4) Define guardrails & trade-offs; instrumentation and update cadence.
- 5) Connect to dashboards and weekly review cadence.

## KPI Tree Example — E-commerce (Revenue)

- Revenue = Sessions × Conversion Rate × Average Order Value (AOV).
- CR =  $P(\text{view} \rightarrow \text{cart}) \times P(\text{cart} \rightarrow \text{checkout}) \times P(\text{checkout} \rightarrow \text{purchase})$ .
- AOV =  $\Sigma(\text{price} \times \text{qty})/\#\text{orders}$ ; Levers: bundles, upsell, free-shipping threshold.

# E-commerce — numeric illustration

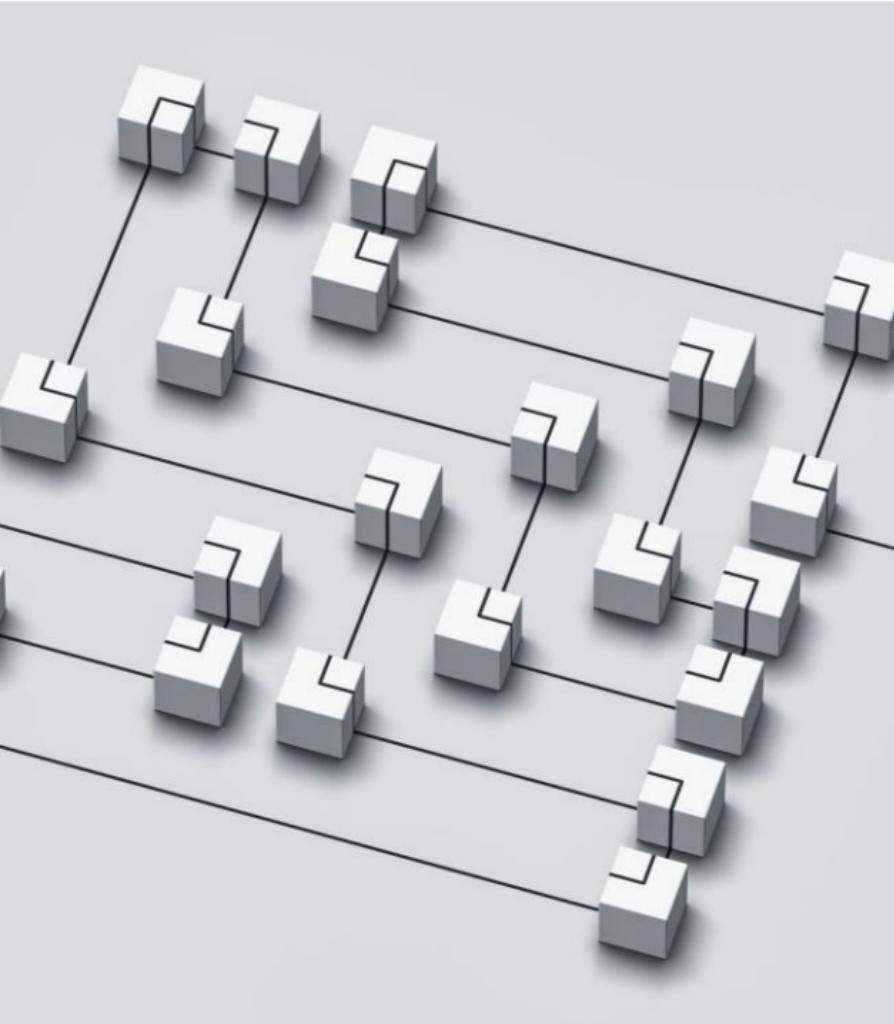
- Current: 1,000,000 sessions/month; CR = 2.0%; AOV = 420k  
→ Revenue  $\approx$  8.4bn (VND).
- Target: +15% revenue → 9.66bn. Two options:
- O1: raise CR to 2.3% (AOV fixed); O2: raise AOV to 483k (CR fixed).

# KPI Tree Example — SaaS (MRR)

- $MRR = \#Customers \times ARPA \times (1 - Churn)$ .
- $\#Customers = New\ Signups \times Activation\ Rate \times Conversion\ to\ Paid$ .
- Levers: pricing tiers, onboarding, success motions, retention plays.

# Guardrails & measurement quality

- Canonical definitions (returns/cancellations; order edit dates).
- Windows for CR/AOV by channel/device; remove bot traffic.
- Health metrics: latency, payment errors, inventory availability.



# Measurement Plan & Instrumentation

- Events: page\_view, add\_to\_cart, begin\_checkout, purchase (with parameters).
- Schema: user\_id, session\_id, device, source, campaign, value.
- Fix update frequency and dashboard ownership.

# Dashboard Design

Board 1: North Star + driver-tree heatmap (red/yellow/green).

Board 2: Funnel by device/channel;  
Board 3: AOV by category.

Features: cohort filters, annotate events (sales, UI changes).



## KPI Tree – common pitfalls

Overlapping branches/double counting; missing guardrails.

No clear owners/levers; weak causal validation.

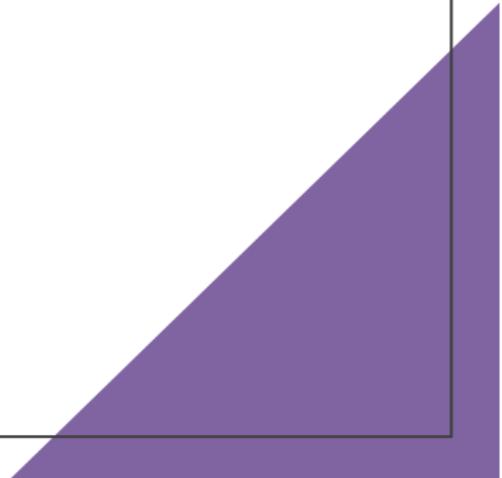
Slow refresh; metrics lag decisions.

# Quick Exercise 2 (10')

- Teams of 2–3: build a KPI Tree for a ride-hailing app.
- North Star: Successful trips/day or GMV.
- Identify three key drivers and associated levers.

# PART D — PROBLEM FRAMING

Decision-centered framing & KPIs



# Why Problem Framing matters

- Up to 80% of analytics failures stem from misframed problems/KPIs.
- Guides the right analysis/model and avoids wasted effort.
- Enables measurement of impact and consistent decision-making.

# Problem Statement Template

- Context: describe the current situation and goal.
- Decision: the choice (D) and candidate options (O).
- Criteria: the set of KPIs to compare options (C) — Decision-Options-Criteria.

# From business question → analytics task

- Question: 'Grow revenue by 15% without increasing ads' →
- Task 1 (diagnostic): find funnel bottlenecks & affected cohorts.
- Task 2 (prescriptive): prioritize improvements with highest ROI.



# Define target variable & leakage

- Target: purchase within 7 days since session (binary).
  - Avoid leakage: exclude variables generated after the decision point (e.g., applied voucher).
  - Time-based splits; monitor drift & cohort stability.
-

# Constraints, assumptions, risks

- Constraints: engineering resources, rollout time, SLA.
- Assumptions: stable seasonality; data reliability.
- Risks: ad policy changes, supply chain disruptions, behavior shifts.

## Success criteria & guardrails

- Primary: +15% revenue; Secondary: CR drop <1pp; stable AOV.
- Technical: lift  $\geq X\%$ , SE/CI acceptable; Product: latency  $< 200\text{ms}$ .
- Ethics: do not worsen experience for vulnerable groups; transparent pricing.



# Prioritization — Impact/Effort matrix

- List opportunities; estimate impact and implementation cost.
- Rank: quick wins, big bets, fillers, avoid.
- Create a 6–8 week backlog with clear milestones.

# A/B Testing — basics

- Set the randomization unit (user/session) and primary metric.
- Avoid peeking; ensure group independence.
- Estimate test duration from traffic and minimum detectable effect.

# Rule of thumb for sample size

- Approx. for CR:  $n \approx 16 \times p(1-p) / \Delta^2$  per group.
- Example:  $p=2\%$ ,  $\Delta=0.3\text{pp}$  ( $0.003$ )  $\rightarrow n \approx 34,844$  per group.
- Note: use proper calculators for multiple KPIs/adjustments.

# Quick Exercise 3 (10')

- Write a problem statement: goal, decision, criteria.
- Propose 3 options; estimate impact & relative effort.
- Pick one for A/B; define KPIs and guardrails.



# Anti-patterns & fixes

- Analytics work not tied to a decision → ask: ‘What decision will this inform?’
  - Vague/unmeasurable KPIs → standardize definitions & sources.
  - Technical success, business failure → experiment & canary releases.
-

# **WRAP-UP & NEXT STEPS**

References, homework, what's next

---

# Key Takeaways



Business analytics turns data into insight and action.



CRISP-DM provides a reliable blueprint for end-to-end projects.



KPI Trees connect the North Star to actionable levers.



Problem framing centers analytics on decisions and measurable success.

# Suggested References

- Davenport & Harris — Competing on Analytics.
- Provost & Fawcett — Data Science for Business.
- Google Analytics / Amplitude Taxonomy docs (event design).
- O'Reilly — Practical A/B Testing; CRISP-DM guide.

# Homework

- Pick a familiar digital product (web/app).
- Build a 3-level KPI Tree and list three actionable levers.
- Write a 1-page problem statement using DOC (Decision-Options-Criteria).

# Prep for Next Lecture

- Hands-on SQL & EDA: connect event and order data.
- Build cohorts, funnels & RFM; prepare a KPI dashboard.
- Intro to notebooks & data quality checklists.

# Q&A

- Open discussion: biggest barriers to KPI Trees in practice?
- How to balance guardrails when optimizing a specific driver?

# **APPENDIX — TEMPLATES**

Canvases / Checklists / Glossary

---

# Template — Problem Framing Canvas

- Context & Goals; Decision; Options; Criteria & KPIs.
- Constraints; Assumptions; Risks; Experiment/Measurement plan.
- Owner; Timeline; Required data preparation.



# Template — KPI Tree (blank)

- North Star: \_\_\_\_\_
  - Branch 1 → driver → lever → owner
  - Branch 2 → driver → lever → owner
  - Guardrails: \_\_\_\_\_
-



# Checklist — CRISP-DM

- BU: SMART goals, KPIs/guardrails, stakeholders & RACI.
  - DU: inventory, EDA, data definitions, quality issues.
  - DP: missing/outliers, feature list, time-aware splits.
  - M: baselines, evaluation, explainability, versioning.
  - E: business evaluation, experiment plan.
  - D: deployment, monitoring, runbook, rollback.
-

# Glossary

- AOV, CR, CAC, LTV, MRR/ARR, ARPA, Churn, Retention, Cohort.
- Leading/Lagging KPI, Funnel, Uplift, Guardrail, Drift.
- Canary, Blue-Green, Feature Store, Model Card.

# **Q&A**

# **THANK YOU**

[minh.nguyenbinh@hust.edu.vn](mailto:minh.nguyenbinh@hust.edu.vn)



