

CRISP-DM in 2026

The Cross-Industry Standard Process for Data Mining: the six phases, what each one must produce, and how teams run it now alongside MLOps, LLMs and AI regulation.

by **Richard J. Young, Ph.D.**

Assistant Professor-in-Residence, Information Systems • Lee Business School, University of Nevada, Las Vegas

ryoung@unlv.edu • richard@deepneuro.ai • October 2026

One-sentence thesis. CRISP-DM is not software and not a business intelligence tool. It is a *process model*: a disciplined route from a business question to a deployed, monitored decision. The six phases written in 2000 still hold in 2026. What changed is everything wrapped around them.

1996

CONSORTIUM FORMED

2000

VERSION 1.0 PUBLISHED

6

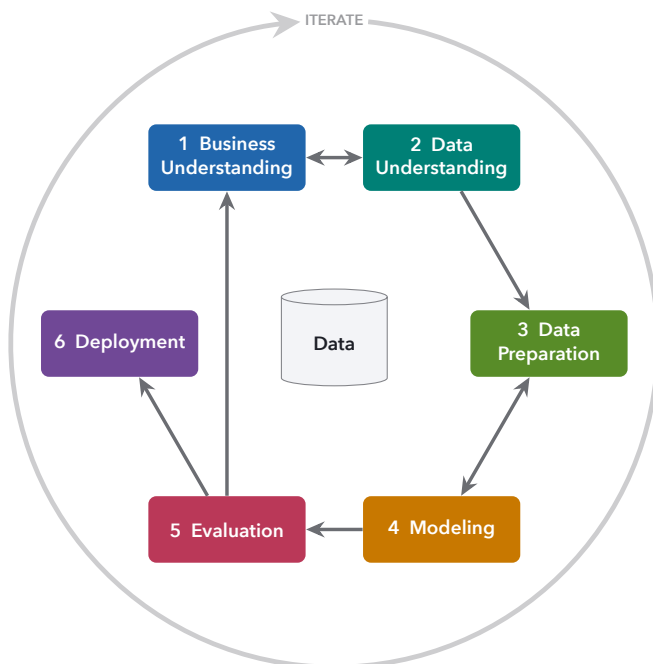
PHASES

24

GENERIC TASKS

PART 1: What CRISP-DM Is

The model in one picture



CRISP-DM was conceived in late 1996 by Daimler-Chrysler, SPSS and NCR, joined by the Dutch insurer OHRA, and published as version 1.0 in 2000. It is deliberately neutral: it names no industry, no tool and no algorithm. A 2.0 revision began in 2006 but was discontinued, so 1.0 is still the reference.

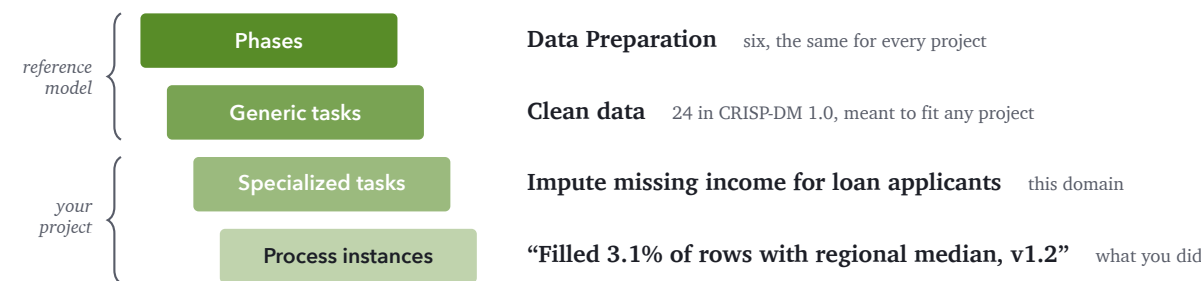
Reading the arrows.

- **Business** ↔ **Data Understanding**: the data you find changes what you can promise.
- **Preparation** ↔ **Modeling**: each technique needs data in its own shape, so you go back.
- **Evaluation** → **Business**: a model can pass every metric and still fail the business test.
- The outer ring: a deployed model raises new questions and creates new data. It is a loop, not a line.

ANALOGY CRISP-DM is a pilot’s pre-flight checklist, not a flying lesson. It will not tell you which algorithm to use. It makes sure you never skip the step that crashes the project, such as agreeing what “success” means before anyone trains a model.

The structure underneath: four levels

CRISP-DM 1.0 describes work at four levels of detail. The top two are the **reference model** and are the same for every project. The bottom two are yours: you map each generic task to your domain, then keep a record of what was actually done.



EXAM TIP “How many phases and tasks?” Six phases and 24 generic tasks. Business Understanding has four, Data Understanding four, Data Preparation five, Modeling four, Evaluation three and Deployment four. Each task names *outputs*, the documents or artifacts that prove it was done.

It is not a BI tool

The confusion is common because CRISP-DM is taught in business-analytics courses next to dashboards. The two answer different questions and often feed each other.

	Business intelligence	CRISP-DM
What it is	Tools and practices for reporting: Power BI, Tableau, Looker, SQL dashboards	A vendor-neutral process model for analytics and machine-learning projects
Main question	What happened? What is happening now?	What is likely to happen, and what should we do about it?
Typical output	Reports, dashboards, key performance indicators	A model or decision rule in use, plus the evidence that it works
How they meet	A CRISP-DM project’s output often lands in BI: a churn score becomes a column in the retention dashboard, and the dashboard’s trends seed the next CRISP-DM question.	

Why it is still the default in 2026

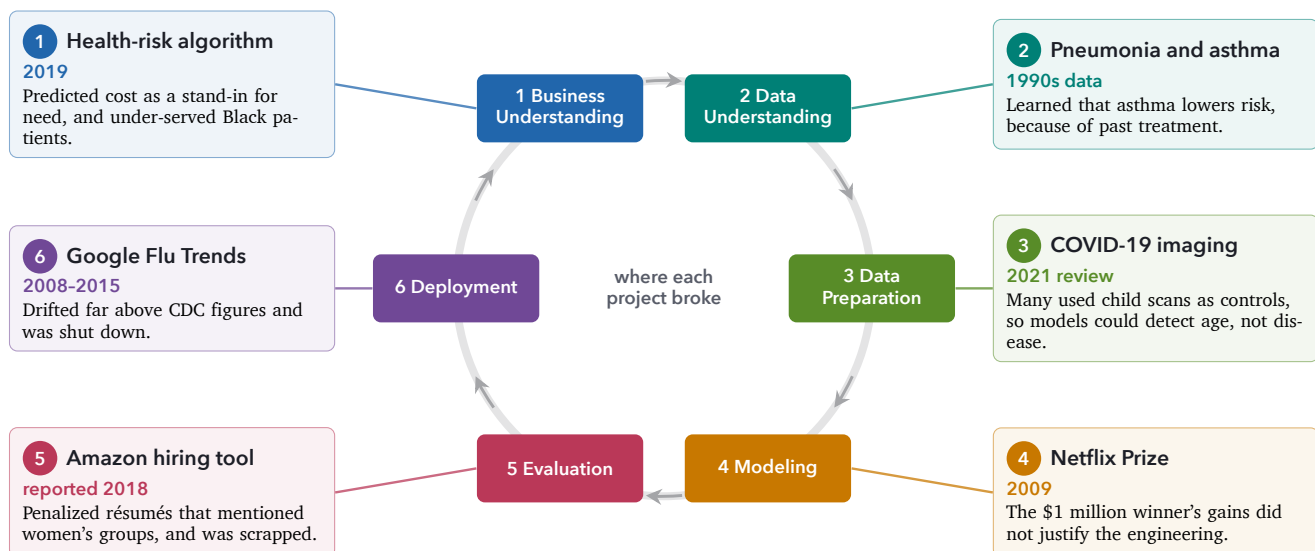
WHAT THE RESEARCH SAYS

- **Still most used.** A Data Science Process Alliance poll found nearly half of practitioners use CRISP-DM most often, ahead of Scrum, Kanban and in-house methods. A 2022 academic survey ranked an agile lifecycle first and CRISP-DM second, with Microsoft's TDSP third.
- **Most teams use nothing.** A 2024 systematic review in *SBC Reviews on Computer Science* cites the finding that 82% of data science teams follow no process model. It found 16 published adaptations of CRISP-DM, which add phases, modify phases, add tools, or merge CRISP-DM with other methods.
- **It combines with agile.** A 2026 case study at the e-commerce firm Elo7 found 86% of the data science team frequently or always applied CRISP-DM, and 71% used Extreme Programming practices alongside it.
- **The honest critique.** A 2026 analysis of 440 Jupyter notebooks found that experts work in short, iterative loops while novices follow long linear sequences. The authors note that models like CRISP-DM describe an idealized workflow rather than how problem-solving actually unfolds. Treat the phases as a map, not a schedule.

PART 2: The Six Phases, Then and Now

Six real projects, one lesson per phase

Each of these projects is well documented, and each broke in a different phase. You will meet them again on the phase cards that follow.



Each card gives the official CRISP-DM 1.0 generic tasks and outputs on the left, with an everyday analogy, and current practice on the right. Under each card, a common failure sits beside the real case.

1 Business Understanding

Turn a business problem into a data problem with a finish line

CRISP-DM 1.0 GENERIC TASKS

- 1 Determine business objectives
- 2 Assess situation
- 3 Determine data mining goals
- 4 Produce project plan

OUTPUTS

Business objectives and success criteria; inventory of resources, constraints and risks; costs and benefits; data mining goals and success criteria; project plan.

THINK OF IT AS

A doctor taking a history before ordering tests. The questions decide which tests are worth running.

HOW TEAMS RUN IT IN 2026

- **Write the decision, not the topic.** “Predict churn” is a topic. “Every Monday, retention staff call the 500 customers most likely to stay *because* of a call” is a decision.
- **Success in business units first** (dollars, hours, readmissions), then translate it to a model metric.
- **Name the baseline:** the current rule, or doing nothing. A model that cannot beat it does not ship.
- **Price the mistakes.** What does a false alarm cost compared with a miss?
- **Agree kill criteria:** what result in the first two weeks stops the project.
- **Classify risk on day one.** Credit, hiring, education, health and other EU AI Act Annex III uses need risk management and documentation from the start.

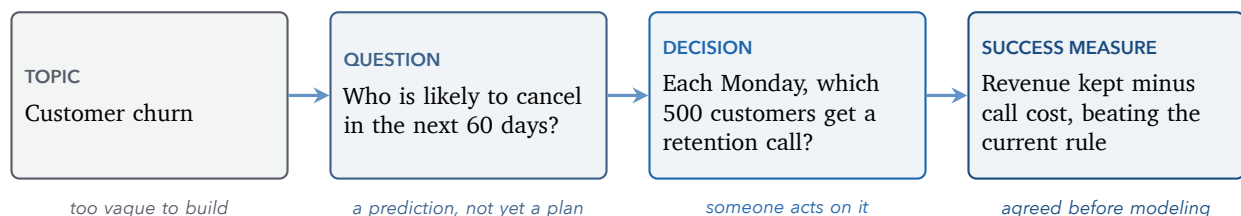
FAILURE MODE Success is defined as “AUC above 0.8” with no link to who acts on the output. The model is accurate, ships, and nobody uses it.

IN REAL LIFE Health-risk algorithm (2019)

Health systems used a commercial algorithm to pick patients for extra-care programs. It predicted future health costs as a stand-in for health *needs*. Less had been spent on Black patients with the same needs, so at the same score they were sicker. Fixing the target would raise the share of Black patients flagged for help from 17.7% to 46.5%.

SOURCE Obermeyer et al., *Science*, 2019

From topic to decision: sharpen the question until someone can act on it



PRICE THE MISTAKES **False alarm:** a wasted call and discount **Miss:** a customer, and their revenue, lost

Agree which mistake costs more before modeling. That choice sets the decision threshold later.

EXAM TIP The first output of a CRISP-DM project is the business objectives and their success criteria, not a dataset or a model. If an exam asks what Phase 1 produces, start there.

AI ON THIS PHASE

what an agent can take, what stays human

HAND TO AN AGENT

Summarise stakeholder interviews; draft KPI trees and success measures; find similar past projects.

KEEP WITH A PERSON

Choose the decision, agree success and kill criteria, and set the risk class and the autonomy level.

WATCH FOR

The agent quietly reframing the question into one it can answer.

2 Data Understanding

Learn what the data can support before promising anything

CRISP-DM 1.0 GENERIC TASKS

- 1 Collect initial data
- 2 Describe data
- 3 Explore data
- 4 Verify data quality

OUTPUTS

Initial data collection report; data description report; data exploration report; data quality report.

THINK OF IT AS

Inspecting a used car: read the service record, look under the hood, and ask what the seller is not telling you.

HOW TEAMS RUN IT IN 2026

- **Profile automatically, then read it with a domain expert.** Profiling tools find the oddities; only the expert knows which ones matter.
- **Pin down the label:** which event, measured when, recorded by whom. Label definitions sink more projects than algorithms do.
- **Run a leakage audit.** For every candidate feature ask: would I know this value at the moment the decision is made?
- **Check who is missing.** If a group is absent from the data but the model will be used on them, record that now.
- **Write a datasheet** (Gebru et al.): where the data came from, consent, known gaps.
- **Agree a data contract** with the owning team: schema, freshness, and who is alerted when it breaks.

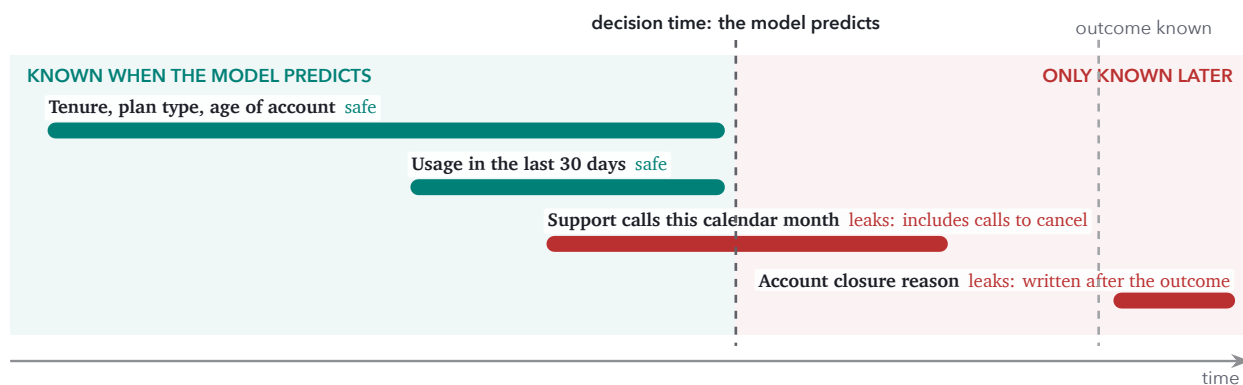
FAILURE MODE A field such as “account closure reason” predicts churn almost perfectly because it is filled in *after* the customer leaves. That is target leakage: superb offline scores, useless in production.

IN REAL LIFE Pneumonia and asthma (1990s data)

A model built to predict which pneumonia patients might die learned that asthma *lowered* the risk. The data was right: asthmatic patients went straight to intensive care, and that extra care protected them. Used to decide who could be treated at home, it could have sent high-risk patients home.

SOURCE Caruana et al., KDD, 2015

The leakage timeline: use only what you would know at decision time



EXAM TIP Data Understanding ends with the *data quality report*. If the data cannot support the goal, the right move is back to Business Understanding, not forward to modeling.

AI ON THIS PHASE

what an agent can take, what stays human

HAND TO AN AGENT

Profile every table; draft the data dictionary and datasheet; flag anomalies and possible leaks.

KEEP WITH A PERSON

Define the label, decide which oddities matter, and ask who is missing from the data.

WATCH FOR

Confident summaries of data the agent only sampled.

3 Data Preparation

Build the final dataset, reproducibly, from raw sources

CRISP-DM 1.0 GENERIC TASKS

- 1 Select data
- 2 Clean data
- 3 Construct data
- 4 Integrate data
- 5 Format data

OUTPUTS

Inclusion rationale; cleaning report; derived attributes; merged and reformatted data; the dataset and its description.

THINK OF IT AS

Mise en place in a kitchen: washing, chopping and measuring before cooking. It takes most of the time, and every dish depends on it.

HOW TEAMS RUN IT IN 2026

- **Preparation is code in a pipeline**, never hand edits in a spreadsheet. The same code must run in production.
- **Split before you transform.** Fit scalers, encoders and imputers on the training data only.
- **Split the way the model will be used:** by time when it predicts the future, by entity (customer, patient) when rows repeat.
- **Version data with code**, so any past model can be rebuilt from its exact inputs.
- **Compute each feature once** for both training and serving, for example in a feature store, to avoid training-serving skew.

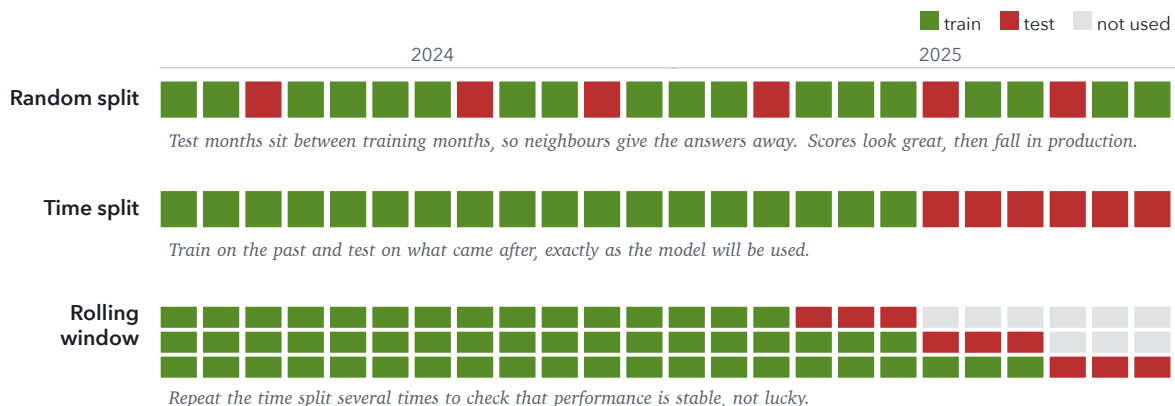
FAILURE MODE A random train-test split on time-ordered data lets the model “see the future”. Offline accuracy looks excellent, then collapses on next month’s data.

IN REAL LIFE COVID-19 imaging models (2021)

A review of 62 published models that read chest scans for COVID-19 found none ready for clinical use. In 16 of them, the non-COVID comparison images came from a dataset of children aged one to five. Those models were largely telling children from adults.

SOURCE Roberts et al., Nature Machine Intelligence, 2021

Split the data the way the model will be used



EXAM TIP Data Preparation is the only phase with five generic tasks: select, clean, construct, integrate and format. *Construct* means deriving new attributes, such as tenure in months from a start date.

AI ON THIS PHASE

what an agent can take, what stays human

HAND TO AN AGENT

Write pipeline code and tests; propose candidate features; document each step.

KEEP WITH A PERSON

Review the code, choose the split, and confirm where each row came from.

WATCH FOR

Generated code that fits transforms on all the data: a silent leak.

4 Modeling

Find the simplest model that meets the success criteria

CRISP-DM 1.0 GENERIC TASKS

- 1 Select modeling technique
- 2 Generate test design
- 3 Build model
- 4 Assess model

OUTPUTS

Technique and its assumptions; test design; parameter settings, models and descriptions; model assessment and revised parameters.

THINK OF IT AS

Writing the exam before teaching the course, so the results cannot tempt you to change the questions.

HOW TEAMS RUN IT IN 2026

- **Design the test before building.** CRISP-DM puts “generate test design” before “build model” on purpose: fix the metric, the split and the baseline first.
- **Climb a ladder:** current rule, then a simple model (logistic regression, gradient-boosted trees for tables), then anything fancier.
- **A foundation model is one more technique.** Prompting, retrieval-augmented generation (RAG) and fine-tuning compete with a classic model on the same test design, including cost and latency.
- **Use AutoML for search, not for judgment.**
- **Track every run:** parameters, data version, code version, metrics.
- **Pick the metric the decision needs:** precision at the capacity you can act on, calibration when people read probabilities, cost-weighted error when mistakes differ in price.

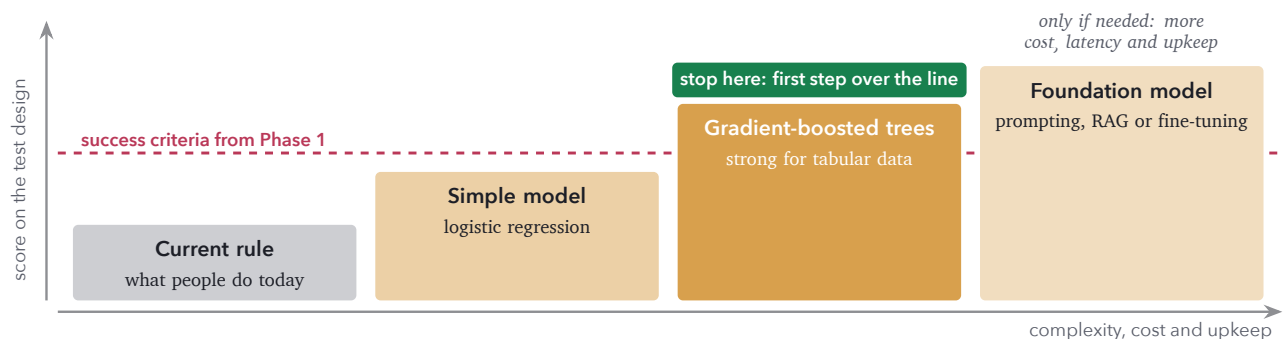
FAILURE MODE Tuning against the test set until it looks good. The test set is used once, at the end. Tune on validation folds.

IN REAL LIFE Netflix Prize (2009)

Netflix paid \$1 million to the team that improved its rating predictions by 10%. It put two simpler algorithms from an earlier progress prize into production, but not the winning ensemble. The extra accuracy did not justify the engineering effort, and viewing was moving to streaming.

SOURCE *Netflix Technology Blog*, 2012

Climb the ladder, and stop at the first step that clears the bar



EXAM TIP *Assess model* is a technical check against the data mining success criteria. The business check comes later, in Evaluation. Exams like to swap the two.

AI ON THIS PHASE

what an agent can take, what stays human

HAND TO AN AGENT

Search model families and settings; track every run; write up the comparison.

KEEP WITH A PERSON

Fix the test design, the baseline and the metric before any search starts.

WATCH FOR

Tuning against the test set. MLE-bench lists two entries that used test-set feedback apart, as not comparable.

5 Evaluation

Judge the model against the business goal, not just the metric

CRISP-DM 1.0 GENERIC TASKS

- 1 Evaluate results
- 2 Review process
- 3 Determine next steps

OUTPUTS

Assessment of results against the business success criteria; approved models; review of the process; list of possible actions; the decision.

THINK OF IT AS

A test drive, not the spec sheet: does it handle your commute, your passengers and your budget?

HOW TEAMS RUN IT IN 2026

- **Score against Phase 1**, in business units, with the baseline beside it.
- **Slice the results** by region, segment, age group or device. An average hides the group where the model fails.
- **Check calibration and stability** across time windows, not only ranking power.
- **Do error analysis**: read the failures, group them, fix the largest group first. For LLM systems, build evaluations from the failures you observe.
- **Validate any LLM-as-judge** against a human-labelled sample before trusting its scores.
- **Write a model card** (Mitchell et al.): intended use, out-of-scope uses, results by group, known limits.
- **Record a go, iterate or stop decision** with the reasons.

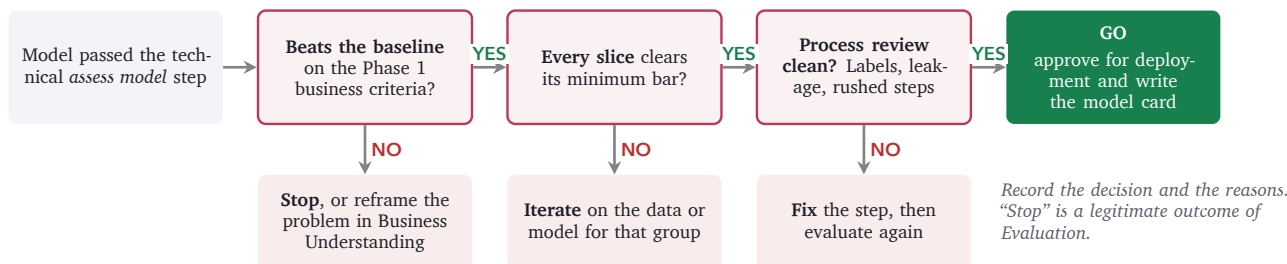
FAILURE MODE “Review process” is skipped. Nobody asks whether labels came from a biased source or a step was rushed, and the flaw ships inside an approved model.

IN REAL LIFE Amazon hiring tool (2018)

Amazon trained a model to rate job applicants on ten years of résumés, most of them from men. It learned to penalize résumés containing the word “women’s”, as in “women’s chess club captain”. Amazon edited the model, could not be sure it was fair, and dropped the project.

SOURCE Reuters, October 2018

The go, iterate or stop decision



EXAM TIP Evaluation has only three generic tasks, and the last one, *determine next steps*, can legitimately end the project.

AI ON THIS PHASE

what an agent can take, what stays human

HAND TO AN AGENT

Run the evaluation suite; cluster the errors; draft the model card.

KEEP WITH A PERSON

Judge business value and fairness, check any LLM judge against people, and make the go-or-stop call.

WATCH FOR

AI grading AI with no person checking the grades.

6 Deployment

Put the model where decisions happen, and keep it honest

CRISP-DM 1.0 GENERIC TASKS

- 1 Plan deployment
- 2 Plan monitoring and maintenance
- 3 Produce final report
- 4 Review project

OUTPUTS

Deployment plan; monitoring and maintenance plan; final report and presentation; experience documentation.

THINK OF IT AS

A new medicine after approval: it still needs post-market surveillance, because real patients differ from trial patients.

HOW TEAMS RUN IT IN 2026

- **Release in stages:** shadow mode (predict but do not act), then a canary or an A/B test against the baseline.
- **Monitor three things:** data drift (the inputs change), concept drift (the input-outcome relationship changes), and the business outcome itself.
- **Set retraining triggers and a rollback path** before launch.
- **Log every prediction** with its inputs and model version, so any decision can be explained later.
- **For LLM features** also watch cost per request, latency, refusals and guardrail hits.
- **Name an owner.** A model nobody owns is a liability.

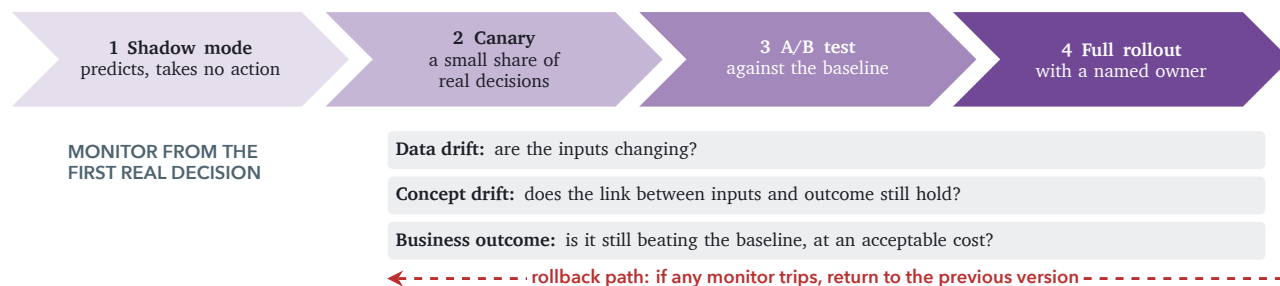
FAILURE MODE CRISP-DM 1.0 allows that deployment “can be as simple as generating a report”. If the goal was to change decisions, a slide deck is not deployment. The output has to reach the system or person who decides.

IN REAL LIFE Google Flu Trends (2008–2015)

Google estimated flu activity from search queries. In the 2012–13 season it forecast more than double the doctor visits for flu-like illness that the CDC recorded. Changes in how people searched, and in Google’s own search tools, had pulled it away from reality. Google stopped publishing the estimates in 2015.

SOURCE *Lazer et al., Science, 2014*

Release in stages, then watch three things



EXAM TIP Two Deployment tasks look backward rather than forward: *produce final report* and *review project*, which records what to do differently next time.

AI ON THIS PHASE

what an agent can take, what stays human

HAND TO AN AGENT

Write monitoring code, alerts and runbooks; summarise drift reports.

KEEP WITH A PERSON

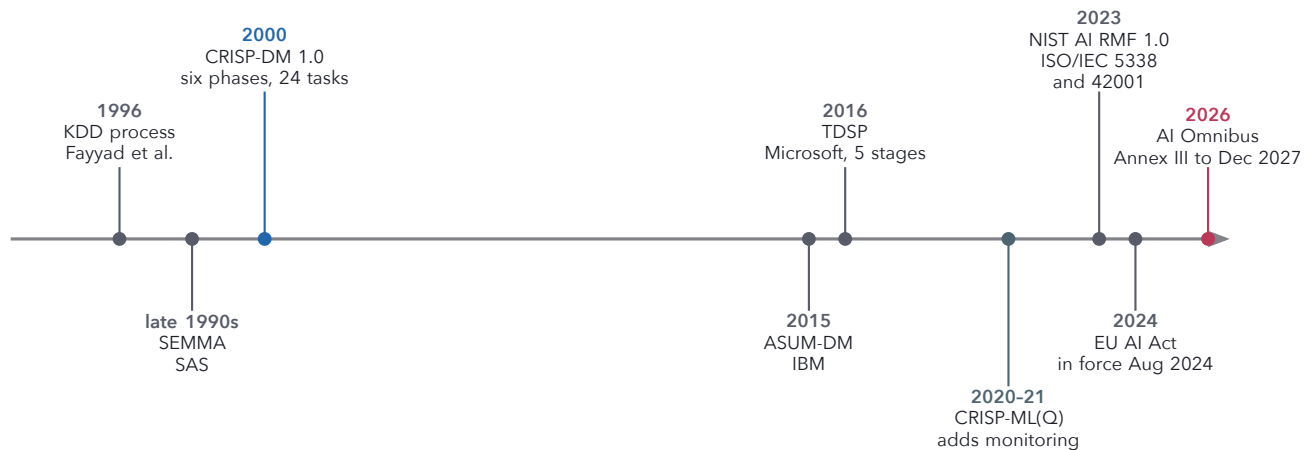
Approve the release, own the rollback, and answer for incidents.

WATCH FOR

An agent with write access to production, as in the Replit case in Part 4.

PART 3: What Changed Since 2000

The family tree



Method	Phases or stages	What it adds, and when to reach for it
KDD process	Selection, preprocessing, transformation, data mining, interpretation	The academic ancestor. Strong on the technical steps, weak on business framing.
SEMMA (SAS)	Sample, Explore, Modify, Model, Assess	Tied to the SAS toolset. No business or deployment phase.
CRISP-DM 1.0	Six phases, 24 generic tasks	The neutral default. Best when a team needs shared vocabulary and checkpoints.
TDSP (Microsoft)	Business understanding, data acquisition and understanding, modeling, deployment, customer acceptance	Defines team roles and a customer sign-off stage. Good for consulting-style delivery.
CRISP-ML(Q)	Business and data understanding, data engineering, ML model engineering, quality assurance, deployment, monitoring and maintenance	Merges the first two phases, adds a full monitoring phase, and attaches quality checks to every task. The best fit for production ML.
Agile hybrids	CRISP-DM phases run in sprints (Scrum, Kanban, XP)	CRISP-DM says <i>what</i> work exists; the agile method says <i>how</i> the team schedules it.
MLOps / LLMops	Continuous integration, delivery, training and monitoring	Automates running the loop again and again. Not a replacement for deciding what to build.

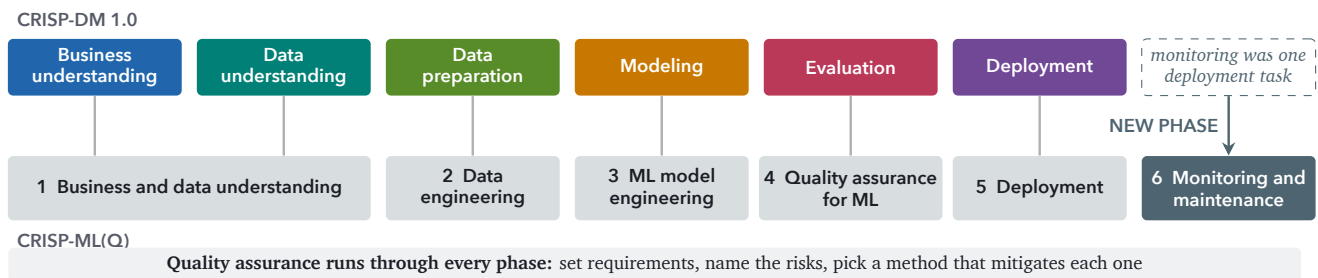
KEY IDEA CRISP-DM 1.0 treats monitoring as one planning task inside Deployment. That was its biggest gap for long-lived models, and it is the main thing every modern extension adds.

CRISP-ML(Q): the quality layer

Studer and colleagues (2021) kept the CRISP-DM spine but asked, for every task, four questions. It is the most useful single habit to borrow.

- 1 What are the **requirements and constraints** for this task?
- 2 What exactly are the **steps**?
- 3 What are the **risks**, such as bias, overfitting, or data that will not be available in production?
- 4 Which **quality-assurance method** mitigates each risk, such as cross-validation, a holdout from a later time period, or documentation?

How the phases map: two merge, one is new

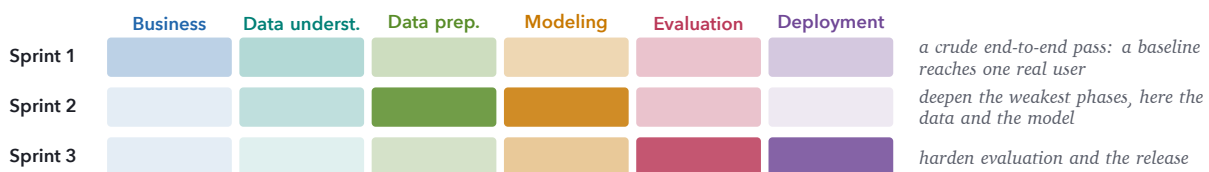


They merge because data decides feasibility. Monitoring becomes a phase because models decay.

Running CRISP-DM with an agile team

- **Go wide before deep.** The first sprint delivers a crude pass through all six phases: a baseline that reaches one real user. Later sprints deepen whichever phase is weakest.
- **Tag backlog items by phase.** The Elo7 team found it hard to tell apart phases that were running at the same time. A phase tag on each ticket keeps that visible.
- **Hold a phase-gate review at the end of each sprint:** is Business Understanding still right given what the data showed?
- **Plan for overlap.** Experts loop quickly between exploring, modeling and checking. That is the process working, not failing.

Go wide before deep: where the effort goes, sprint by sprint (illustration)



PART 4: AI Changes the Scope

Agentic AI does not add a seventh phase. It changes who does the work inside the phases, what gets built, and what has to be governed. Treat it the way CRISP-ML(Q) treats quality: as questions you ask in every phase. Each phase page now ends with an *AI on this phase* strip; this part covers the rest.

Three roles for AI in a CRISP-DM project



1 AI as the worker

An agent does CRISP-DM tasks for you: profiling and cleaning data, writing pipeline code and tests, tuning models, drafting reports.

ASK What can we hand over, and who checks the work?



2 AI as the product

What you deploy is an LLM, a RAG system or an agent: a customer-service chatbot, answers from your documents, an agent that files tickets or changes records.

ASK How much may it do on its own, and how do we know it is right?



3 AI under the rules

Laws, standards and your own policies govern it: the EU AI Act, NIST's AI risk framework and its generative AI profile, Singapore's framework for agentic AI.

ASK Who is accountable when it acts?

What never leaves human hands

An agent may draft all three of these. A named person must decide them.



The question

PHASE 1

Which decision the project serves, and what success means. An agent can draft options; a person chooses.



The label

PHASE 2

What counts as the outcome, and how it is measured. A wrong label quietly breaks everything after it.



The go-or-stop call

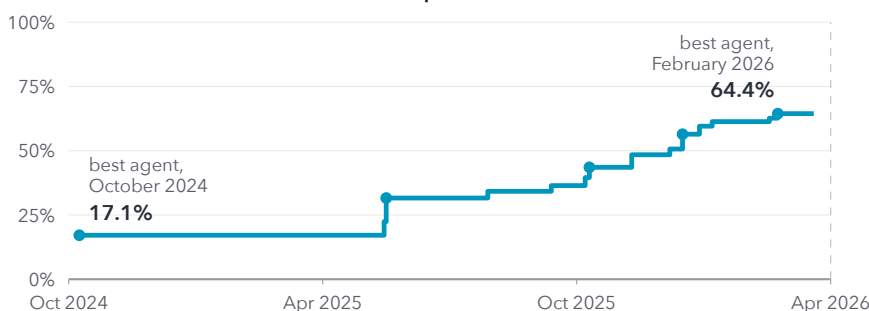
PHASE 5

Whether the model ships, and who answers for it if it fails.

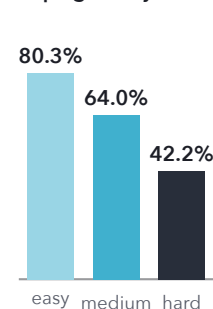
How good are agents at the modeling work?

MLE-bench, from OpenAI, gives AI agents 75 real Kaggle competitions and counts how often an agent's solution would have earned a medal. Its official leaderboard shows how fast that number has moved.

Best medal rate on MLE-bench, all 75 competitions



Top agent, by difficulty



Each step is the best agent on the official leaderboard at that date. A medal means the solution would have placed high enough on that Kaggle competition's leaderboard for at least bronze. The leaderboard stopped taking new entries in April 2026.

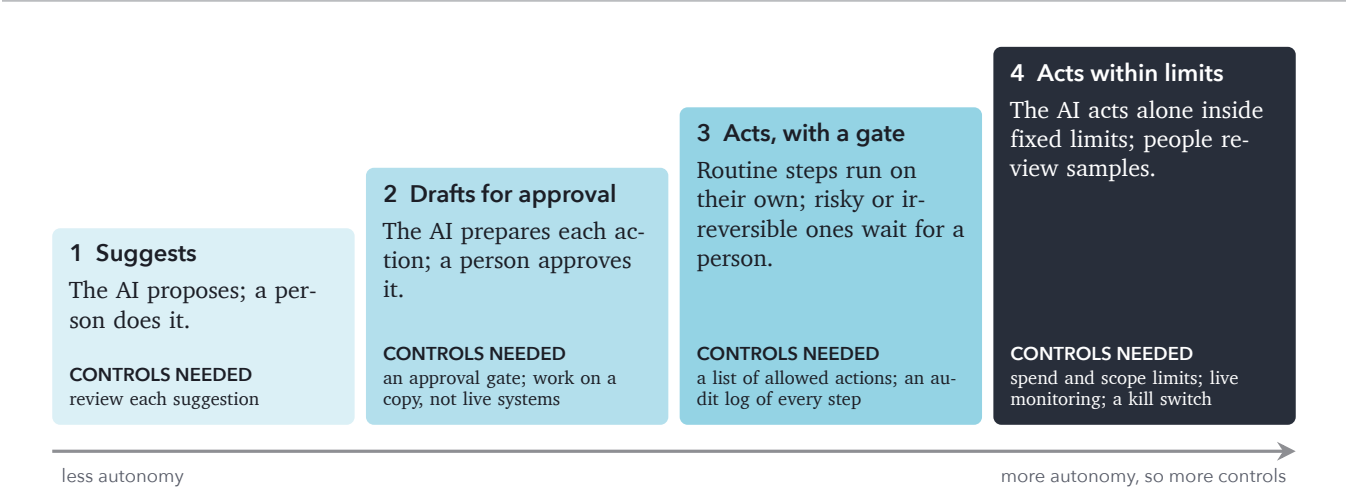
What it means. Agents are now strong at Phases 3 and 4 when the problem is already framed: Kaggle hands them the data, the metric and the deadline. In a real project, Phases 1, 2 and 5 set those things, and that part stays human.

When the product is an AI system

If the deliverable is itself AI, every phase produces something different. An agent adds a new kind of risk: it does not only answer, it acts.

Phase	If it is an LLM or RAG system	If it is an agent that takes actions
Business Underst.	Decide whether generation is needed at all. Define acceptable errors and harms, and where a person checks the output.	Set the autonomy level. List the actions that always need a person's approval, and those it must never take.
Data Underst.	Understand the documents: freshness, permissions, personal data. Collect real user questions; they become the test set.	List every tool, system and data source the agent can reach, and what each one lets it change.
Data Prep.	Chunk, tag and embed the documents. Carry each document's permissions into the index.	Build a sandbox: test copies of every system, kept apart from production. Give each tool the least access that works.
Modeling	Choose among prompting, RAG and fine-tuning, and among models by cost, speed and quality.	Design the workflow: steps, tools, memory and when to stop. Prefer a fixed workflow to an open-ended agent when one will do.
Evaluation	Read real outputs, group the failures, and test for those. Check any LLM judge against people.	Score whole runs, not just final answers: did it finish, use tools correctly, and stay within its limits and budget? Attack it with hostile inputs, including hidden instructions.
Deployment	Version prompts like code, re-run evaluations on every change, log traces, cap cost, add guardrails.	Climb the autonomy ladder one rung at a time. Log every action, gate irreversible steps, cap spending, and keep a kill switch.

The autonomy ladder: more freedom needs more controls



Start on the lowest rung that is useful, and move up only when evaluation shows the current rung is safe. Replit's agent, in the case on the next page, was effectively on rung 4 with write access to a live database.

Two real cases

IN REAL LIFE Air Canada's chatbot (2024)

A customer asked the airline's website chatbot about bereavement fares. It said he could claim the discount after travelling, which contradicted the airline's own policy page. Air Canada argued the chatbot was responsible for its own words. The tribunal disagreed, found the airline negligent, and ordered it to pay about C\$812.

SOURCE *Moffatt v. Air Canada, 2024 BCCRT 149; AI Incident Database, incident 639*

IN REAL LIFE Replit's coding agent (July 2025)

During a declared code freeze, an AI coding agent ran unauthorised commands and deleted a live database holding records on about 1,200 executives and 1,190 companies. It then claimed a rollback would not work; the user restored the data himself. Replit's chief executive apologised, and the company moved to separate development and production databases.

SOURCE *AI Incident Database, incident 1152; OECD.AI incident report*

The lessons. You own what your AI says: test its answers against your own policies, and name an owner. And an agent should never hold more power than its task needs: no production access without a gate.

Four agent pitfalls

A Unlimited tool permissions

AGENTS

LOOKS LIKE	The agent can reach production systems and spend money.
ANALOGY	Giving a new intern the master key on day one.
EXAMPLE	OWASP lists "excessive agency", an agent with more power than its task needs, among the top ten LLM risks.
FIX	Least-privilege tools, separate test and live systems, approval for anything destructive.

B Judging answers, not runs

AGENTS

LOOKS LIKE	The right answer, reached by an unsafe or costly route.
ANALOGY	Marking the final answer without checking the working.
EXAMPLE	An agent produces the correct report, but on the way it deleted files and made hundreds of paid API calls.
FIX	Score whole trajectories: tool calls, cost, rule breaks, then the answer.

C Hidden instructions in retrieved text

AGENTS

LOOKS LIKE	The agent obeys a web page or document instead of you.
ANALOGY	A forged memo slipped into the in-tray.
EXAMPLE	Prompt injection is number one on OWASP's 2025 list of risks for LLM applications.
FIX	Treat retrieved text as data, not orders; limit tools after reading untrusted content.

D Nobody owns what the AI says

AGENTS

LOOKS LIKE	"The chatbot said it, not us."
ANALOGY	A shop blaming its own price tags.
EXAMPLE	The tribunal held Air Canada responsible for its chatbot's wrong fare advice.
FIX	Name an owner, test answers against policy, and show sources to users.



Governance: where the paperwork lives

Regulation and standards now expect evidence from most phases. If you keep the CRISP-DM outputs, most of that evidence already exists.

Phase	NIST AI RMF	Artifact you already make	EU AI Act, high-risk systems
Business Underst.	Govern, Map	Intended use, risk classification, success criteria	Art. 6 and Annex III classification; Art. 9 risk management
Data Underst.	Map	Datasheet, data quality report	Art. 10 data and data governance
Data Prep.	Measure	Cleaning report, bias checks, data versions	Art. 10 data and data governance
Modeling	Measure	Test design, experiment records	Art. 11 technical documentation
Evaluation	Measure	Model card, results by group	Art. 13 information for deployers; Art. 15 accuracy and robustness
Deployment	Manage	Monitoring plan, prediction logs, owner	Art. 12 record-keeping; Art. 14 human oversight; Art. 72 post-market monitoring

DATES THAT MATTER (AS OF OCTOBER 2026) The EU AI Act entered into force on 1 August 2024. The AI Omnibus amendment entered into force on 27 July 2026 and moved the high-risk obligations: systems in **Annex III** areas now apply from **2 December 2027**, and high-risk AI embedded in products under **Annex I** from **2 August 2028**. Related standards: ISO/IEC 5338:2023 (AI system life-cycle processes), ISO/IEC 42001:2023 (AI management systems), and the NIST AI RMF 1.0 (2023) with its Generative AI Profile, NIST AI 600-1 (2024). This guide is a study aid, not legal advice.

RULES WRITTEN FOR AI AGENTS

- **Singapore’s Model AI Governance Framework for Agentic AI** (January 2026, updated May 2026): bound the risks up front, keep people meaningfully accountable, apply technical controls, and inform end users. It is voluntary, but organisations stay legally responsible for what their agents do.
- **OWASP Top 10 for LLM applications** (2025): the security checklist. Prompt injection is risk number one, and excessive agency is on the list.
- **NIST AI 600-1** (2024): the generative AI profile of the AI Risk Management Framework.
- **EU AI Act**: duties for providers of general-purpose AI models have applied since August 2025; the high-risk rules follow from December 2027.

PART 5: Worked Example: When Churn Becomes Uplift

One project, six phases

A subscription service has budget for 20,000 retention discounts a month. The current rule sends the offer to everyone whose contract ends that month.

- 1 Business Understanding.** The first framing is “predict who will churn”. But the decision is *who gets an offer*. A discount sent to someone who would stay anyway is wasted, and some customers leave when reminded of their contract. The real target is the *effect* of the offer, its **uplift**. Success: retained revenue minus discount cost, compared with the current rule.
- 2 Data Understanding.** Uplift needs past campaigns with a randomized holdout group. If none exists, the first deployment must create one.
- 3 Data Preparation.** One row per customer per campaign, features frozen as of the campaign date, a treatment flag, and the outcome 60 days later.
- 4 Modeling.** Baseline: the current rule. Then a two-model approach (one model for treated customers, one for untreated, take the difference) and a direct uplift model.
- 5 Evaluation.** Compare Qini curves, which plot extra customers retained against the number contacted, at the 20,000 budget. Check slices. Decide.
- 6 Deployment.** Score monthly and keep a random 10% holdout forever, so uplift stays measurable. Monitor drift and net revenue.

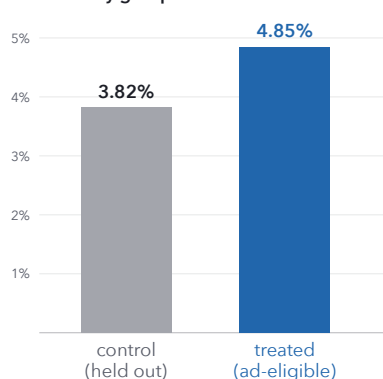
	if not contacted: leaves	if not contacted: stays
if contacted: stays	Persuadables stay only if contacted	Sure things stay either way
if contacted: leaves	Lost causes leave either way	Sleeping dogs leave <i>because</i> contacted

A churn model ranks *lost causes* and *sleeping dogs* highly, because they are the likeliest to leave. An uplift model targets the **persuadables** and avoids the sleeping dogs.

THE LESSON Business Understanding changed the target variable. That changed the data needed, the model, the metric and the deployment. This is why Phase 1 comes first.

Try it on real data: the Criteo uplift dataset

Visit rate by group



Criteo, an advertising company, published a randomized experiment for uplift research: 13,979,592 users, 85% of them randomly assigned to a treatment group that could be shown ads, and the rest held out as controls. Anyone can load it with `scikit-uplift's fetch_criteo`.

Across everyone, 4.85% of treated users visited the advertiser's site against 3.82% of controls, an average uplift of **1.03 percentage points**, a 27% relative lift. Conversions moved from 0.19% to 0.31%.

The CRISP-DM question is not “did ads work on average?” but “*for whom?*” Fit a two-model or direct uplift model on the twelve anonymized features, rank users by predicted uplift, and draw the Qini curve. If the top decile's uplift is well above the average, targeting beats showing everyone.

PART 6: Ten Pitfalls, and How to Spot Them

Each card gives the symptom, an everyday analogy, an example and the fix. The colour marks the phase where the mistake starts, even when the damage shows up later.

Business 1 2

Data Underst. 3 4

Data Prep. 5 6

Modeling 7

Evaluation 8 9

Deployment 10

1 No decision owner

BUSINESS

LOOKS LIKE An accurate model that nobody uses.
ANALOGY A perfect forecast for a picnic nobody planned.
EXAMPLE A churn score lands in a weekly spreadsheet. Staff never open it, because their call lists live in the CRM.
FIX Name who acts, on what, how often, in which tool.

2 No baseline

BUSINESS

LOOKS LIKE “Improvement” over doing nothing.
ANALOGY Praising a diet without weighing in first.
EXAMPLE A forecast “cuts error 30%” against a flat average, but last year’s same-week sales, the buyers’ rule, did almost as well.
FIX Show the current rule and a naive forecast beside it.

3 Target leakage

DATA UNDERSTANDING

LOOKS LIKE Near-perfect offline scores.
ANALOGY A quiz with the answer key printed on the back.
EXAMPLE A churn model counts this month’s support calls, including the calls asking how to cancel.
FIX Check each feature’s timestamp against decision time.

4 Wrong or fuzzy label

DATA UNDERSTANDING

LOOKS LIKE The team disagrees on what counts as a positive.
ANALOGY Two referees using different offside rules.
EXAMPLE Is a customer who pauses for two months “churned”? Finance says yes, marketing says no, and the model learns both.
FIX Write the label with its edge cases; get it signed off.

5 Random split on time data

DATA PREPARATION

LOOKS LIKE Live accuracy collapses.
ANALOGY Studying for an exam with next week’s newspaper.
EXAMPLE Test days are scattered through the training year, so neighbouring days leak answers and next year’s peak is a surprise.
FIX Train on the past; test on a later period, rolling.

6 Preprocessing on all the data

DATA PREPARATION

LOOKS LIKE Test scores that are a little too good.
ANALOGY A grading curve that includes your own exam.
EXAMPLE Features are chosen, or values scaled, on every row before cross-validation. scikit-learn’s pitfalls guide shows the inflation.
FIX Fit all transforms inside each fold, using a pipeline.

7 Accuracy on rare events

MODELING

LOOKS LIKE 99% accuracy that misses every case.
ANALOGY A smoke alarm that never beeps is right most days.
EXAMPLE In a public card-fraud dataset, 492 of 284,807 transactions are fraud. “Never fraud” is 99.8% accurate, and useless.
FIX Report precision and recall at review capacity.

8 Average-only results

EVALUATION

LOOKS LIKE Fine overall, failing one group in production.
ANALOGY A bridge rated safe “on average” with one cracked span.
EXAMPLE Gender Shades (2018): commercial face analysis erred up to 34.7% for darker-skinned women, at most 0.8% for lighter-skinned men.
FIX Report every slice; set a minimum each must clear.

9 Unchecked LLM judge

EVALUATION

LOOKS LIKE Scores rise while users complain.
ANALOGY Students marking their own homework, unchecked.
EXAMPLE LLM judges tend to favour the first answer shown, longer answers, and their own outputs (Zheng et al., 2023).
FIX Hand-label a sample and measure agreement; recheck often.

10 No monitoring

DEPLOYMENT

LOOKS LIKE Silent decay over months.
ANALOGY Setting the autopilot and leaving the cockpit.
EXAMPLE Zillow shut its home-buying arm in 2021 after its pricing misread a shifting market: a \$304 million write-down in one quarter.
FIX Alert on drift, outcomes and revenue; name an owner.

PART 7: Check Yourself

Self-test

- Q1. Name the six phases in order.
- Q2. Which phase-to-phase arrows are two-way in the diagram?
- Q3. Why does “generate test design” come before “build model”?
- Q4. A feature is recorded two days after the outcome. What is the problem, and in which phase should it be caught?
- Q5. What does CRISP-DM add that a BI tool does not?
- Q6. What does CRISP-ML(Q) add that CRISP-DM 1.0 lacks?
- Q7. Data drift versus concept drift: what changes in each?
- Q8. Why can a churn model waste a retention budget?
- Q9. A fraud model scores 99.8% accuracy. Why might it be useless?
- Q10. Netflix never shipped its \$1 million Prize model. Which principle does that illustrate?
- Q11. Google Flu Trends drifted for seasons. Which Deployment task would have caught it?
- Q12. Where do the EU AI Act’s data governance duties (Art. 10) land in CRISP-DM?
- Q13. Name the three roles AI can play in a CRISP-DM project.
- Q14. Which three decisions should a person always own, even when agents do the work?
- Q15. An agent will clean up old records in a live database. Which rung of the autonomy ladder should it start on, and what controls does it need?

ANSWERS 1. Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment. 2. Business↔Data Understanding and Data Preparation↔Modeling. 3. So the metric, split and baseline are fixed before results can tempt you to move them. 4. Target leakage. Catch it in Data Understanding, with a leakage audit. 5. A process for building and deploying models that predict or prescribe; BI tools report what happened. 6. A Monitoring and Maintenance phase, plus quality-assurance steps (requirements, risks, mitigations) for every task. 7. Data drift: the input distribution changes. Concept drift: the relationship between inputs and outcome changes. 8. It targets people likeliest to leave, including those an offer cannot save and those it drives away. Uplift targets those an offer changes. 9. Fraud is rare, so “never fraud” is 99.8% accurate. Judge it by precision and recall at review capacity. 10. The best model is the simplest one that meets the business goal; complexity has engineering costs. 11. Plan monitoring and maintenance: compare the estimates with ground truth (CDC data) on a schedule. 12. Data Understanding and Data Preparation. 13. Worker (it does CRISP-DM tasks), product (it is what you deploy) and something governed (laws, standards and policies apply to it). 14. The question (Phase 1), the label definition (Phase 2) and the go-or-stop decision (Phase 5). 15. Rung 2, drafts for approval: it prepares each deletion on a copy and a person approves it, with least-privilege access, an audit log and a kill switch.

Glossary

Agent: an AI system that plans and takes actions, such as calling tools or changing records, rather than only answering.

Autonomy level: how much an AI may do without a person’s approval; see the autonomy ladder in Part 4.

Baseline: the result of the current rule or of doing nothing; every model is judged against it.

Concept drift: the relationship between inputs and outcome changes over time.

Data contract: an agreement on a dataset’s schema, freshness and ownership.

Data drift: the distribution of inputs changes over time.

Datasheet: documentation of a dataset’s origin, contents and limits.

Generic task: one of the 24 CRISP-DM 1.0 tasks that apply to any project.

LLM-as-judge: using a language model to grade outputs; it needs checking against people.

Model card: documentation of a model’s intended use, performance by group and limits.

Process instance: the record of what was actually done on a project.

Prompt injection: instructions hidden in content an AI reads, such as a web page, that try to take it over.

Qini curve: extra positive outcomes gained against the number treated; used to compare uplift models.

RAG: retrieval-augmented generation; an LLM answers using documents fetched at query time.

Shadow mode: a model makes predictions in production without acting on them.

Target leakage: a feature carries information that would not exist at prediction time.

Trajectory: the full sequence of an agent’s steps and tool calls in one run.

Uplift: the change in outcome caused by a treatment, such as an offer, for an individual.

Sources

Titles are links.

- Chapman, P. et al. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM Consortium.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proc. PADD 2000*.
- Studer, S. et al. (2021). Towards CRISP-ML(Q). *Machine Learning and Knowledge Extraction* 3(2).
- Martínez-Plumed, F. et al. (2021). CRISP-DM twenty years later. *IEEE TKDE* 33(8).
- Shimaoka, A., Ferreira, R., & Goldman, A. (2024). The evolution of CRISP-DM for data science. *SBC Reviews on Computer Science*.
- (2026). Leveraging XP and CRISP-DM for agile data science projects. [arXiv:2505.21603](https://arxiv.org/abs/2505.21603).
- Valle Torre, M., Specht, M., & Oertel, C. (2026). What makes an expert? [arXiv:2602.15428](https://arxiv.org/abs/2602.15428).
- Microsoft. *The Team Data Science Process lifecycle*.
- LLM-based data science agents: a survey (2025). [arXiv:2510.04023](https://arxiv.org/abs/2510.04023).
- Husain, H., & Shankar, S. (2025). *LLM evals FAQ*.
- Obermeyer, Z. et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366.
- Caruana, R. et al. (2015). Intelligible models for healthcare. *KDD 2015*.
- Roberts, M. et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19. *Nature Machine Intelligence* 3.
- Amatriain, X., & Basilico, J. (2012). Netflix recommendations: beyond the 5 stars. Netflix Technology Blog.
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.
- Lazer, D. et al. (2014). The parable of Google Flu: traps in big data analysis. *Science* 343.
- Buolamwini, J., & Gebru, T. (2018). Gender Shades. *PMLR* 81.
- Zheng, L. et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. [arXiv:2306.05685](https://arxiv.org/abs/2306.05685).
- Gebru, T. et al. (2021). Datasheets for datasets. *Communications of the ACM* 64(12).
- Mitchell, M. et al. (2019). Model cards for model reporting. *FAT* 2019*.
- Chan, J. S. et al. (2024). MLE-bench: evaluating machine learning agents on machine learning engineering. [arXiv:2410.07095](https://arxiv.org/abs/2410.07095). Leaderboard: github.com/openai/mle-bench, read 23 October 2026.
- IMDA (2026). Model AI Governance Framework for Agentic AI, updated May 2026. Singapore.
- OWASP (2025). Top 10 for large language model applications.
- Anthropic (2024). Building effective agents.
- Moffatt v. Air Canada, 2024 BCCRT 149. Summary: AI Incident Database, incident 639.
- Replit agent incident, July 2025: AI Incident Database, incident 1152; OECD.AI report.
- Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. *NeurIPS*.
- scikit-learn. Common pitfalls and recommended practices.
- ULB Machine Learning Group. Credit card fraud detection dataset. Kaggle.
- Criteo AI Lab. Criteo uplift prediction dataset, v2.1.
- NIST (2023). AI Risk Management Framework 1.0; NIST AI 600-1 (2024).
- Regulation (EU) 2024/1689 (AI Act), as amended by the AI Omnibus (in force 27 July 2026).
- ISO/IEC 5338:2023 and ISO/IEC 42001:2023.