CSE2525 · DATA MINING · 2026–27

Introduction to Data Mining

What data mining is, how this course is run, and why you should tread with caution — this year with an information-access focus.

A. Anand (head lecturer) · with J. Urbano & N. Tömen · lecture 1 study deck

PART I

There are no correct answers — only correct design decisions

Data mining is a craft of choices under messy data. The next two stories — a cholera map in 1854 and an ancient library — are really design decisions about what structure to look for.

QUESTION

London, 1854: hundreds are dying of cholera and no one knows why. All you have is a list of the **addresses of the dead**. How would you find the cause?

No germ theory yet, no microscope data — just locations on a map.

Think — then → for the answer.

LONDON, 1854

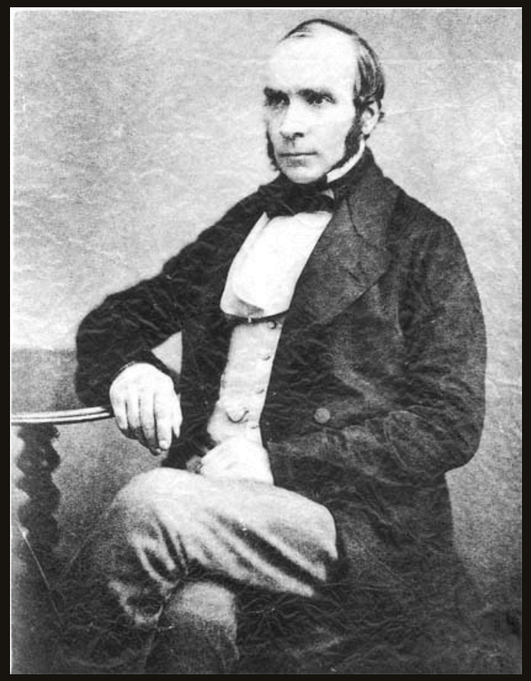
The first data miner — John Snow

THE PROBLEM

A cholera outbreak tore through Soho — **~600 people dead in days**. The accepted theory was *"bad air"* (miasma). Nobody knew the real cause.

— Snow, a physician, suspected the water — but had no proof.

— All he had was a list of addresses of the dead. How do you find a cause in that?

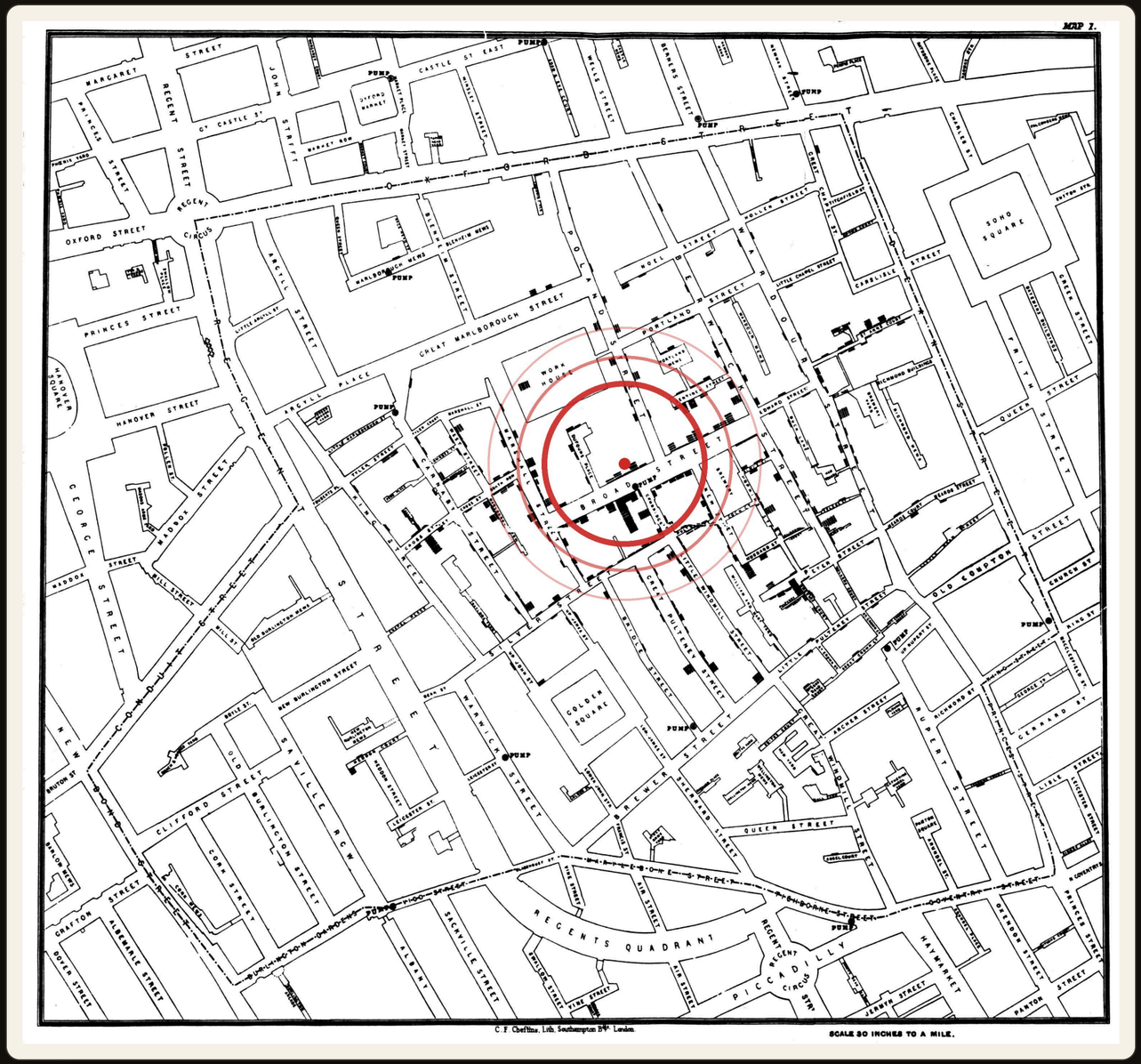


HE PLOTTED EVERY DEATH ON A MAP



Each red dot is a cholera death. Snow marked them street by street — and a dense cluster began to emerge in the middle of Soho, while whole blocks nearby stayed clear.

ONE PUMP EXPLAINED IT ALL



The deaths ringed one water pump on Broad Street. Snow had its handle removed — and the outbreak stopped. The design decision — plot deaths in *space*, not in a table — made the cause visible.

ALEXANDRIA, 3RD CENTURY BC

Another origin story — the first retrieval system

~500,000 scrolls
AN UNUSABLE PILE

→

The Pinakes catalogue
GENRE · AUTHOR · TITLE

→

Any scroll, found
INFORMATION RETRIEVAL

HOW THE PINAKES ORGANIZED IT

Callimachus catalogued the Library of Alexandria — the Pinakes (“tables”), arguably the first information-retrieval system. It split ~half a million scrolls into about ten genres (rhetoric, law, epic, history, medicine…); within each, authors were listed alphabetically, each with a short biography; and every work was recorded by its title, opening line, and line count. Any scroll was now locatable — the first real index.

THE COURSE DEFINITION

What data mining is — in this course

WORKING DEFINITION

Data mining is the **craft of deciding what structure to look for** in **large, messy, unlabelled** data — and **finding it at scale**.

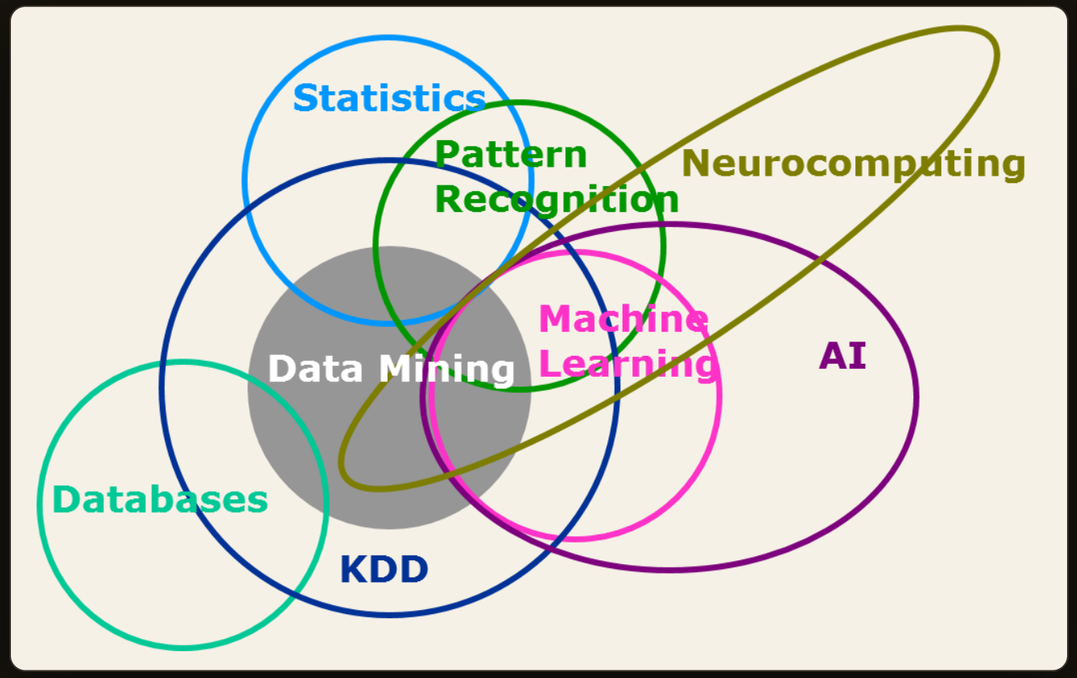
— Deciding what to look for — the design decision, as in Snow's map and the Pinakes.

— Large, messy, unlabelled · at scale — the part of the problem this course focuses on.

A sharper slice of the broader textbook definition ("extract useful information from data"): we emphasise scale and no labels.

BLOGS.SAS.COM

Data mining sits at an intersection



It borrows from statistics, databases, machine learning, pattern recognition and AI — the overlap at the heart of Knowledge Discovery in Databases (KDD).

THE BROADER TEXTBOOK VIEW

The textbook definition

CLASSIC DEFINITION

The primary goal of data mining is to **extract useful information from a large volume of data** and transform it into an **understandable structure** for further use.

— “…searching and analyzing a large batch of raw data to identify patterns and extract useful information.”

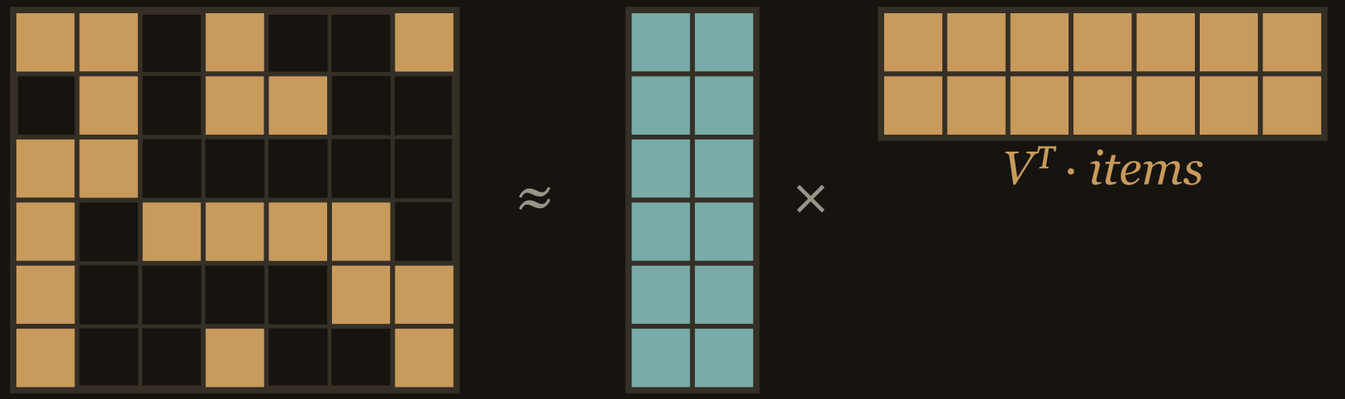
— “…the development of models for data in order to extract information from it.”

— “…analyzing data from different perspectives and summarizing it into useful information.”

— And crucially — it is done by humans.

GUESS FIRST · ① RECOMMENDATION

How does Spotify know what you'll like?



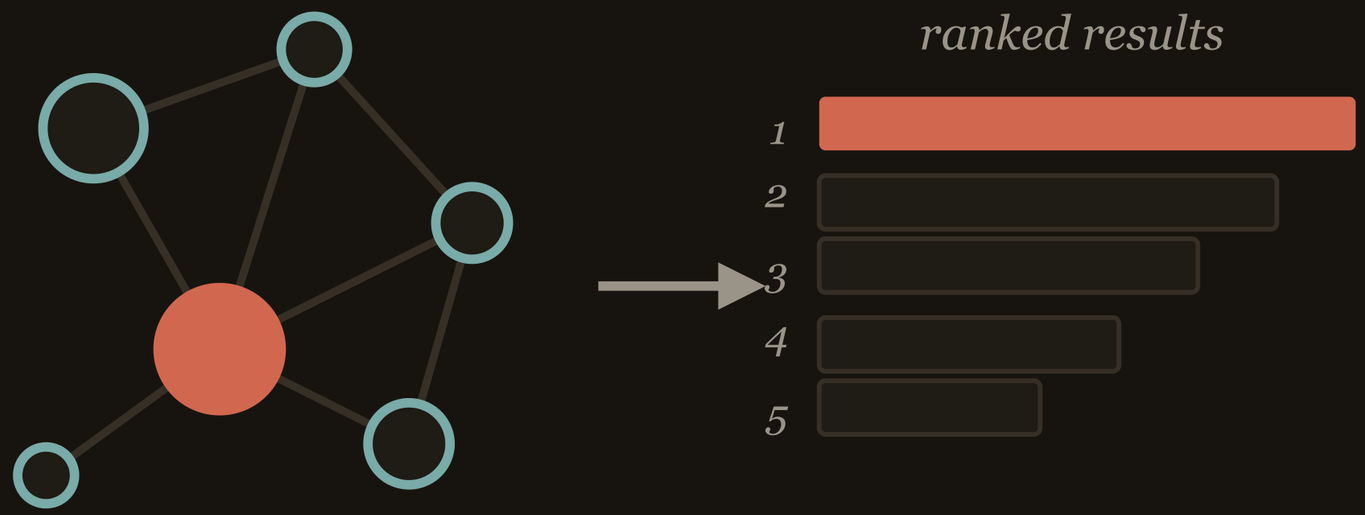
R · ratings (mostly ?) *U · users*

similar users like similar items

Distances / similarity | Embeddings | Matrix factorization | Clustering | Dimensionality reduction (PCA)

GUESS FIRST · ② WEB SEARCH

How does Google rank the entire web for your query?

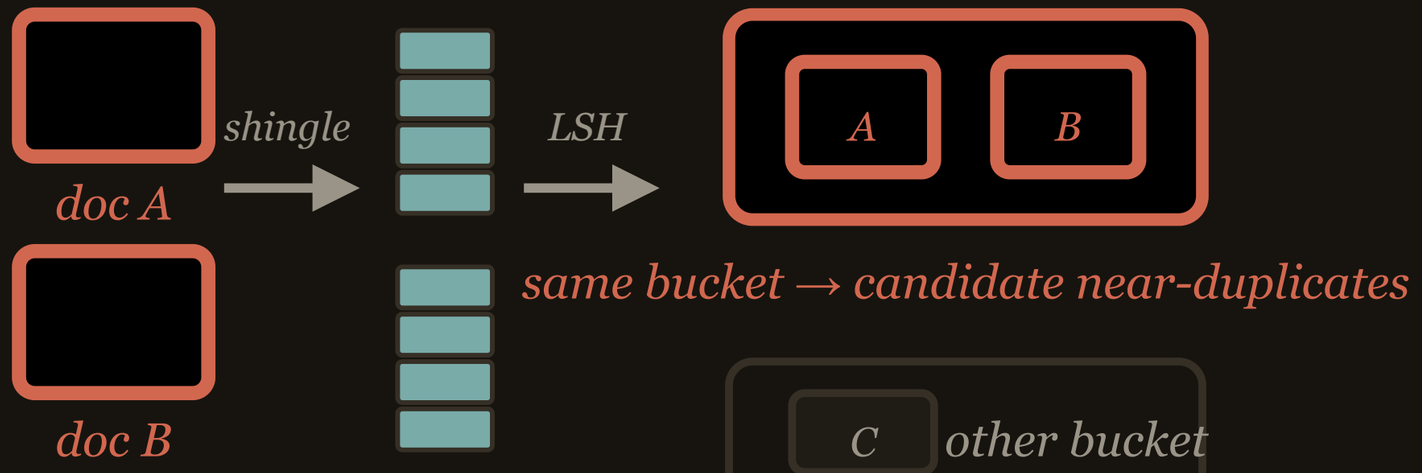


as a link graph — node size = PageRank

Inverted index | Graph mining · PageRank | Graph embeddings | Ranking | Retrieval | Evaluation

GUESS FIRST · ③ AT SCALE

How do you spot near-duplicate pages among billions?



Jaccard distance | Shingling | MinHashing | LSH / hashing

Fast-filter → slow-check

SAME PIPELINE, THREE DIFFERENT JOBS

Data mining vs. machine learning vs. data management



Data management — store & serve the data **Data mining** — exploration & question-formulation (and making sense of the result)

Machine learning — fitting the model

One shared pipeline — data mining frames the question, machine learning fits the model.

RECAP

Where data mining lives

THE DIVISION OF LABOUR

Machine learning owns the model-fitting stage — which algorithm, how to train it, the resulting model.

Data mining owns the exploration and question-formulation stage — deciding what to look for and what "similar", "structure", or "relevant" should mean — and then making sense of the result. Often the data is unlabelled, so the learning here is unsupervised.

A HEALTHCARE EXAMPLE

Mining first, then predicting

DATA MINING PHASE

Analyse a large database of patient records — symptoms, diagnostics, treatments, outcomes — with **clustering**, **association rules**, **pattern discovery** to *uncover patterns* (e.g. fever + cough + shortness of breath → a specific respiratory illness).

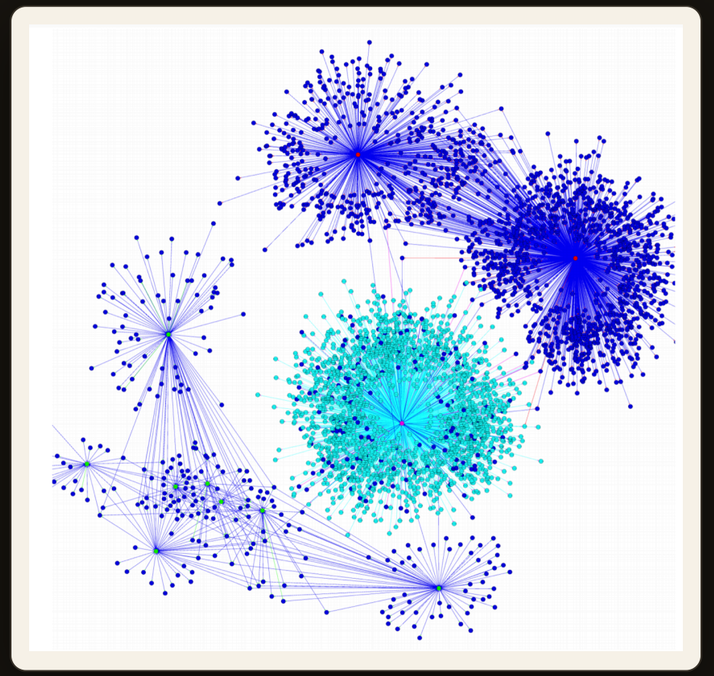
MACHINE LEARNING PHASE

Use those discovered patterns to **train a model**. A new patient's symptoms → the model predicts likely diagnoses and suggests treatments from historical success rates.

Data mining finds the structure; machine learning predicts with it.

CYBER SECURITY · MALWARE FAMILIES

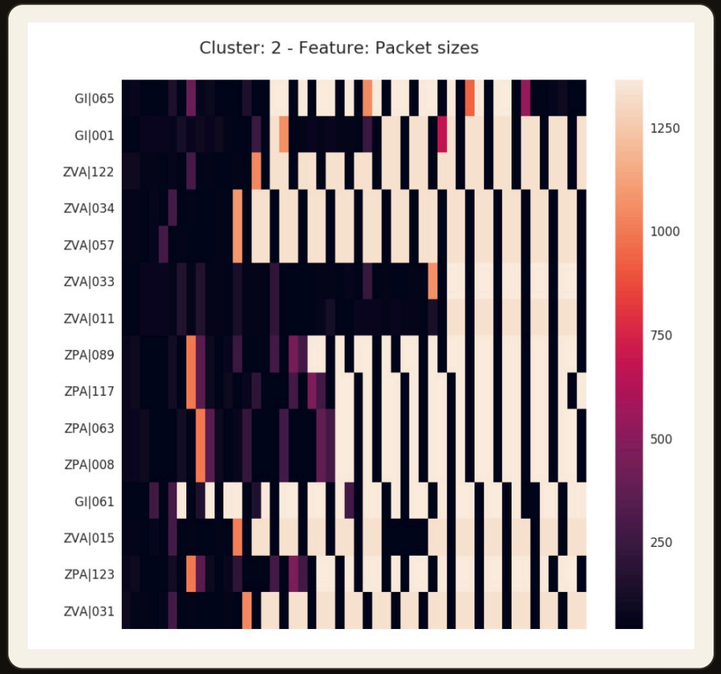
A real example: network / graph data



Connections between hosts form a huge graph. Clustering it reveals behavioural profiles of malware families — no labels needed (Nadeem et al., 2020).

MAKING THE RESULT INSIGHTFUL

Visualise groups with heatmaps



Row = connection · column = time · cell colour = bytes transferred. A cluster becomes a readable behavioural signature.

THE WHOLE LOOP

These are all data-mining skills

THE DATA-MINING LOOP

Take the data as is, but select features using domain knowledge → cluster with clustering methods → make the gained knowledge insightful (e.g. heatmaps).

Feature choice, the right algorithm, and a human-readable result — that end-to-end judgement is the craft.

THE THREE V'S

Why is data mining hard?

Volume

- Too big to fit / scan naively
- Need **scalable** algorithms

Velocity

- Data arrives fast / streams
- Need **approximate yet accurate** estimates

Variety

- Many types & distributions
- Must **adapt** methods to each

These three pressures are exactly why we need hashing, indexing, embeddings and dimensionality reduction.

PART II

The course & logistics

The team, what's new this year, and how everything runs — goals, schedule, labs, grading, prerequisites.

WHO TEACHES THIS COURSE

Three instructors, three verticals

Avishek Anand · head lecturer

- Assoc. prof., Web Information Systems
- IR, NLP, Explainable AI
- Teaches: **distances, embeddings, graphs** — text & graph mining

Julián Urbano

- Asst. prof., Computer Science (TU Delft)
- Music IR, retrieval evaluation, significance testing
- Teaches: **retrieval, rankings, evaluation**

Nergis Tömen

- Asst. prof., Intelligent Systems
- Bio-inspired vision, neuromorphic computing
- Teaches: **matrices, PCA, decompositions, recommenders**

WHAT'S DIFFERENT THIS YEAR

This year: an information-access focus

THE SHIFT

The course is reoriented around information access — how we find, rank, and evaluate the right information inside massive data, not just describe it.

Content swap vs. last year: out go anomaly detection and sketching; in come retrieval, ranking, and evaluation.

— One lab assignment this year — down from two.

— No Kaggle competition this year.

— New reading: Introduction to Information Retrieval (Manning, Raghavan & Schütze) for selected chapters.

WHAT YOU SHOULD WALK AWAY WITH

Course goals — three strands in every lecture

THE PROMISE

Each lecture runs three strands at once — the **concepts**, the **algorithms**, and the **skills** to apply them.

Concepts

- The ideas & theory
- Which concept fits which data type

Algorithms

- Implement them from scratch
- Strengths, weaknesses, and *when* to use them

Skills

- Pick the right method for a problem
- Make it work on real, messy data

The skills strand lives mainly in the lab project — where you apply concepts & algorithms end-to-end.

THE SPINE OF THE WHOLE COURSE

Three verticals: Distances · Matrices · Retrieval

Distances · Avishek

"When are two things similar — and who decides?"

- Similarity & metrics, DTW
- Text & graph embeddings
- Clustering · Hashing · MinHash / LSH

Matrices · Nergis

"What hidden structure explains the data?"

- Representation & operations
- Factorization & decompositions
- Dimensionality reduction (PCA)

Retrieval · Julian

"How do I find the right thing fast — and how do I know I did?"

- Indexing & the inverted index
- Ranking · Evaluation
- Near-duplicate / similar-item search

This year's shift: counting folds into Distances, and Retrieval becomes its own vertical — the information-access focus. Every topic hangs off one of these three.

MONDAYS 13:45 · THURSDAYS 10:45 · EEMCS AMPERE · LABS TUE 08:45

Schedule

Week	Monday 13:45	Thursday 10:45
9 Nov	Intro	Distances
16 Nov	Retrieval	Matrices & PCA
23 Nov	Rankings	Evaluation
30 Nov	Embeddings	Matrix decomposition
7 Dec	no lecture	Recommender systems
14 Dec	Hashing	Sketching
21 Dec–1 Jan	❄ Christmas break — no teaching	
4 Jan	Invited lecture	Indexing
11 Jan	Graphs	Bias & explanations
18 Jan	Exam preparation	no lecture
25 Jan	📝 **Data Mining EXAM · Mon 25 Jan 13:30**	

⚑ The first lab is in Week 5 (Tue 8 Dec) — there are no lab sessions until Week 5.

17 LECTURES · 1 LAB

How the teaching works

- 17 lectures — 13 content, 2 invited, 1 intro, plus Q&A sessions.
- In the room we do case studies — how do you actually apply what you learned to a real problem?
- Complete schedule lives on Brightspace.

WHERE THE CONTENT COMES FROM

Course material

ON BRIGHTSPACE (REQUIRED)

Lecture slides (after each class), lab exercises (start of each week), and reading materials — selected book chapters and papers.

— Free via the TU Delft digital library: *Mining of Massive Datasets* · *Data Mining*.

— New this year — *Introduction to Information Retrieval* (Manning, Raghavan & Schütze) for the retrieval / ranking / evaluation chapters. Free online: nlp.stanford.edu/IR-book

— Only the selected chapters uploaded to Brightspace are examinable.

MANDATORY · IN GROUPS

Lab sessions

- One lab assignment this year, in groups — Tuesday mornings (08:45), help and feedback on the spot.
- Two parts: algorithm implementation (Distances / Matrices / Retrieval) and a pipeline (transform → analyse → visualise), built from scratch — no scikit-learn. No Kaggle this year.
- Each group submits one solution, then peer-reviews 2 other groups' assignments on Peer.
- Register your group on Brightspace this week — little flexibility: groups are exported to Peer soon so grading stays on time.
- Missed or failed? An "Assignment repair" resit is due 2 weeks after the original deadline.
- Help: Queue, Mattermost, Answers EWI (tag CSE2525), Weblab, Peer. Please do not use e-mail.

LAB 3 × 10% · EXAM 70%

Assessment & grading

HOW THE GRADE IS MADE UP

Lab = 30%, split into three 10% parts — algorithm implementation (auto-graded), the pipeline (peer-reviewed), and your peer reviews. Exam = 70%. No Kaggle this year.

To pass (TER Art 29, sub 4): each graded element must be > `5.0`, *and* the final computed grade (exam + lab) must be > `5.8`. Check the TER.

THE ASSESSMENT

Final exam

- Final exam = 70%, on WebLab (Osiris + WebLab registration). Open and multiple-choice questions.
- No programming questions this year · closed book, calculator allowed · one resit.
- A practice exam will be available — one only this year, due to staff & topic changes.
- No partial transfer from previous years. mytimetable.tudelft.nl is authoritative for dates.

WHAT WE ASSUME YOU ALREADY KNOW

Expected prior knowledge

— Discrete maths — sets, intersections, unions.

— Linear algebra — matrix multiplication, projections, eigenanalysis.

— Probability & statistics — Gaussians, correlation, covariance.

— Graph theory — adjacency matrix, degree, clique, bipartite graph, shortest path.

— Data structures — hash tables and indexes. Programming — Python.

— Machine learning — logistic regression, random forest, SVM, …

Based on CSE1100/1305/1200/1205/1210 and CSE2510/2520. Not enforced — you are responsible for your own study success.

RECAP OF PART I

Logistics in one slide

KEY POINTS

One mandatory lab — 3 × 10% (algorithm + pipeline + peer reviews); exam 70%. Pass needs > 5.0 on each element and > 5.8 overall (TER Art 29.4). Register your lab group this week.

This year’s focus: information access — retrieval, ranking and evaluation replace anomaly detection and sketching. Feedback only: dm-cs-ewi@tudelft.nl.

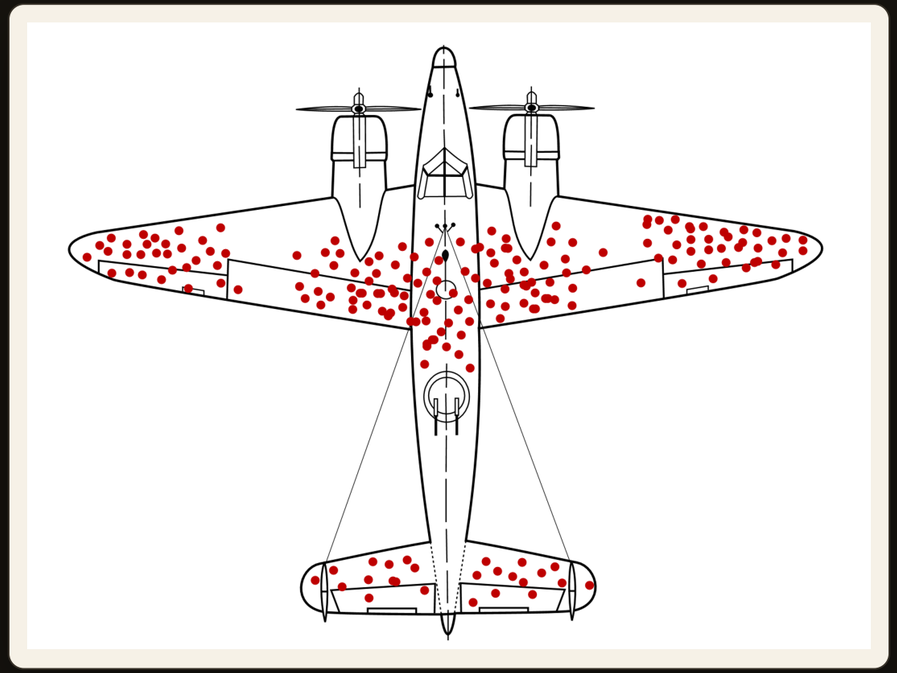
PART III

A word of caution

Don’t fool yourself. The rest of this lecture is cautionary tales about results that look great and are wrong.

WWII · ABRAHAM WALD

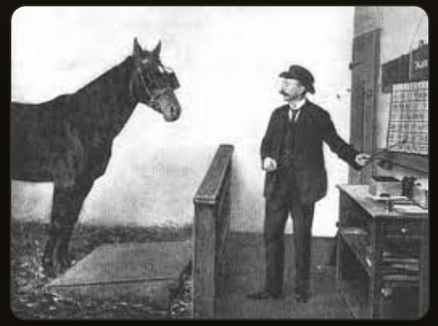
Survivorship bias



Red dots = damage on planes that returned. Where should you add armour? Not where the dots are — where they aren’t (engines, cockpit): planes hit there never came back. Your data is only the survivors.

THE HORSE THAT COULD “COUNT”

The Clever Hans effect



Hans seemed to do arithmetic — but was really reading his trainer’s unconscious cues. Models do this too: right answer, wrong reason.

TOO GOOD TO BE TRUE

Shortcuts in learning — Wikipedia citations

THE TASK

Automatically predict which Wikipedia sentences **need a citation**. A neural network scored **100% on the test set** — while humans agreed only **<55%** of the time.

WHAT SHOULD YOU ASK?

100% accuracy on a task humans find genuinely hard is too good to be true — a red flag, not a trophy.

The catch: sentences needing a citation already carried the marker "[ ]". The model learned *"if the bracket is there, predict citation-needed"* — a shortcut, not understanding.

THE CANONICAL ML PARABLE

Shortcut learning — the “Russian tanks”

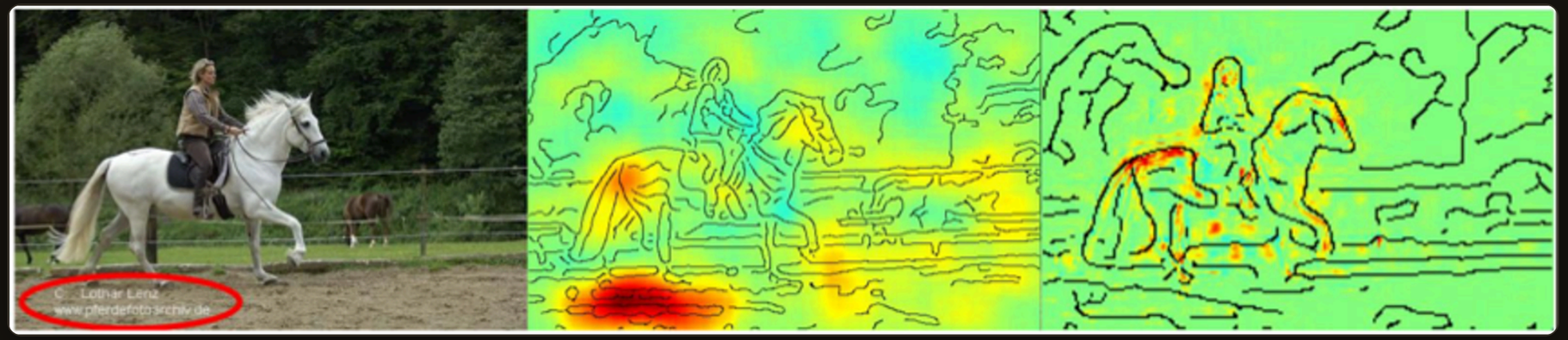




100% accurate in training, worse than random on new data. The network had learned: “if the sky is blue, no tank; otherwise, tank.” It classified the weather, not the tanks.

EXPLAINABLE AI REVEALS THE TRICK

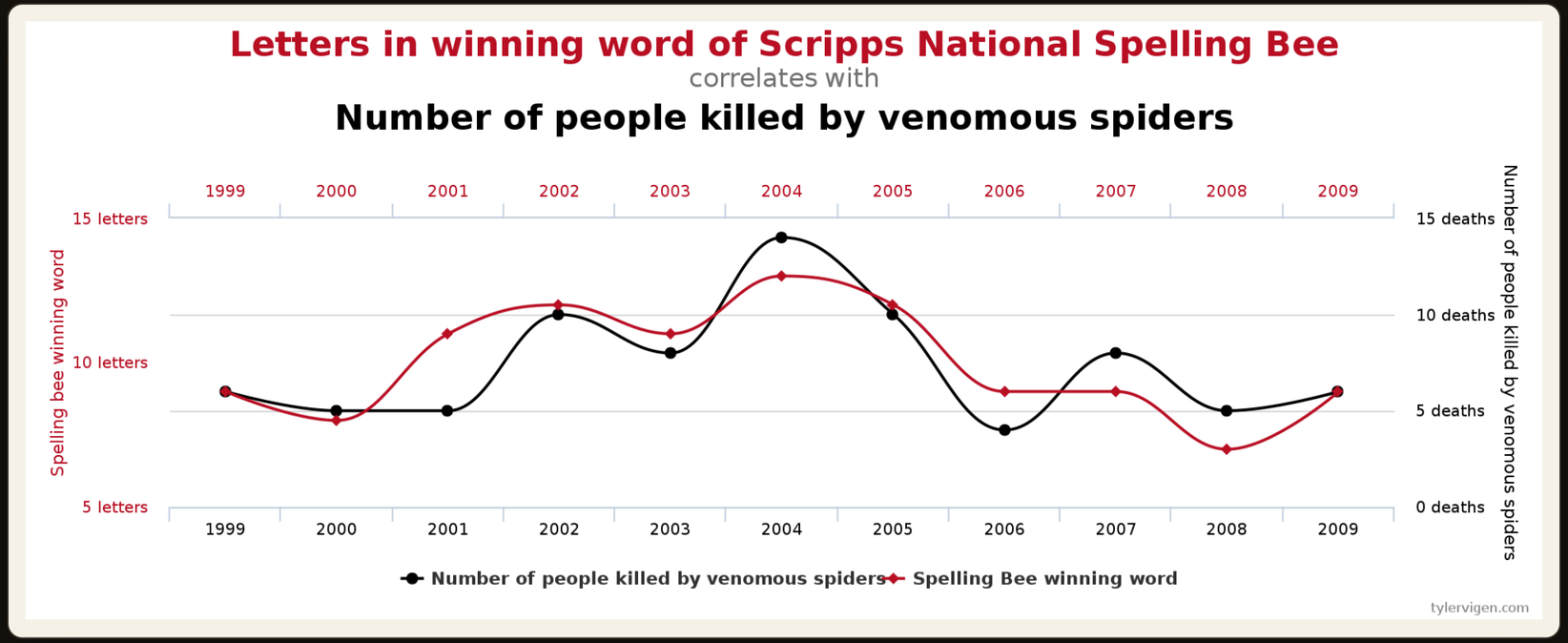
Shortcuts in image models



A classifier labels this “horse” with high confidence. Saliency maps show *why*: it fires on the copyright watermark in the corner. If “© …” → predict horse.

MULTIPLE COMPARISONS

You can always find what you're looking for



Test enough hypotheses and some will "correlate" by pure chance. The multiple-comparisons problem is the cause of many data-mining errors — significance you went fishing for isn't significance.

LECTURE 1 · WRAP-UP

Summary

TAKE AWAY

Data mining = extract useful information from a large volume of data and turn it into an understandable structure. It focuses on the input and output ends of the ML pipeline, often unsupervised.

Organised around three verticals — Distances · Matrices · Retrieval. Grading: labs 30%, exam 70%.

It's called the skill of the 21st century — but tread with caution. Survivorship bias, Clever Hans, shortcut learning and multiple comparisons are waiting to fool you.