


Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

# Course Introduction & What Is Data Mining?

*Lecture 1 · Wednesday 16 September 2026*

**Dr. Innocent Nyalala** | BS Semester 5 (DSAI) | IIT Madras Zanzibar

# Session Outline

IIT Madras Zanzibar

1. (0–12 min) The course: who, when, what, how you are assessed
2. (12–20 min) Why data mining: the data flood, in numbers
3. (20–35 min) What data mining is: definition and the 7-step KDD process
4. (35–45 min) What kinds of data, and what kinds of patterns
5. (45–60 min) First pattern by hand: support, confidence and lift on 5 receipts
6. (60–70 min) Applications, ethics, and why scale changes everything
7. (70–75 min) Exercise, summary and what's next

# Learning Outcomes

IIT Madras Zanzibar

*By the end of this session, you will be able to:*

| | |
|---|---|
| **Explain** | what data mining is, in one sentence, and why we need it now |
| **List** | the 7 steps of the KDD process and say what happens at each |
| **Name** | the main kinds of data and the main kinds of patterns we can mine |
| **Compute** | support, confidence and lift for a rule, by hand |
| **Recognise** | two ethical risks: misleading accuracy, and unfair errors |

# Course Introduction & What Is Data Mining?

IIT Madras Zanzibar

→ **The Course: People, Plan, Assessment**

Why Data Mining? The Data Flood

What Is Data Mining? The KDD Process

What Can Be Mined, and Which Patterns

A First Pattern by Hand: Support, Confidence, Lift

Applications, Ethics and Scale

In-Class Exercise

# Course Logistics

IIT Madras Zanzibar

| Item | Details |
|---|---|
| **Course** | Z3003 Data Mining: Scaling to Large Data · BS Semester 5 (DSAI) |
| **Part 1** | Dr. Innocent Nyalala · 16 Sep – 13 Nov 2026 (8 weeks) |
| **Part 2** | Prof. Mahendran · from 18 Nov (advanced classification and clustering, deep learning, outliers) |
| **Sessions** | Wed 3:25–4:40 pm (75 min) · Thu 2:00–3:15 pm (75 min) · Fri 9:00–9:50 am (50 min) |
| **Tutorial** | Thu 4:45–5:30 pm, every week — hands-on notebooks; bring a laptop |
| **Texts** | Leskovec, Rajaraman & Ullman, Mining of Massive Datasets (free: mmds.org) · Han, Pei & Tong, Data Mining, 4th ed. |
| **Software** | Python 3, Jupyter, pandas, NumPy, scikit-learn, mlxtend; PySpark ideas for scale |

# Where This Course Is Going: The 8-Week Map

IIT Madras Zanzibar

| Week | Core topic | Scaling angle |
|---|---|---|
| 1 | KDD process, data types, summary statistics | why an exact median does not scale; approximation |
| 2 | Correlation, graphics, distance, data cleaning | sampling; comparing millions of pairs |
| 3 | Normalization, discretization, PCA, sampling | fewer columns and rows, same information |
| 4 | Data warehouses, cubes and OLAP | pre-computing summaries; iceberg cubes |
| 5 | Frequent patterns: Apriori, FP-Growth | fewer passes over the data |
| 6–7 | Classification: trees, Naïve Bayes, evaluation | split-finding on large tables; ensembles |
| 8–9 | Clustering: k-means, hierarchical, DBSCAN | mini-batch and streaming variants |

# Assessment Structure

IIT Madras Zanzibar

**20%**

**Quiz 1**

16–18 Oct, in class, 60 min · Weeks 1–4 · calculators allowed

**20%**

**Mini-project**

dataset checkpoint 16 Oct · due Fri 13 Nov, 11:59 pm · notebook + 1-page summary

**60%**

**Quiz 2 + Final**

Part 2, run by Prof. Mahendran

**Mini-project in one line:** choose a real dataset, clean and reduce it, apply one mining technique from Weeks 5–9, and explain what you found to a non-expert. A rubric is published with the brief.

# Policies & Expectations

IIT Madras Zanzibar

**1 Attendance**

all three weekly sessions and the Thursday tutorial

**2 Integrity**

work is your own; cite all external code, data and text

**3 Deadlines**

fixed; no late work unless an extension was agreed in advance

**4 AI tools**

allowed to explain and debug; declare use in a 'Tools used' cell

**5 Participate**

examples are solved together — ask, try, be wrong early

**6 Handover**

from 18 Nov, Prof. Mahendran leads Part 2

**Common thread:** you should be able to explain every line you submit.

# Course Introduction & What Is Data Mining?

The Course: People, Plan, Assessment

→ **Why Data Mining? The Data Flood**

What Is Data Mining? The KDD Process

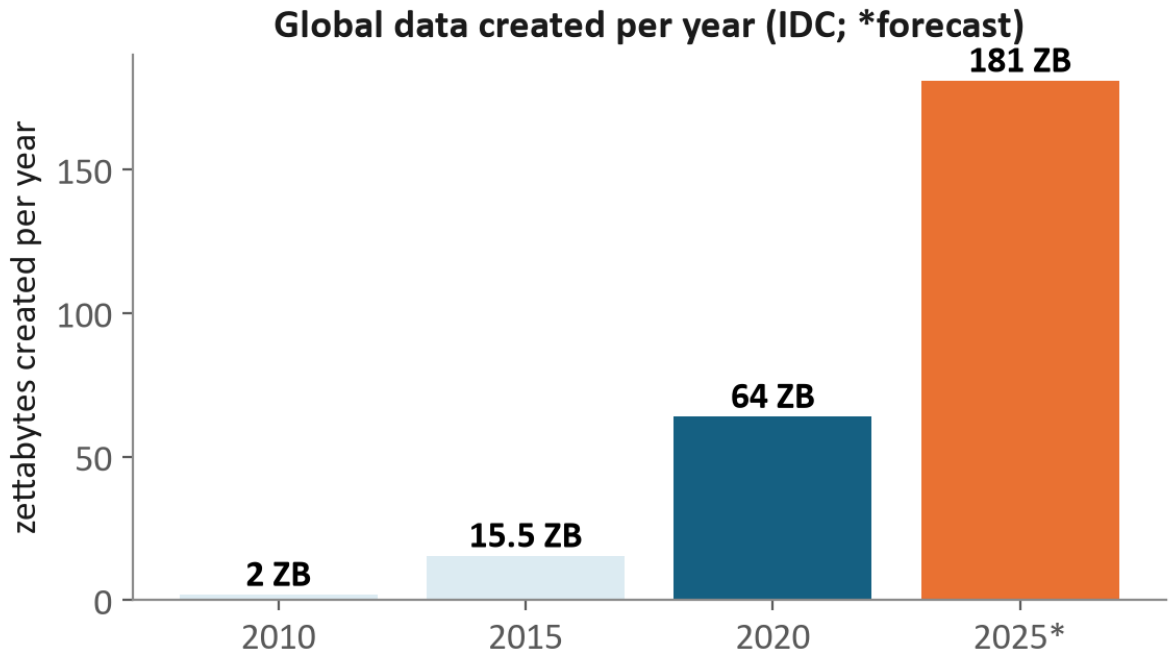
What Can Be Mined, and Which Patterns

A First Pattern by Hand: Support, Confidence, Lift

Applications, Ethics and Scale

In-Class Exercise

# Why Data Mining? We Are Drowning in Data




1. **1 zettabyte**
   = $10^{21}$ bytes = one trillion gigabytes

2. **Growth**
   about 90× in 15 years

3. **The gap**
   people cannot read it; computers must search it

**Drowning in data, starving for knowledge:** data mining turns rows nobody can read into patterns people can act on.

# Close to Home: Data in Tanzania and Zanzibar

IIT Madras Zanzibar

**1 Mobile money**

65.7 M accounts (Mar 2025), up from 29.8 M in 2020; billions of payments a year

**2 Agriculture**

farm surveys: crop, size, yield, rainfall, inputs

**3 Health**

clinic visits, malaria tests, vaccination records

**4 Tourism**

arrivals, hotel bookings, reviews

**5 Trade**

port and customs logs: cloves, seaweed, imports

**6 Our running dataset**

500 Zanzibar farms → later 10,000 farms, 200,000 seasons

**Common thread:** each record is small; together they hide patterns no person could find by reading.

# Course Introduction & What Is Data Mining?

IIT Madras Zanzibar

The Course: People, Plan, Assessment

Why Data Mining? The Data Flood

→ **What Is Data Mining? The KDD Process**

What Can Be Mined, and Which Patterns

A First Pattern by Hand: Support, Confidence, Lift

Applications, Ethics and Scale

In-Class Exercise

# What Is Data Mining?

IIT Madras Zanzibar

> Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.
>
> *Han, Pei & Tong, Data Mining: Concepts and Techniques, 4th ed.*

**Valid**

holds on new data, not only on the sample

**Novel**

not already known — 'bigger farms produce more' is not news

**Useful**

someone can act on it — 'stock urea next to maize seed'

**Understandable**

a person can read it and explain it

**Analogy — gold mining:** move tonnes of rock (data) to find a few grams of gold (knowledge). Strictly, we mine knowledge from data, so it is also called Knowledge Discovery from Data (KDD).

# Data Mining Is Not Just a Database Query

IIT Madras Zanzibar

| | Database query | Classical statistics | Data mining |
|---|---|---|---|
| **You supply** | the exact question | one hypothesis | the data and a goal |
| **You get** | a count, a list, a sum | accept / reject the hypothesis | patterns you had not thought of |
| **Example** | how many clove farms in Pemba North? | does fertilizer raise yield? | which farms behave alike? what predicts failure? |
| **Risk** | none — it is exact | small, if well designed | chance patterns: search enough and something always 'fits' |

**Watch out:** test 1,000 random patterns at the 5% level and about 50 look 'significant' by chance alone.

# The KDD Process: 7 Steps from Raw Data to Knowledge

IIT Madras Zanzibar

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| **Cleaning** | **Integration** | **Selection** | **Transformation** | **Mining** | **Evaluation** | **Presentation** |
| fix errors, fill gaps | join sources | keep the relevant | scale, aggregate | find patterns | keep the real ones | show for action |

**Steps 1–4: preprocessing (Weeks 2–4) — most of the time**

**Steps 5–7: mining and use (Weeks 5–9)**

**Analogy — cooking:** cleaning and cutting vegetables (steps 1–4) takes most of the time; cooking (step 5) is short; tasting and serving (6–7) decide whether anyone eats it.

# KDD Worked Example: Mobile-Money Fraud

IIT Madras Zanzibar

| KDD step | Concrete action on mobile-money payments |
|---|---|
| 1 Cleaning | drop payments with broken timestamps or negative amounts |
| 2 Integration | join SIM-registration records to the payment log (account age) |
| 3 Selection | last 12 months; payments above a minimum value |
| 4 Transformation | time → hour of day; amount → how unusual for this account |
| 5 Mining | train a classifier: fraud or not (Weeks 6–7) |
| 6 Evaluation | how many frauds caught? how many honest customers blocked? |
| 7 Presentation | dashboard: 50 riskiest payments per hour for an analyst |

# Why '99.5% Accurate' Can Catch Zero Fraud

IIT Madras Zanzibar

1. **10,000 payments**
   50 are fraud (0.5%), 9,950 honest
2. **Lazy model**
   always answers 'not fraud'
3. **Accuracy**
   9,950 correct ÷ 10,000 = 99.5%
4. **Frauds caught**
   0 of 50 — useless

$$\text{accuracy} = \text{correct} \div \text{total}$$

*looks excellent when one class is rare*

$$\text{recall} = \text{frauds caught} \div \text{all frauds}$$

*here 0 ÷ 50 = 0 (Week 7)*

**Lesson:** a single accuracy number can hide total failure. Always ask: accurate on what?

# Course Introduction & What Is Data Mining?

The Course: People, Plan, Assessment

Why Data Mining? The Data Flood

What Is Data Mining? The KDD Process

→ **What Can Be Mined, and Which Patterns**

A First Pattern by Hand: Support, Confidence, Lift

Applications, Ethics and Scale

In-Class Exercise

# What Kinds of Data Can Be Mined?

IIT Madras Zanzibar

**1 Tables**

rows × columns: our 500 farms × 9 attributes

**2 Transactions**

one basket per row: a shop receipt

**3 Time series & streams**

daily rainfall; payments per second

**4 Spatial**

farm GPS points; district maps

**5 Text & web**

SMS, reviews, news, web pages

**6 Graphs**

who pays whom; who follows whom

**Common thread:** this course works mostly with tables and transactions; the scaling ideas carry over.

# Our Running Example: 500 Zanzibar Farms

IIT Madras Zanzibar

| farm | region | crop | size (ha) | yield (t) | rain (mm) | price ($/kg) | fertilizer | revenue ($) |
|---|---|---|---|---|---|---|---|---|
| 1 | Pemba North | Cassava | 0.25 | 6.11 | 1781 | 0.26 | No | 1,841 |
| 2 | North Unguja | Rice | 4.18 | 2.71 | 1766 | 0.69 | Yes | 2,526 |
| 3 | South Unguja | Coconut | 0.68 | 3.05 | 1214 | 0.40 | Yes | 1,224 |
| 4 | Zanzibar Urban | Clove | 1.40 | 1.68 | 1250 | 8.21 | Yes | 13,778 |
| 5 | Pemba North | Clove | 3.60 | 1.51 | 1399 | 8.12 | Yes | 12,271 |
| … | … | … | … | … | … | … | … | … |

1. **500 rows**
   one per farm (an object, record or tuple)

2. **9 columns**
   attributes (features, variables)

3. **Labels and numbers**
   type decides what maths is allowed — Thursday

# What Kinds of Patterns? Five Task Families

IIT Madras Zanzibar

| | | | |
|---|---|---|---|
| 1 | **Description** | summarise: mean yield by crop | **Wk 1–4** |
| 2 | **Association** | items that occur together: maize seed → urea | **Wk 5** |
| 3 | **Classification** | predict a label from examples: repay or default | **Wk 6–7** |
| 4 | **Clustering** | group similar records, no labels | **Wk 8–9** |
| 5 | **Outliers** | find the unusual: a strange payment | **Wk 2, 9** |

**Thursday:** each family worked by hand on a small example.

# Course Introduction & What Is Data Mining?

IIT Madras Zanzibar

The Course: People, Plan, Assessment

Why Data Mining? The Data Flood

What Is Data Mining? The KDD Process

What Can Be Mined, and Which Patterns

→ **A First Pattern by Hand: Support, Confidence, Lift**

Applications, Ethics and Scale

In-Class Exercise

# Association: Which Items Are Bought Together?

IIT Madras Zanzibar

| Receipt | Items bought |
|---|---|
| T1 | {seed, urea} |
| T2 | {seed, hoe, gloves, sacks} |
| T3 | {urea, hoe, gloves, pesticide} |
| T4 | {seed, urea, hoe, gloves} |
| T5 | {seed, urea, hoe, pesticide} |

1. **Transaction**
   one receipt (row)

2. **Item**
   one product in it

3. **Itemset**
   a set of items, e.g. {hoe, gloves}

4. **Rule A → B**
   'if A is bought, B is too'

**Question:** is hoe → gloves a real pattern? We need three numbers: support, confidence and lift.

# Step 1 — Support: How Often Does It Happen?

$$\text{support}(A) = (\text{number of receipts containing } A) \div N$$

*N = total number of receipts = 5*

| Itemset | Receipts that contain it | Count | Support |
|---|---|---|---|
| {hoe} | T2, T3, T4, T5 | 4 | 4/5 = 0.8 |
| {gloves} | T2, T3, T4 | 3 | 3/5 = 0.6 |
| {hoe, gloves} | T2, T3, T4 | 3 | 3/5 = 0.6 |

**Meaning:** 60% of all receipts contain both a hoe and gloves.

# Step 2 — Confidence: How Reliable Is the Rule?

IIT Madras Zanzibar

$$\text{confidence}(A \rightarrow B) = \text{support}(A \cup B) \div \text{support}(A)$$

*$A \cup B$ = a receipt containing A and B together*

1. **Receipts with a hoe**
   T2, T3, T4, T5 → 4
2. **Of those, with gloves too**
   T2, T3, T4 → 3
3. **Confidence(hoe → gloves)**
   3 ÷ 4 = 0.6 ÷ 0.8 = 0.75
4. **Reverse rule**
   gloves → hoe = 3 ÷ 3 = 1.0 — direction matters

# Step 3 — Lift: Is It More Than Popularity?

$$\text{lift}(A \rightarrow B) = \text{confidence}(A \rightarrow B) \div \text{support}(B)$$

*how many times more likely B is when A is present*

| Rule | Confidence | support(B) | Lift | Meaning |
|---|---|---|---|---|
| hoe → gloves | 0.75 | 0.6 | **1.25** | gloves 25% more likely with a hoe |
| seed → urea | 0.75 | 0.8 | **0.94** | urea is just popular — no real link |

**lift > 1**
positive link

**lift = 1**
independent

**lift < 1**
negative link

# Course Introduction & What Is Data Mining?

The Course: People, Plan, Assessment

Why Data Mining? The Data Flood

What Is Data Mining? The KDD Process

What Can Be Mined, and Which Patterns

A First Pattern by Hand: Support, Confidence, Lift

→ **Applications, Ethics and Scale**

In-Class Exercise

# Data Mining Everywhere

IIT Madras Zanzibar

**1 Banking & mobile money**

fraud alerts, credit scoring for small loans

**2 Retail**

what to stock together; who to send offers

**3 Agriculture**

which inputs raise yield; early warning of crop failure

**4 Health**

disease outbreaks from clinic visits; readmission risk

**5 Tourism**

demand forecasting; review analysis

**6 Telecom**

which customers are about to leave

**Common thread:** each is one of the five task families applied to a large table.

# Ethics: Accurate Overall, Unfair to a Group

IIT Madras Zanzibar

| Group (did not re-offend) | Wrongly flagged high-risk |
|---|---|
| Black defendants | 44.9% |
| White defendants | 23.5% |

1. **Same accuracy**
   both groups about equally 'accurate' overall
2. **Different errors**
   one group wrongly flagged twice as often
3. **Local risk**
   a loan model trained on Unguja may fail Pemba farmers

**Rule for this course:** report errors by group, protect personal data, and ask who is harmed when the model is wrong. Source: Angwin et al., 'Machine Bias', ProPublica, 2016.

# Why Scale Changes Everything

IIT Madras Zanzibar

| Items in shop | Possible itemsets ($2^n - 1$) |
|---|---|
| 5 (our receipts) | 31 |
| 20 | 1,048,575 |
| 100 | $\approx 1.3 \times 10^{30}$ |
| 10,000 | more than atoms in the universe |

1. **Brute force**
   count every itemset — impossible beyond a few dozen items

2. **Apriori idea (Week 5)**
   if {hoe} is rare, every set containing a hoe is rare — skip them all

**The course in one line:** same questions as classic data mining, answered with smarter counting so they still work on millions of rows.

# Course Introduction & What Is Data Mining?

IIT Madras Zanzibar

The Course: People, Plan, Assessment

Why Data Mining? The Data Flood

What Is Data Mining? The KDD Process

What Can Be Mined, and Which Patterns

A First Pattern by Hand: Support, Confidence, Lift

Applications, Ethics and Scale

→ **In-Class Exercise**

# In-Class Exercise (5 min, pairs)

IIT Madras Zanzibar

**IN-CLASS TASK**

Use the five receipts: T1 {seed, urea} · T2 {seed, hoe, gloves, sacks} · T3 {urea, hoe, gloves, pesticide} · T4 {seed, urea, hoe, gloves} · T5 {seed, urea, hoe, pesticide}

1. Compute the support of {urea, pesticide}.
2. Compute the confidence of pesticide → urea, and of urea → pesticide.
3. Compute the lift of pesticide → urea. Is the link positive, none, or negative?
4. Name the KDD step in which you would check this rule on next month's receipts.

# Solution

IIT Madras Zanzibar

1. **Support {urea, pesticide}**
   receipts T3, T5 → 2/5 = 0.4

2. **Confidence**
   pesticide → urea = 0.4 ÷ 0.4 = 1.0;  urea → pesticide = 0.4 ÷ 0.8 = 0.5

3. **Lift**
   1.0 ÷ support(urea) = 1.0 ÷ 0.8 = 1.25 > 1 → positive link

4. **KDD step**
   6, pattern evaluation: does it still hold on new data?

# Summary & Exam Checkpoints

IIT Madras Zanzibar

**Data mining**

finding valid, novel, useful, understandable patterns in large data

**KDD: 7 steps**

clean, integrate, select, transform, mine, evaluate, present — steps 1–4 take most time

**Support · confidence · lift**

how often · how reliable · more than popularity? (lift > 1)

**Accuracy is not enough**

rare classes and unfair errors hide behind one number

# What's Next

IIT Madras Zanzibar

**Thu** **Lecture 2 (17 Sep): Tasks & Data Types**
the five task families by hand; nominal, ordinal, numeric attributes

**Tut** **Tutorial 1 (Thu 4:45 pm)**
load the 500-farm file in pandas; classify every column

**Fri** **Lecture 3 (18 Sep): Summary Statistics**
mean, median, spread — and why the median does not scale

**Before tomorrow:** install Python 3 + Jupyter; skim Han et al. Ch. 1 (or MMDS Ch. 1).

# References & Further Reading


- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed., Ch. 1. Morgan Kaufmann.
- Leskovec, J., Rajaraman, A. & Ullman, J. (2020). Mining of Massive Datasets, 3rd ed., Ch. 1 & 6. mmds.org.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3).
- Agrawal, R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. VLDB.
- Angwin, J. et al. (2016). Machine Bias. ProPublica.
- IDC Global DataSphere forecasts; Bank of Tanzania / TCRA quarterly statistics (2024–25).