Big Data and Data Mining

Week 1: Introduction

Fenerbahce University

Instructors

Assist. Prof. Vecdi Emre Levent

Office: 311

Email : emre.levent@fbu.edu.tr

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting
    - Non-trivial
    - Implicit
    - Previously unknown and potentially useful patterns or knowledge

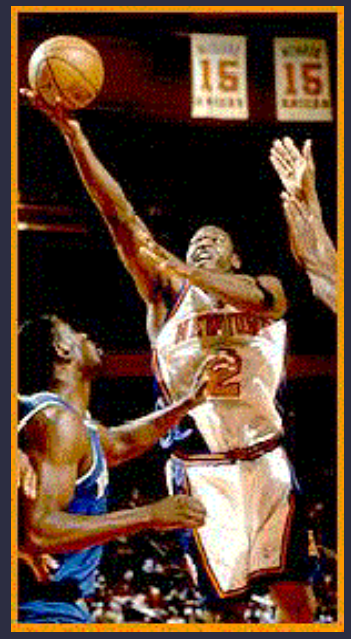
  from huge amount of data

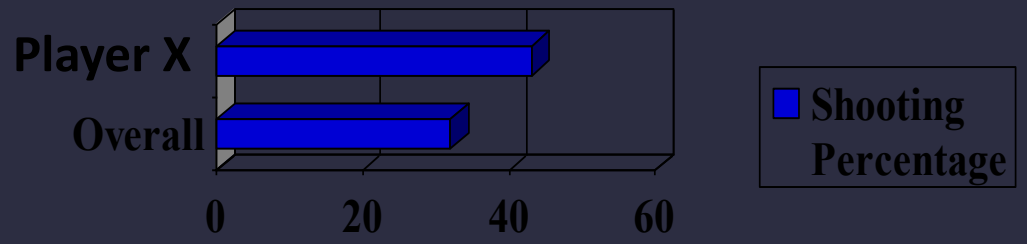
What Is Data Mining?

- Alternative names
  - Knowledge discovery (mining) in databases (KDD)
  - Knowledge extraction
  - Data/pattern analysis
  - Data archeology
  - Data dredging
  - Information harvesting
  - etc…

What is Data Mining?

- Play-by-play information recorded by teams
  - Who is on the court
  - Who shoots
  - Results
- Coaches want to know what works best
  - Plays that work well against a given team
  - Good/bad player
- A data mining tool to answer these questions





Why Data Mining?—Potential Applications

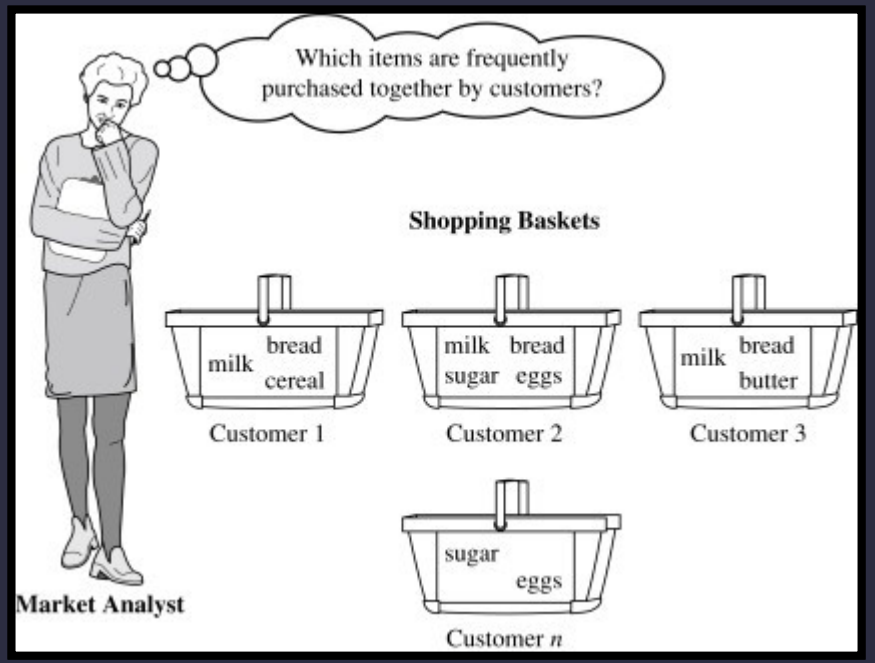
- Data analysis and decision support
  - Market analysis and management
    - Target marketing
      - Segmentation
      - Targeting
      - Positioning
    - Segmentation bases:
      - Demographic,
      - Geographic
      - Psychographic
      - Behavioral

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Customer relationship management (CRM)
      - Unified customer data (contacts, interactions, deals)
      - Pipeline & workflow management (stages, tasks, SLAs)
      - Automation (lead capture, scoring, routing)

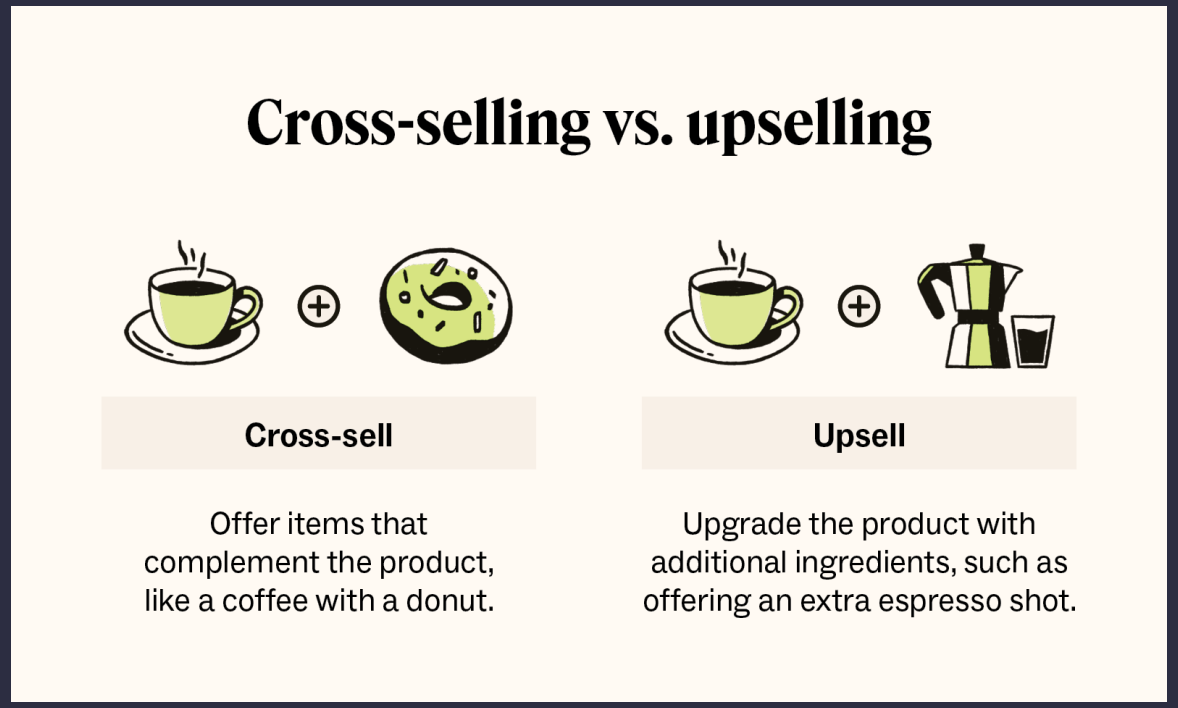
Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Market basket analysis
      - Cross-sell
      - Promotions
      - Shelf/layout optimization



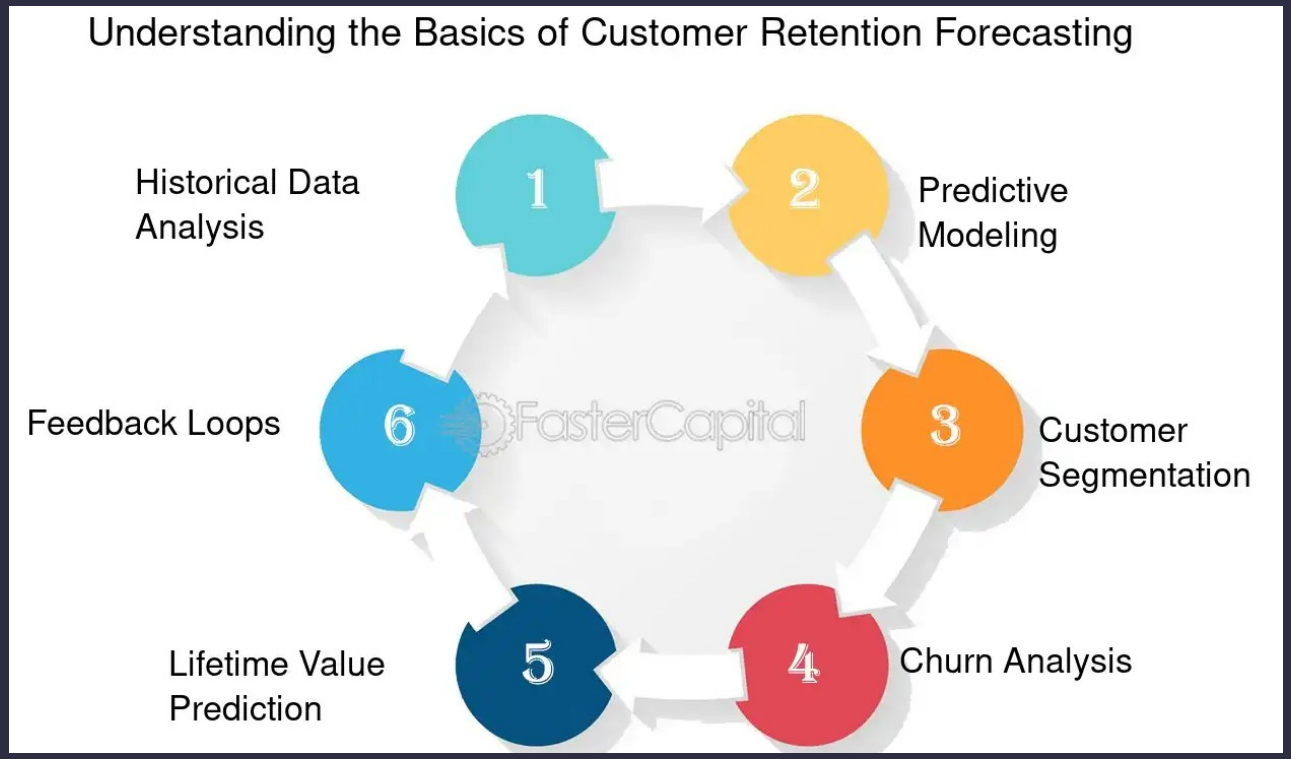
Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Cross selling / Up Selling



Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Risk analysis and management
    - Forecasting
    - Customer retention

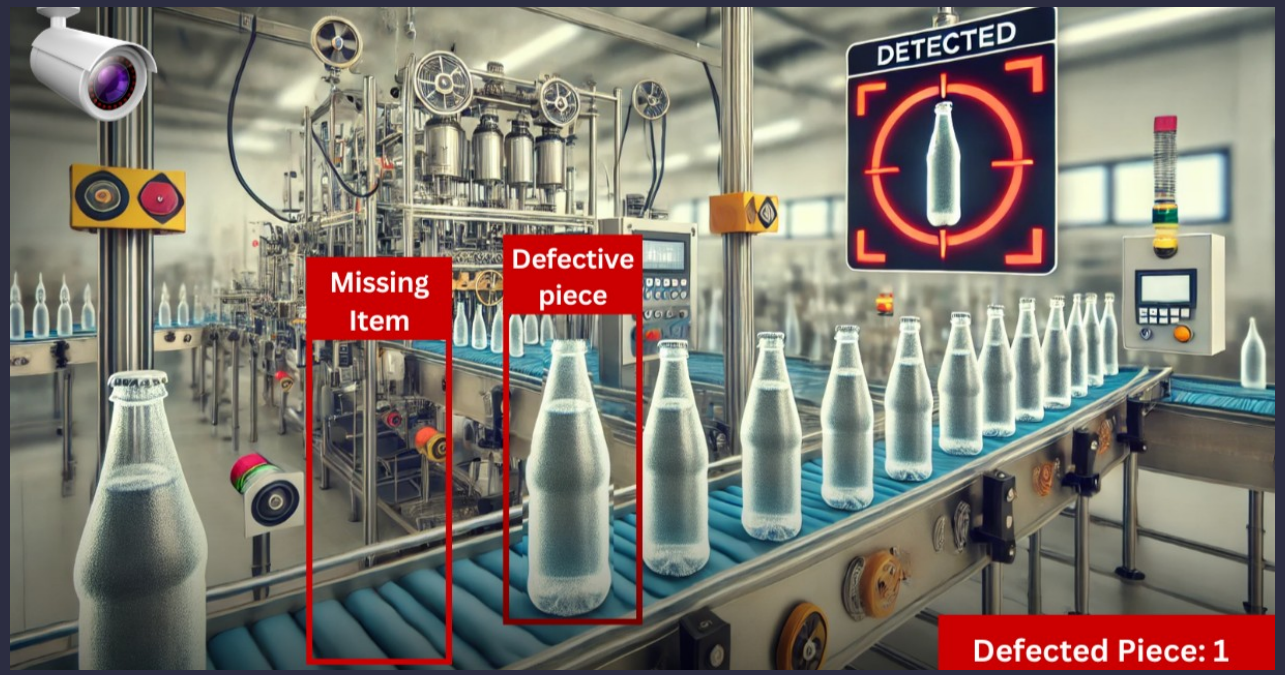


Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Fraud detection
  - Detection of unusual patterns (outliers)
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

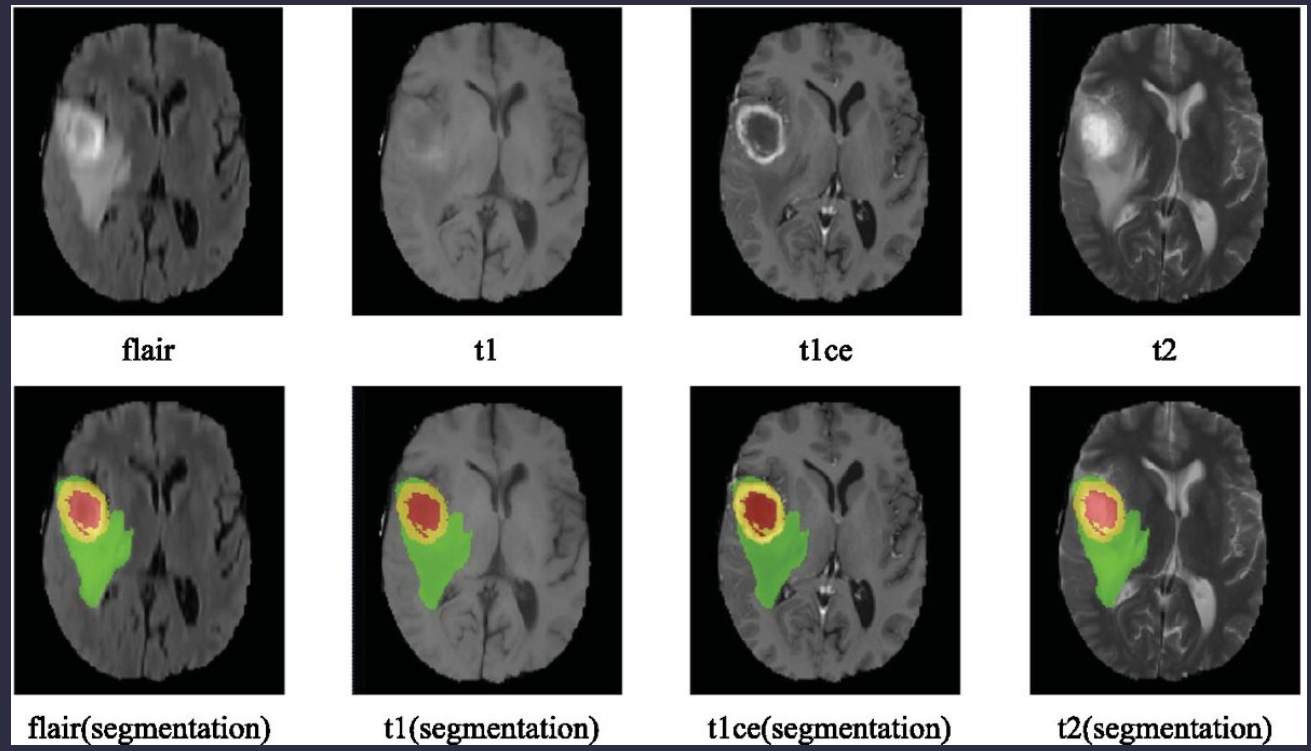
Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Product defect
    - Product defect detection (Production line images)



Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Medical imaging (lesion detection, segmentation)



Why Data Mining?—Potential Applications

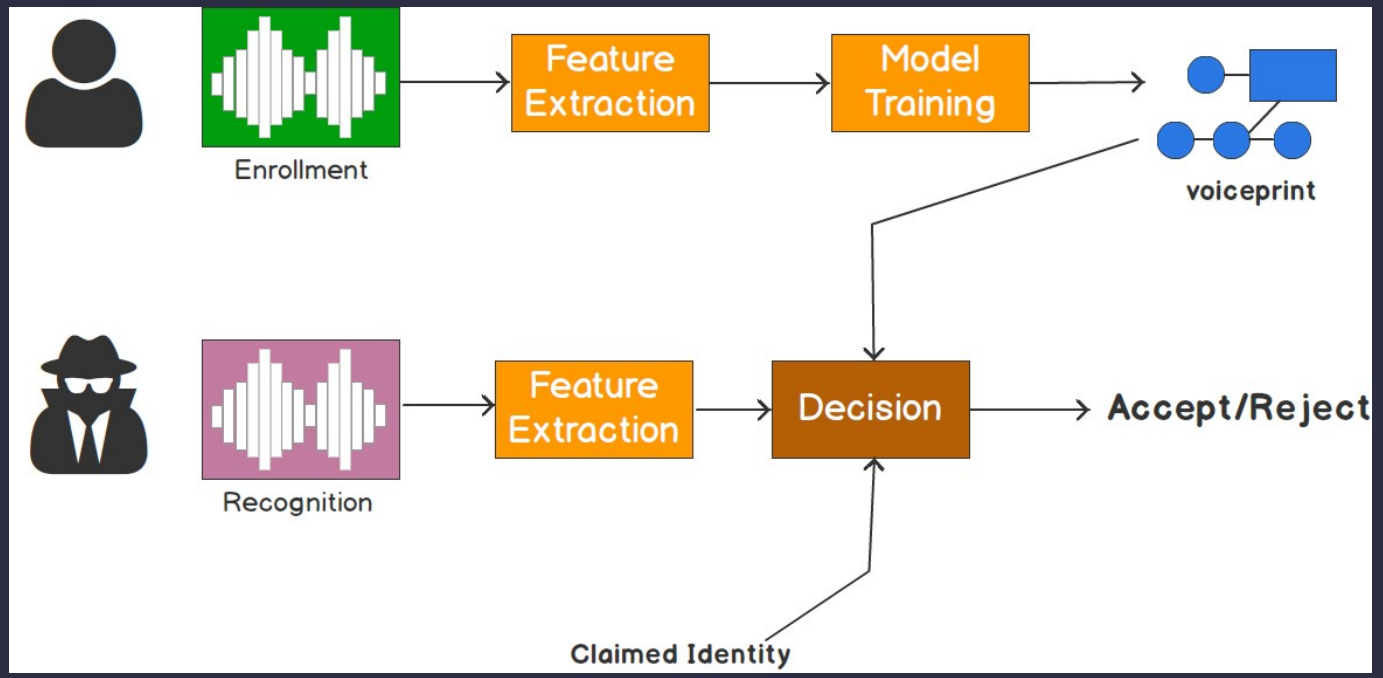
- Data analysis and decision support
  - Store shelf analysis (planogram compliance, stock clearance)

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Call center recordings (sentiment, keyword capture)

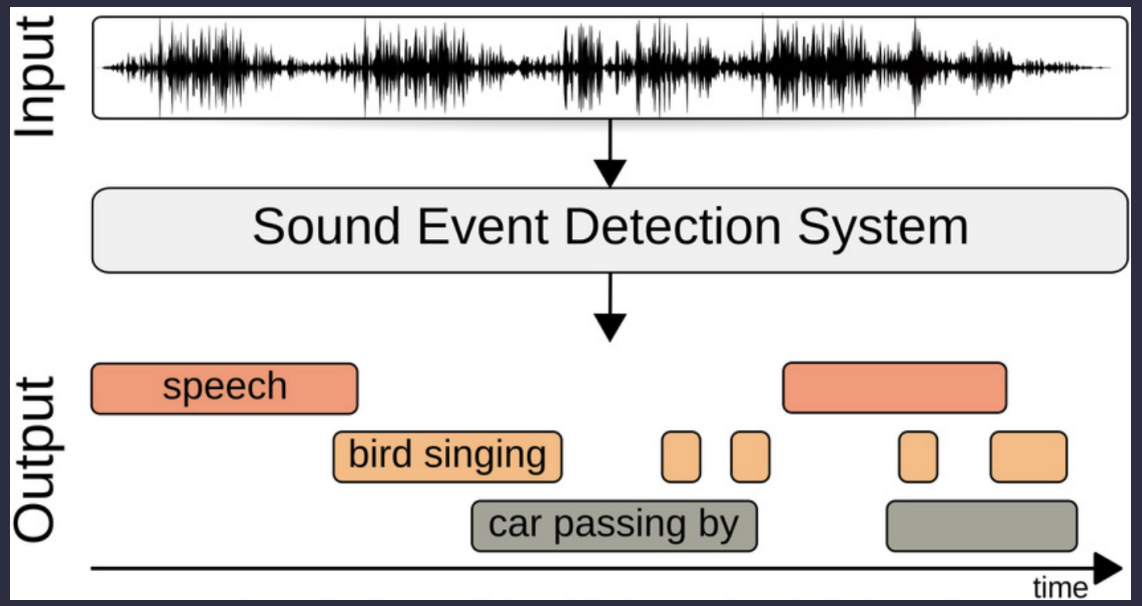
Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Speaker discrimination & verification (diarization, speaker ID)



Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Noise event detection (alarm, glass breakage, machine malfunction)

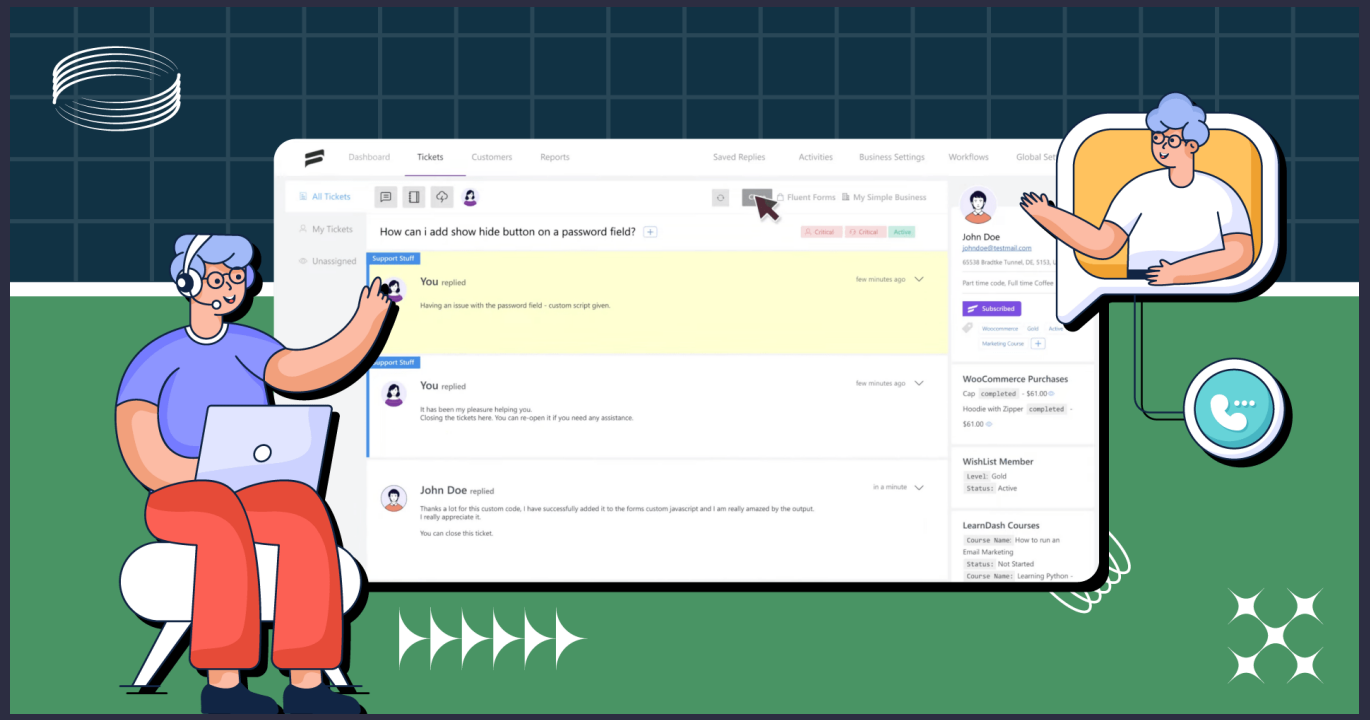


Why Data Mining?—Potential Applications

- Text Based
  - Social media comments (sentiment, topic modeling)

Why Data Mining?—Potential Applications

- Text Based
  - Customer complaints & support tickets (intent, prioritization)

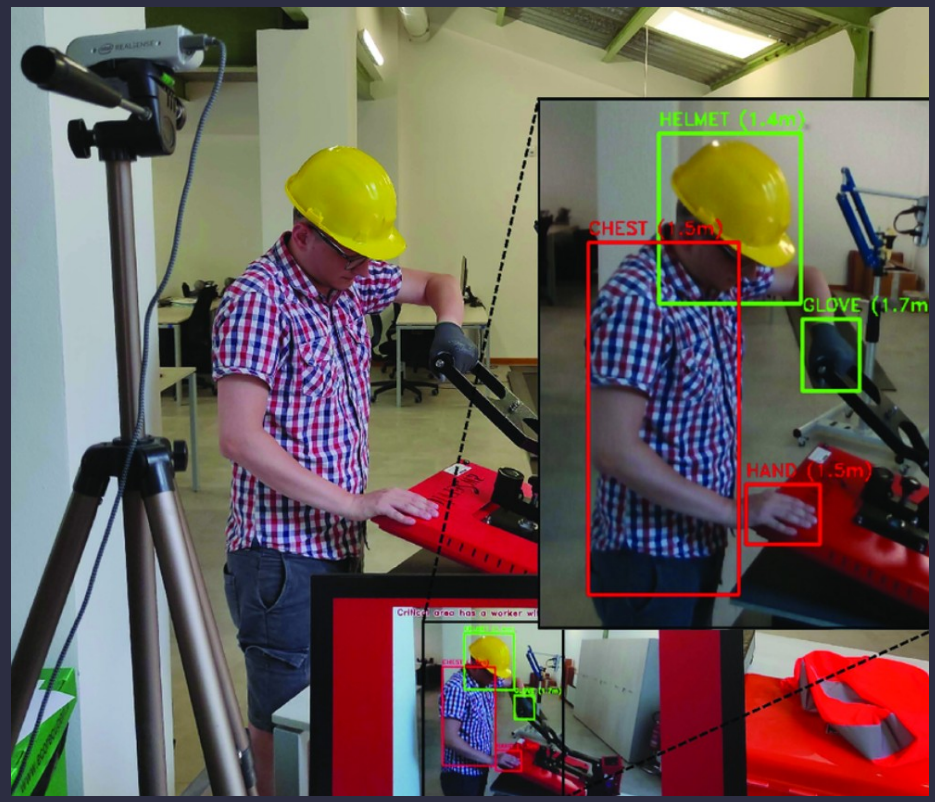


Why Data Mining?—Potential Applications

- Text Based
  - News-academic text summary and information extraction

Why Data Mining?—Potential Applications

- Video Based
  - Security and occupational safety (Dangerous behavior, PPE/PPE monitoring)

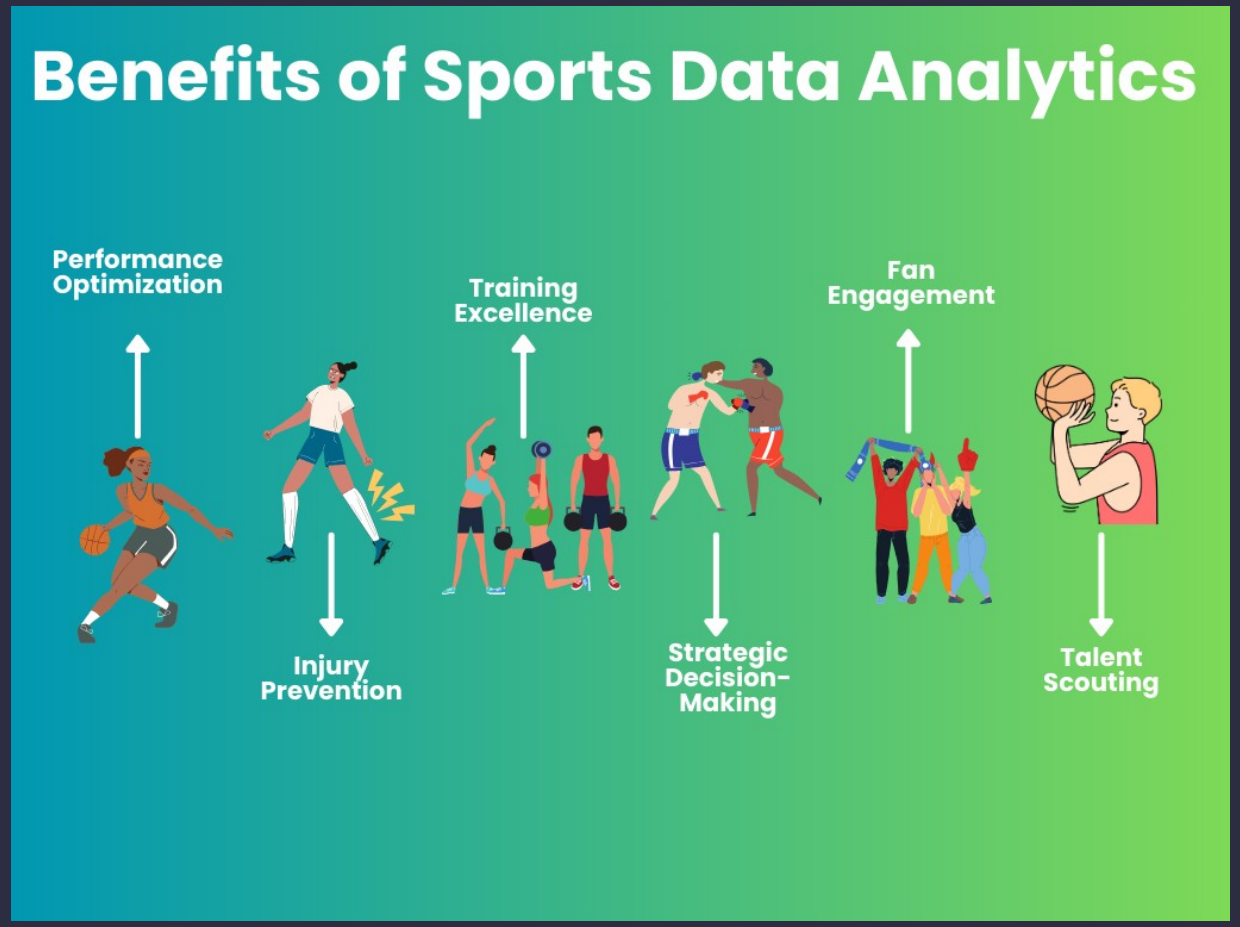


Why Data Mining?—Potential Applications

- Video Based
  - Retail customer flow (count, heat map)

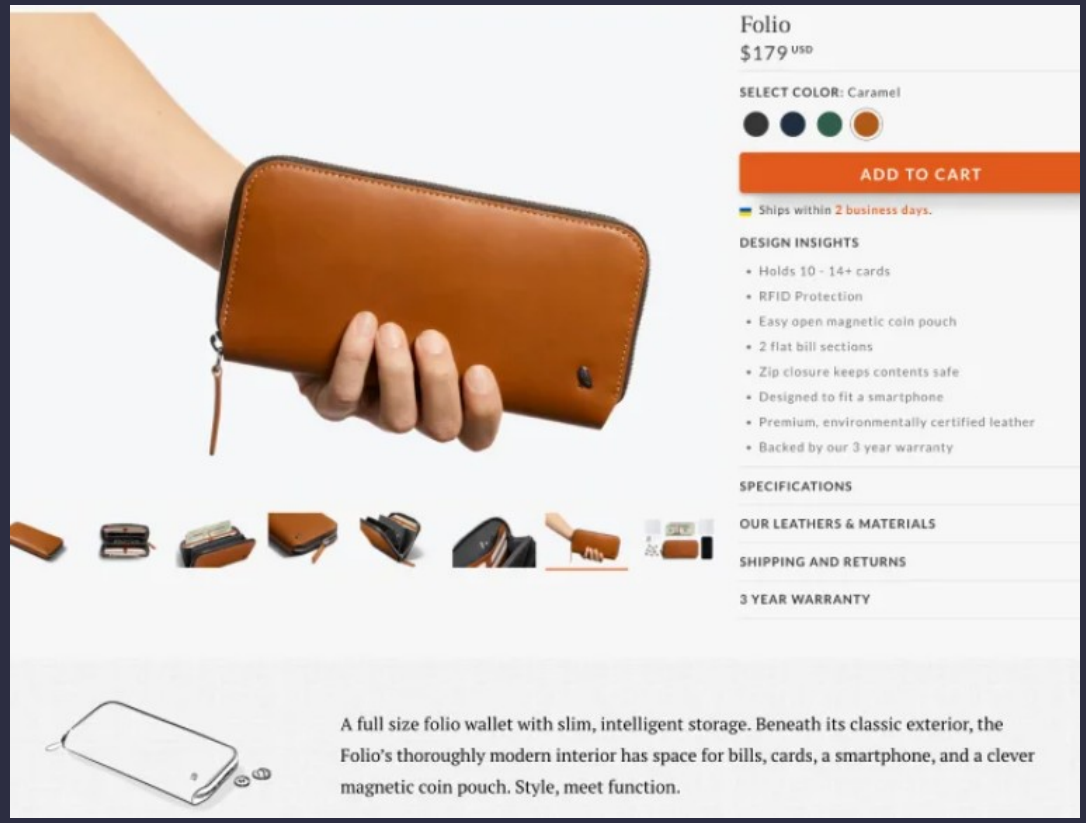
Why Data Mining?—Potential Applications

- Video Based
  - Sports analytics (in-game event detection, performance metrics)



Why Data Mining?—Potential Applications

- Multi-Modal Based
  - E-commerce product pages (visual quality + description consistency)

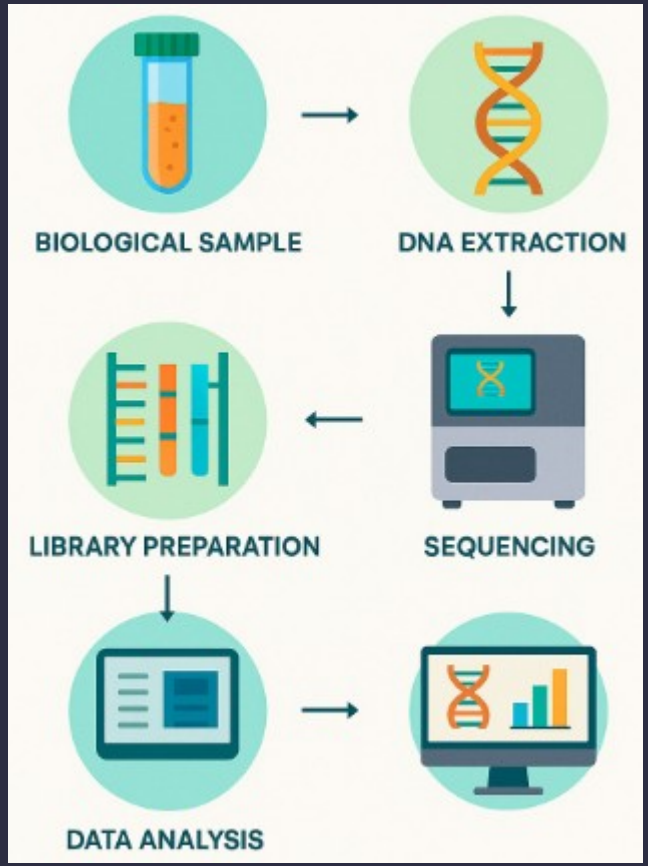


Why Data Mining?—Potential Applications

- Multi-Modal Based
  - Ad effectiveness (video scene + subtitle + tone of voice)

Why Data Mining?—Potential Applications

- Other Applications
  - DNA and bio-data analysis



Course Resources

Website: levent.tc

Courses > Graduate Courses > Big Data and Data Mining

Course Resources

Course Page Content;

- Syllabus
- Lesson Schedule
- Lecture Notes
- Homeworks
- Projects
- Exams

Course Resources

Syllabus;

Lesson hours;

Monday 9.00-14.00

Course Resources

Syllabus;

Between 4-6 homeworks will be given.

2 Quizzes .

Attendance to classes is mandatory at **80** %.

Course Resources

Syllabus;

Evaluation weights

Activities	Rates
Visa	20%
Homework/Quiz	10%
Project	30%
Final	25%

Course Resources

Syllabus;

Letter grade ranges

Term Grade	Weight	Letter grade
90-100	4.00	AA
85-89	3.50	BA
80-84	3.00	BB
75-79	2.50	CB
65-74	2.00	CC
50-64	1:50	DC
45-49	1.00	DD
0 -44	0	FF

Course Resources

Syllabus;

expected effort

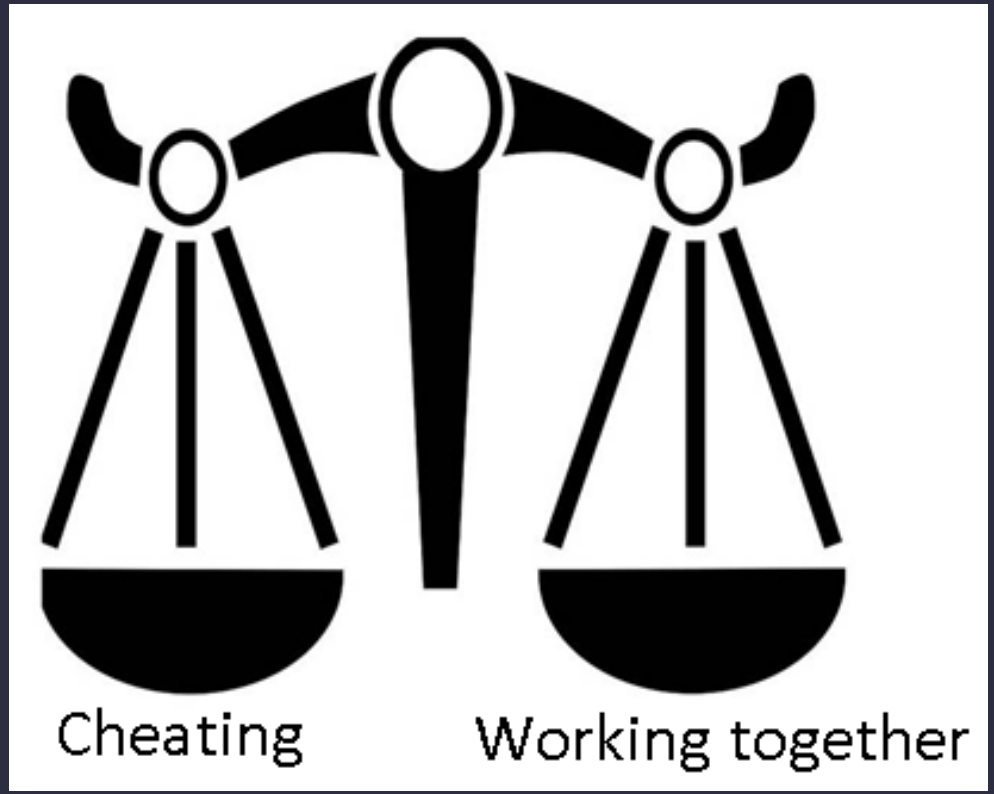
190 hours in total effort is expected.

Contents	Hour	How many times	Subtotal
Course Preparation	2	14	28
Course Repetition	2	14	28
Homework	4	6	24
Project	48	1	48
Classroom Lesson	4	14	56
Midterm and Final	3	2	6

Course Resources

Syllabus;

Academic honesty



Outline

- Introduction: What is data mining?
- Data mining tasks - Clustering, Classification, Rule learning, etc.
- Data mining process: Data preparation/cleansing, task identification
- Introduction to WEKA
- Association Rule mining
- Association rules - different algorithm types
- Classification/Prediction
- Classification - tree-based approaches
- Classification - Neural Networks
- Clustering basics
- Clustering - Statistical approaches
- Clustering - Neuralnet approaches
- Image Classification & Object Detection
- Sound Processing
- Text Mining

Outline

Banking – Credit Card Fraud Detection

- Amount & Purchase Patterns
  - Amount-related derivatives: Z-score, amount-to-pay ratio, number of installments based on the cardholder's average over the last 7/30 days.
  - Micro-amount (e.g., $0–$5) or very large amount flag, outlier indicator.
  - Same-day refund/chargeback rate, number of past chargebacks.
  - Labels Fraud or Not

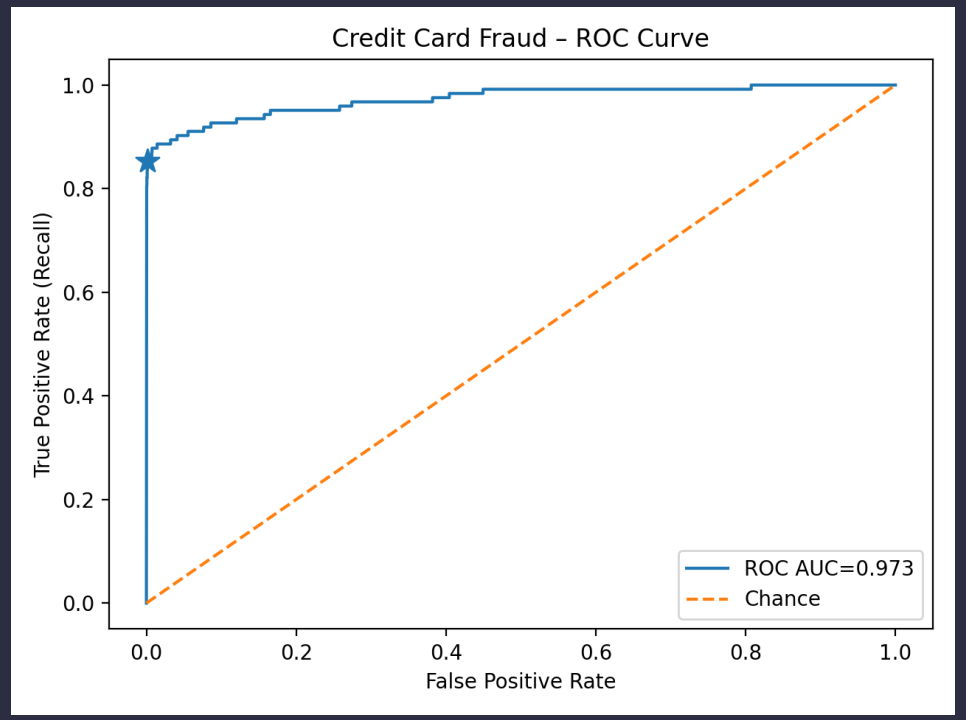
Sample-Bank Transactions

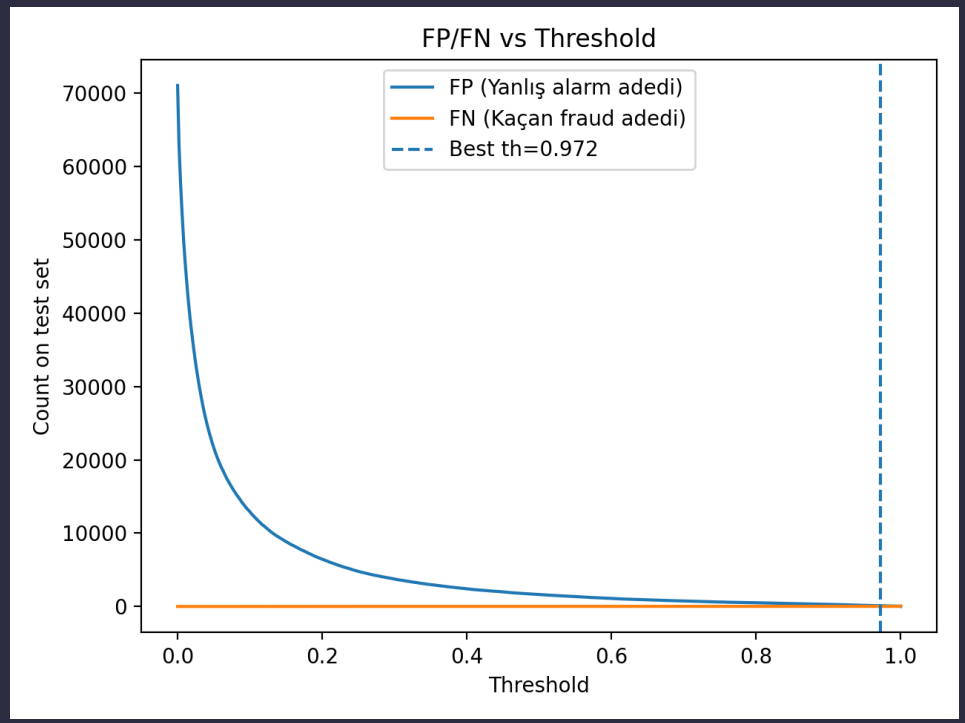
	TRANS_ID	TYPE	DATE	AMOUNT
1	1348	CHEQUE	02-01-2008	-1.669,92
2	1444	CHEQUE	02-01-2008	-11.546,89
3	1407	CHEQUE	04-01-2008	-5.499,39
4	1520	CHEQUE	04-01-2008	-3.101,20
5	1586	CHEQUE	05-01-2008	-10.466,84
6	1466	CHEQUE	06-01-2008	-8.599,08
7	1575	CHEQUE	06-01-2008	-1.600,03
8	1513	CHEQUE	09-01-2008	-2.129,43
9	1505	CHEQUE	10-01-2008	-11.359,36
10	1393	CHEQUE	11-01-2008	-4.013,81
11	1534	CHEQUE	11-01-2008	-3.525,21
12	1305	CHEQUE	12-01-2008	-1.421,15
13	1392	CHEQUE	12-01-2008	-6.829,53
14	1566	CHEQUE	12-01-2008	-2.187,77
15	1	DEPOSIT	13-01-2008	3.474,20
16	1606	CHEQUE	13-01-2008	-5.488,02

Outline

Banking – Credit Card Fraud Detection

- Fraud Classifier
  - Regression Apporach





FPR: Normal Operations flagged as Fraud

Outline

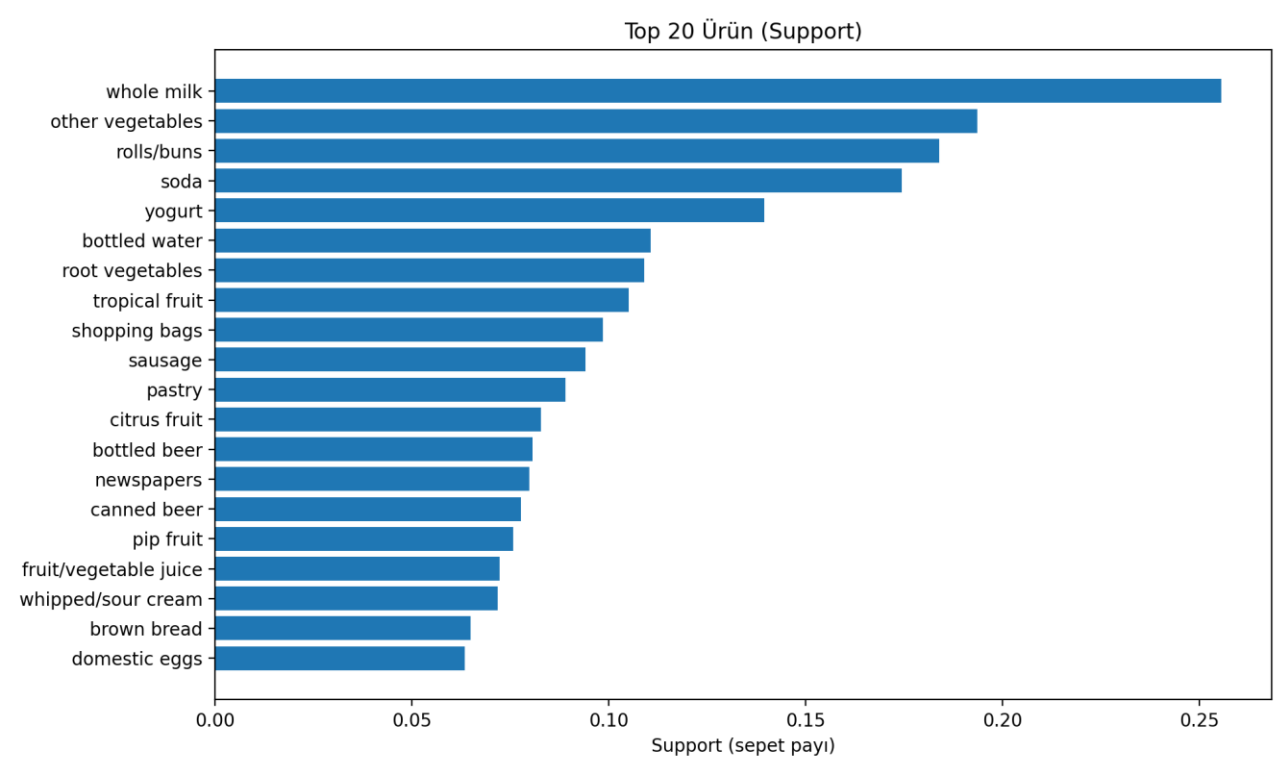
Retail –Basket Analysis

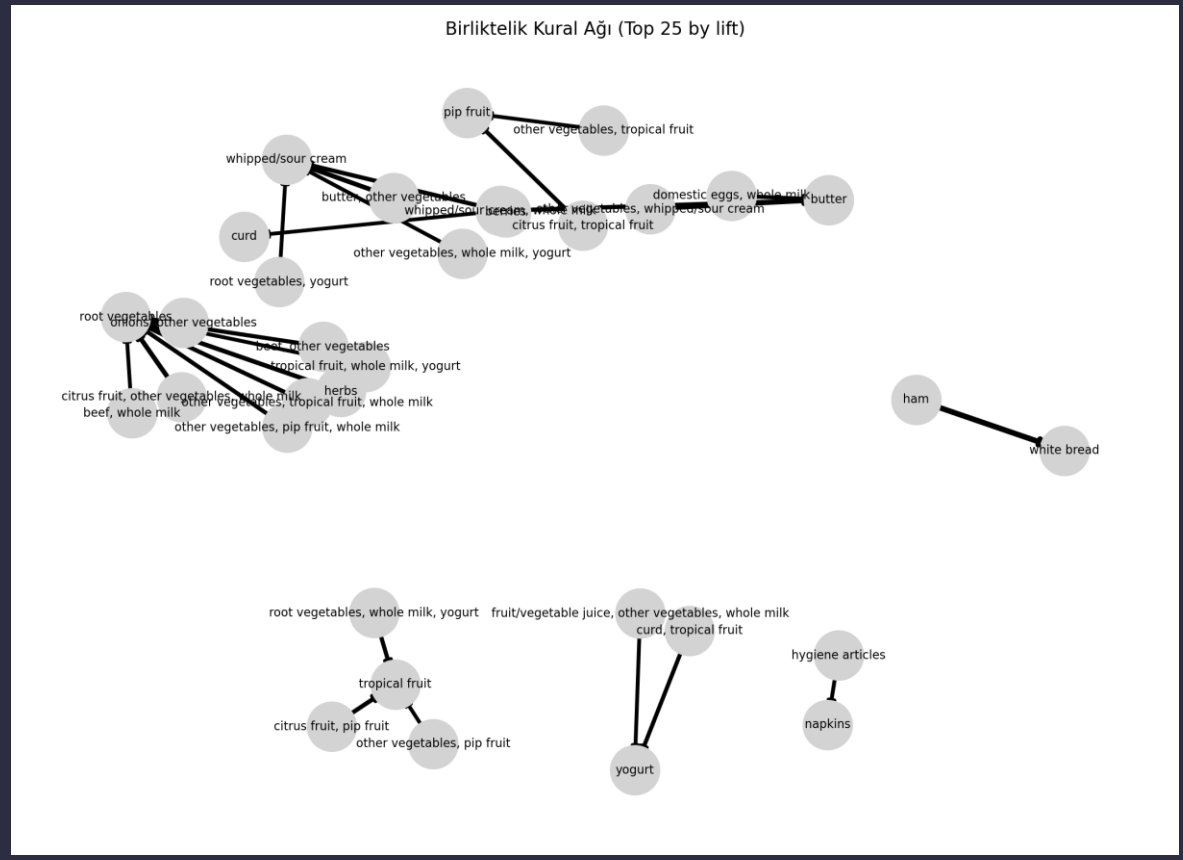
- Dataset: “Groceries” market-basket data
- Format: One row per basket; items listed as text tokens (no customer IDs, prices, or timestamps).
- Task fit: Frequent itemset mining & association rules.
- Use cases: Cross-sell suggestions, shelf layout, promo bundling; good for quick retail demos.

TID	Items in the Basket
1	espresso, sugar, newspaper
2	espresso, sugar, cola
3	espresso, sugar
4	cappuccino, cigarettes
5	cappuccino, sugar
6	cappuccino, sugar, sweets
7	decaf, sugar, chewing_gums
8	decaf, soda, vinegar
9	decaf, sugar, cigarettes

Outline

Retail –Basket Analysis





Outline

Health – Heart Disease Risk Prediction

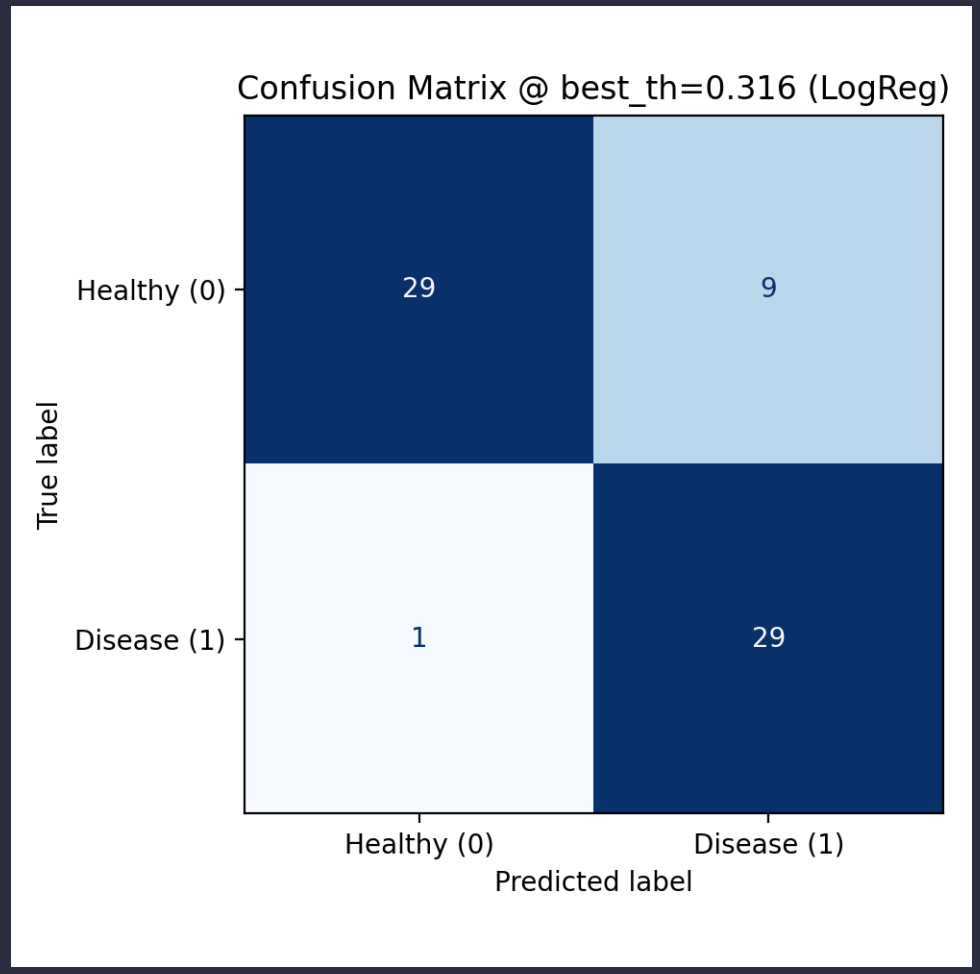
- Dataset: OpenML Heart tabular clinical data (e.g., age, sex, chest pain type, blood pressure, cholesterol, max HR, ST depression).

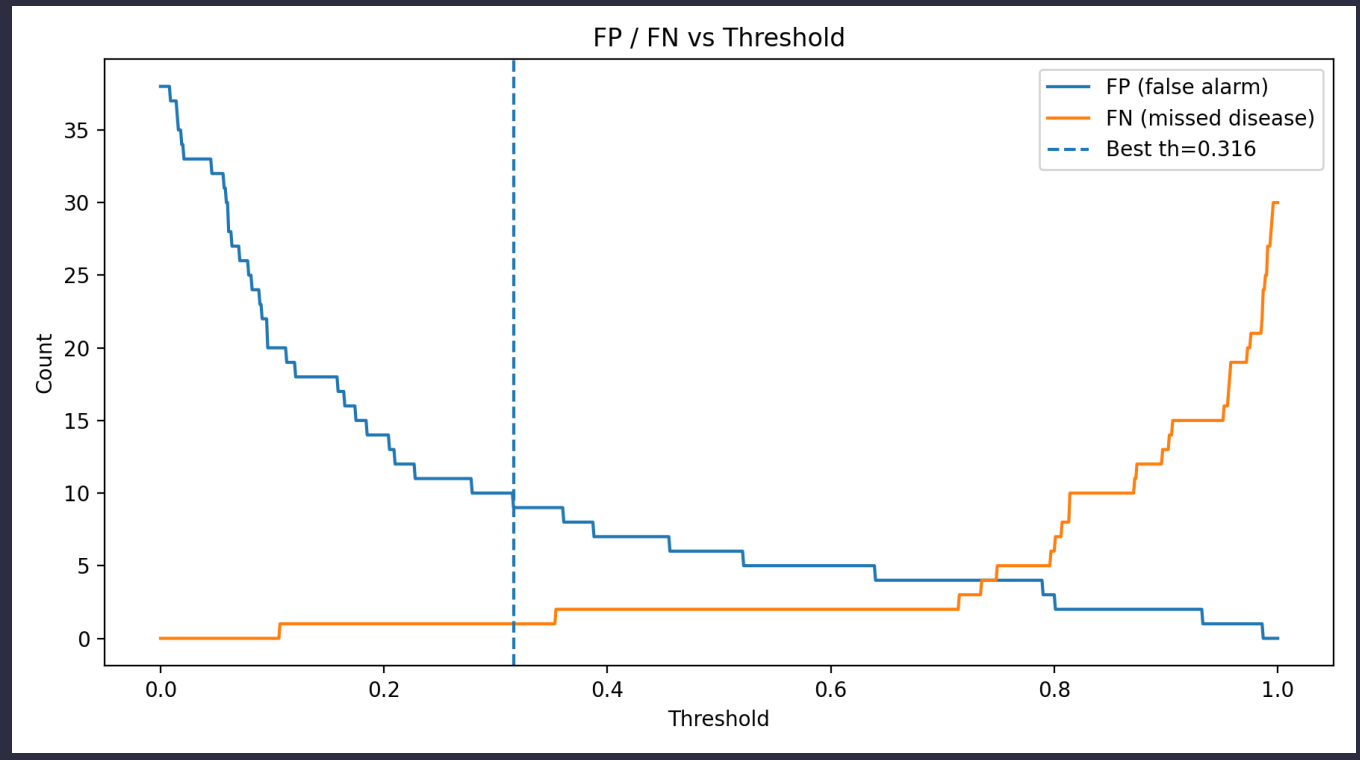
- Target (label): Heart disease presence (binary) — encoded as 1 = disease, 0 = no disease (original strings like “present/absent” are mapped).
- Task: Binary risk prediction / classification; outputs a probability of disease per patient.

Attribute	Type	Description
Age	Continuous	Age of the patient in days
Gender	Discrete	1: women, 2: men
Height (cm)	Continuous	Height of the patient in cm
Weight (kg)	Continuous	Weight of the patient in kg
Ap_hi	Continuous	Systolic blood pressure
Ap_lo	Continuous	Diastolic blood pressure
Cholesterol	Discrete	1: normal, 2: above normal, 3: well above normal
Gluc	Discrete	1: normal, 2: above normal, 3: well above normal
Smoke	Discrete	whether patient smokes or not
Alco	Discrete	Alcohol intake-Binary feature
Active	Discrete	Physical activity-Binary feature
Cardio	Discrete	Presence or absence of cardiovascular disease

Outline

Health – Heart Disease Risk Prediction





Outline

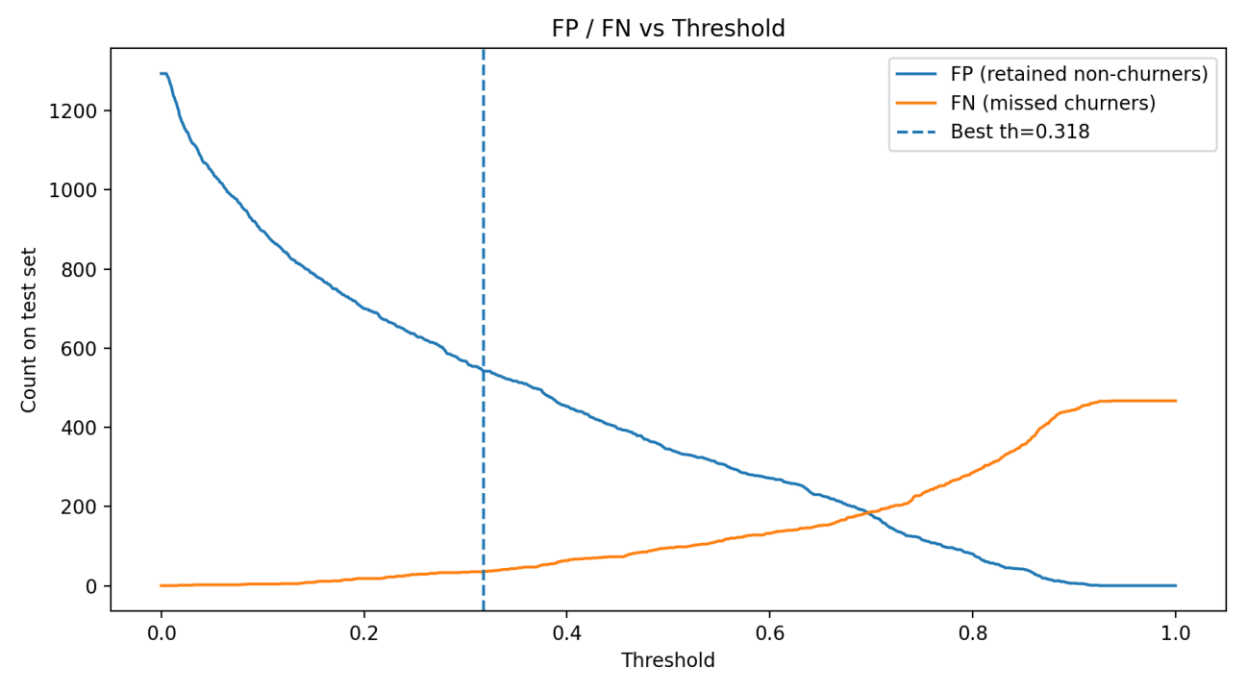
- Telecom – Customer Churn Prediction

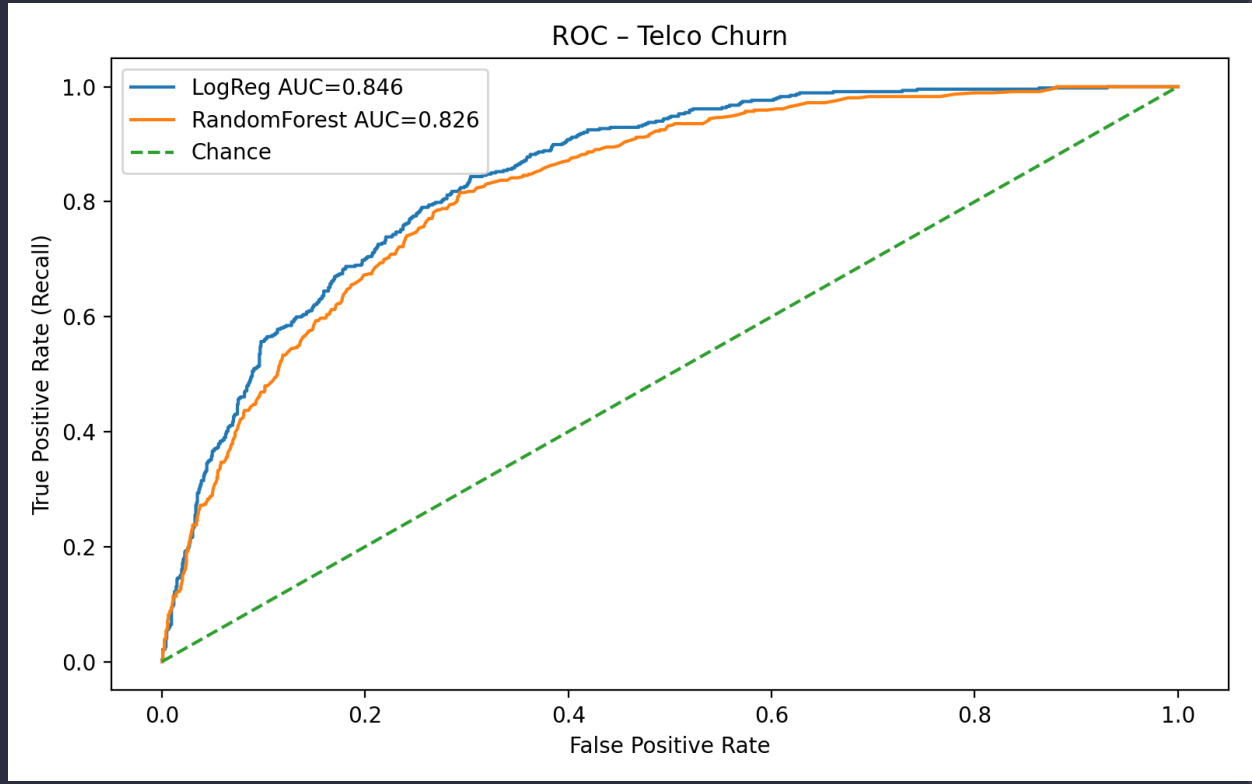
- Dataset: IBM Telco Customer Churn (tabular; ~7k customers, 20+ features).
- Goal: Predict Churn (Yes/No) → probability per customer.
- Models: Logistic Regression & Random Forest (scaling + one-hot).

customer	month	product	tenure	tickets_30d	late_on_bill	churned
78658	Jan	phone	9	0	0	0
93343	Jan	internet & phone	4	2	0	0
54241	Jan	video & internet	15	0	0	1
76227	Jan	internet	9	0	0	0
77751	Jan	video & internet	9	0	1	1
5337	Jan	video & internet	47	0	0	0
37661	Jan	internet	21	1	0	0
55129	Jan	video & internet	14	0	0	0
78658	Feb	phone	10	0	0	0
93343	Feb	internet & phone	5	1	0	0
76227	Feb	internet	10	0	0	0
5337	Feb	video & internet	48	0	0	0
37661	Feb	internet	22	0	0	0
55129	Feb	video & internet	15	0	0	1
57496	Feb	phone	14	1	0	0

Outline

- Telecom – Customer Churn Prediction

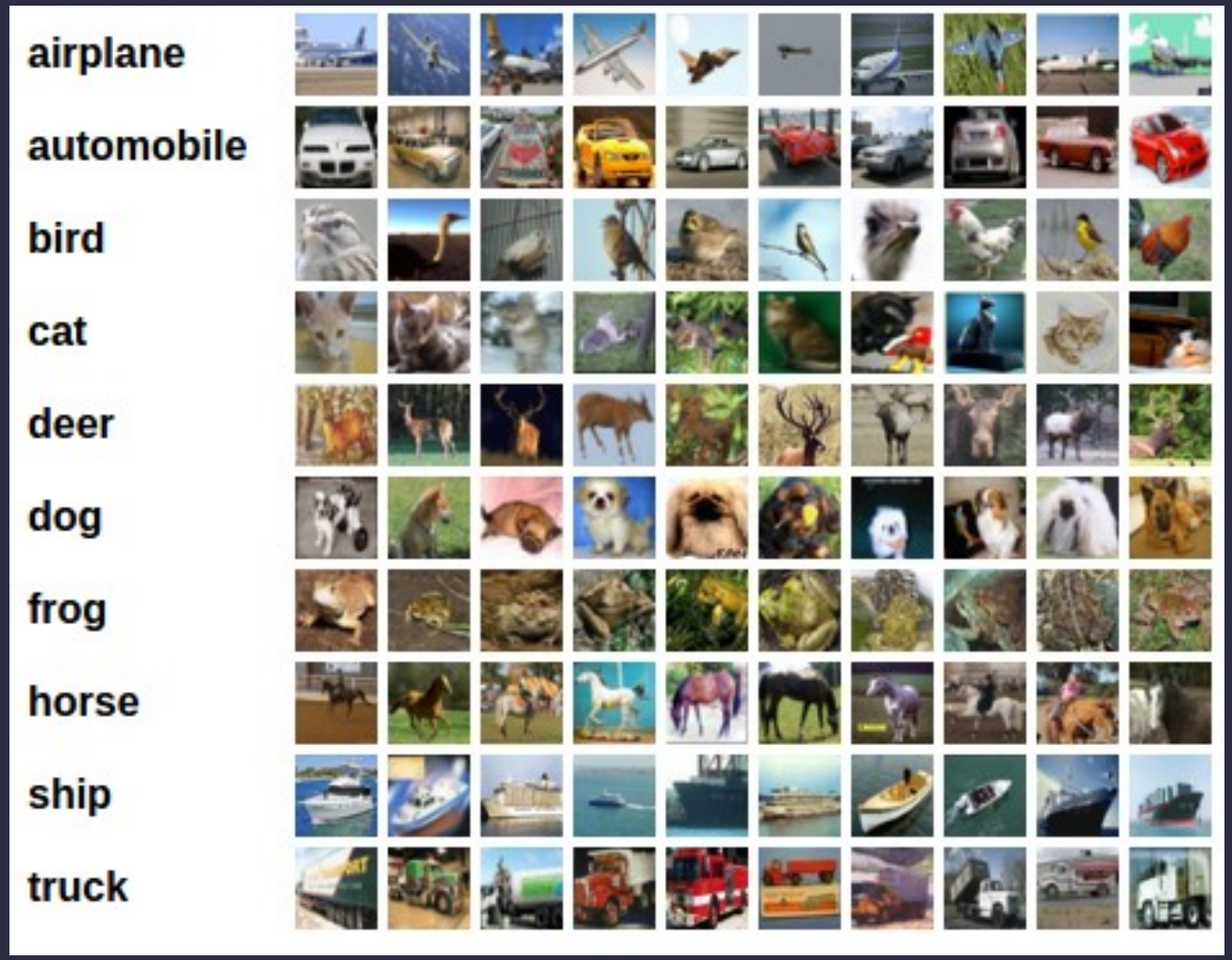




Outline

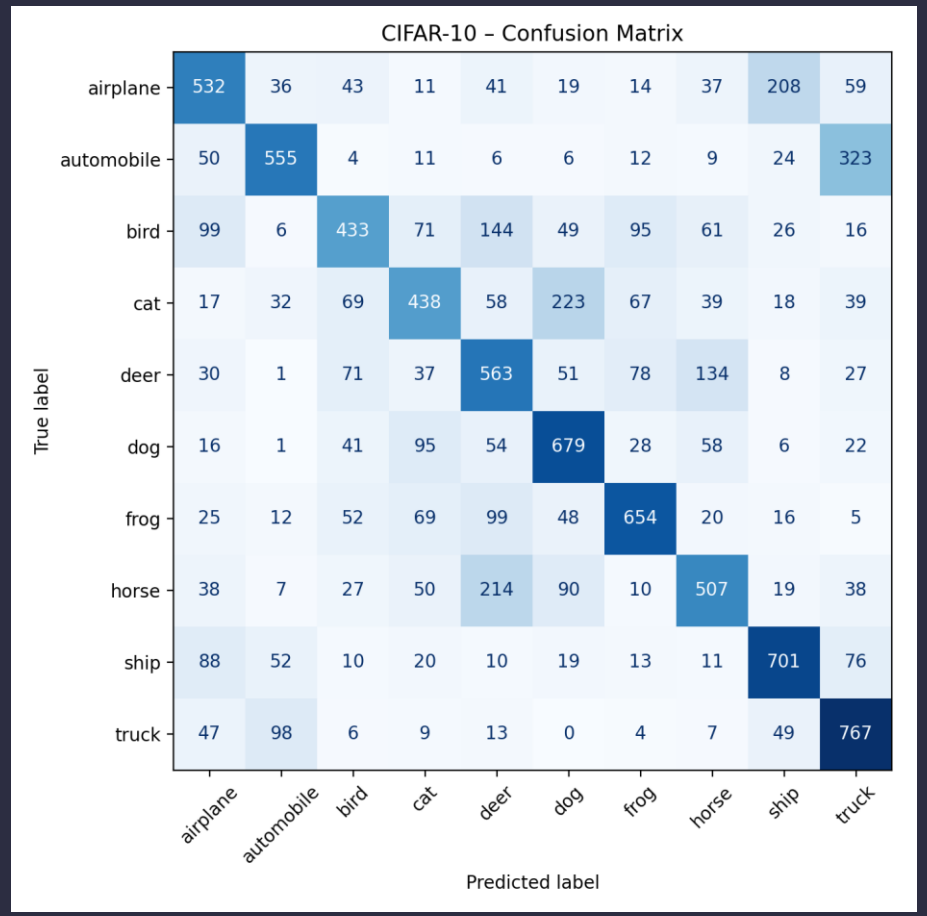
Image Classification

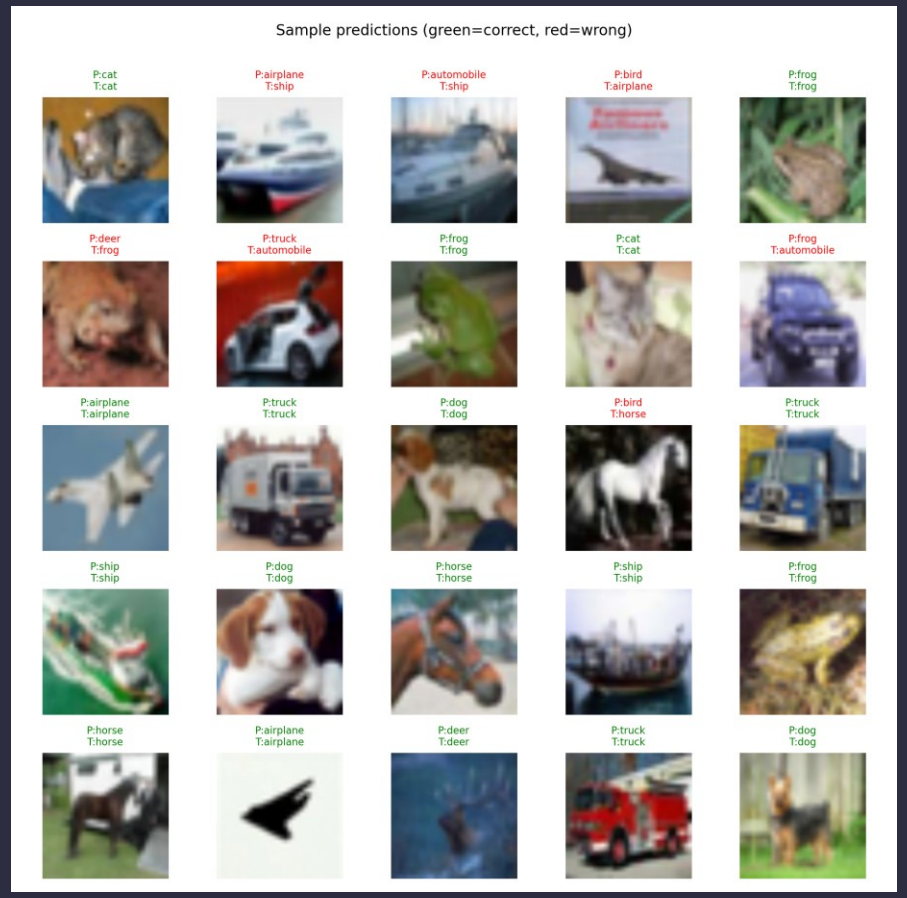
- Dataset: CIFAR-10 — natural color images across 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck).
- Size & format: 60,000 images total
- Multiclass image classification (predict one of 10 labels per image).



Outline

Image Classification





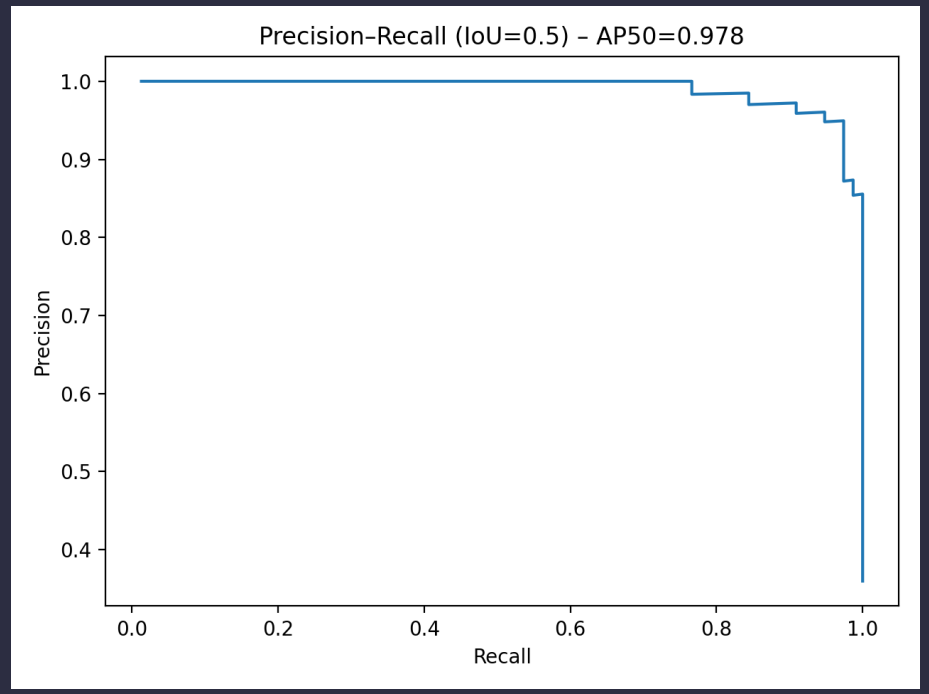
Outline

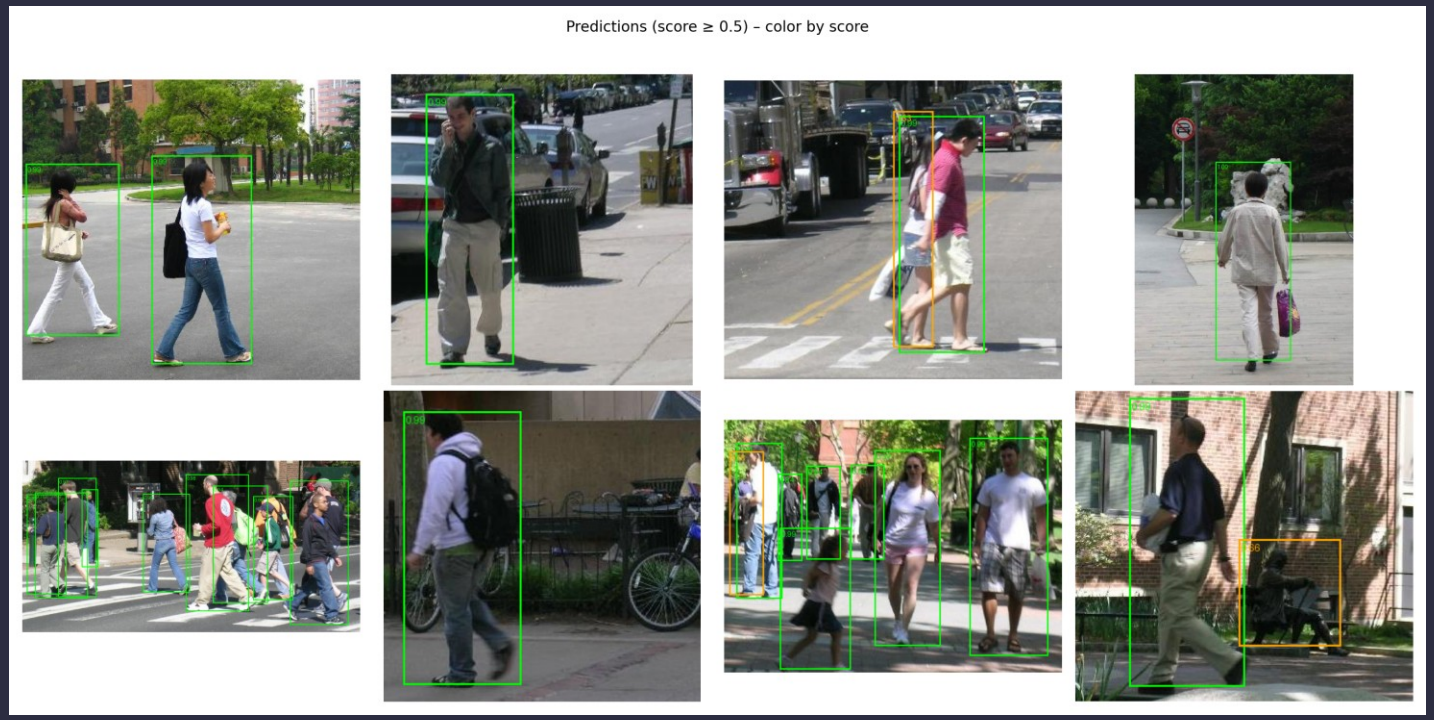
- Image – Object Detection

- Dataset: Penn–Fudan Pedestrians — urban RGB images with pedestrian masks (converted to bounding boxes);
- Task: Object detection — predict bounding boxes and class scores for people in each image.
- Model: Faster R-CNN (ResNet-50 FPN, pretrained)

Outline

- Image – Object Detection





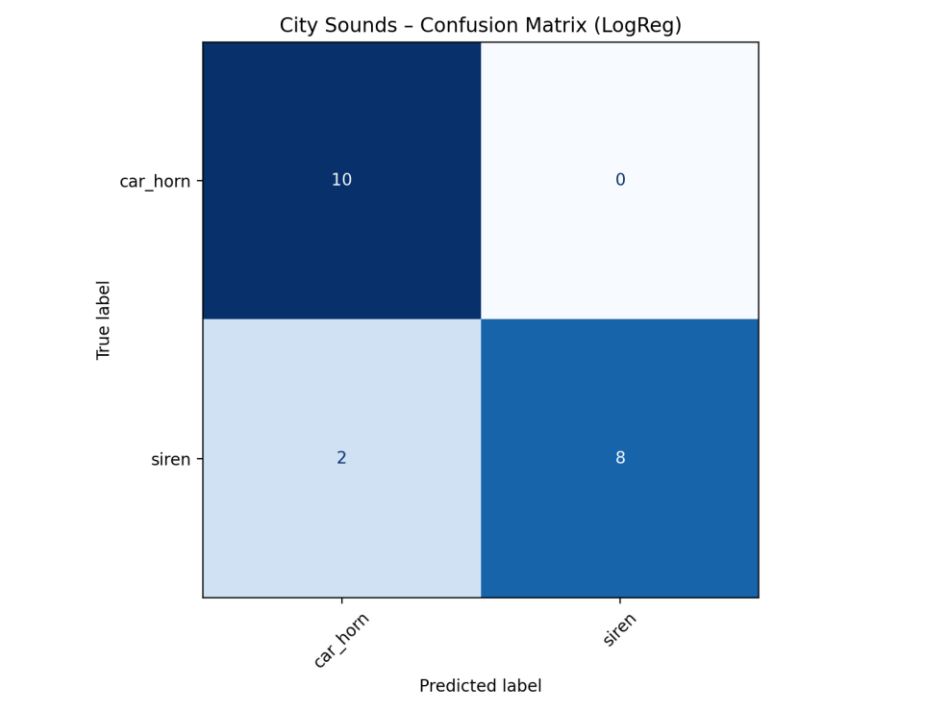
Outline

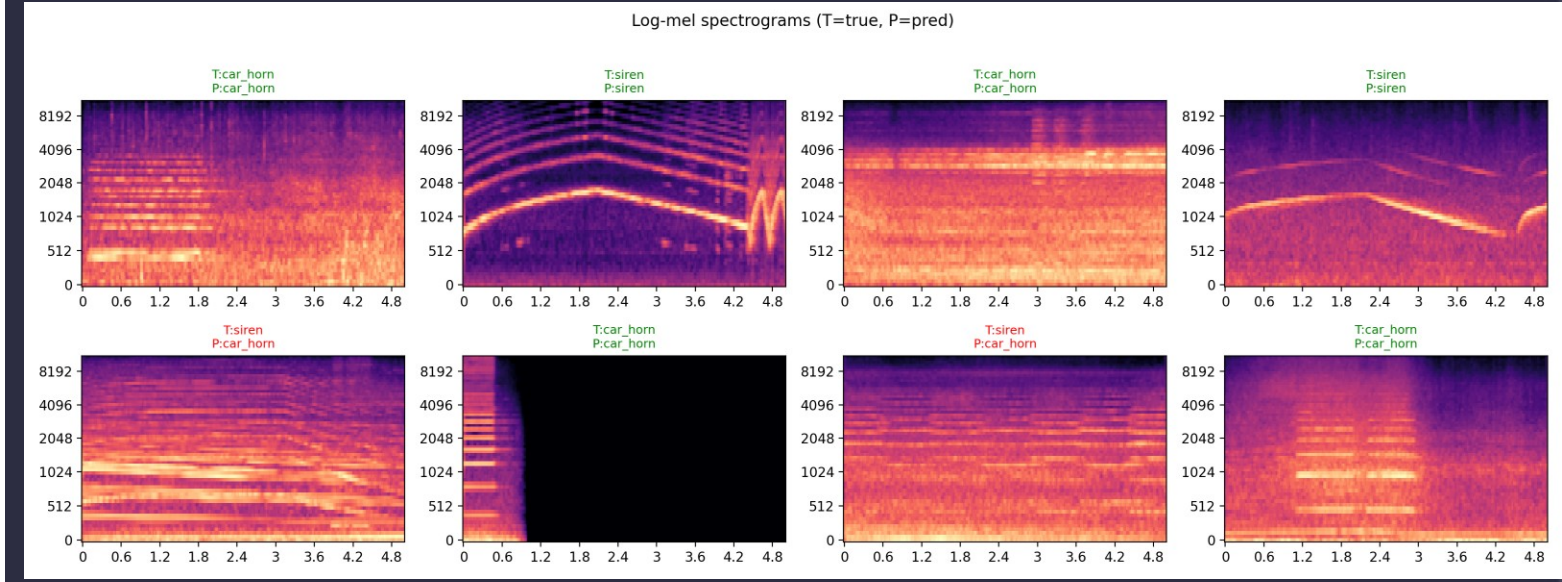
- Sound – City Sounds Classification

- Dataset: ESC-50 (subset of city sounds) — siren, car_horn, drilling, engine_idling, jackhammer, street_music; 5-second WAV clips.Features:
- Log-mel spectrograms

Outline

- Sound – City Sounds Classification





Outline

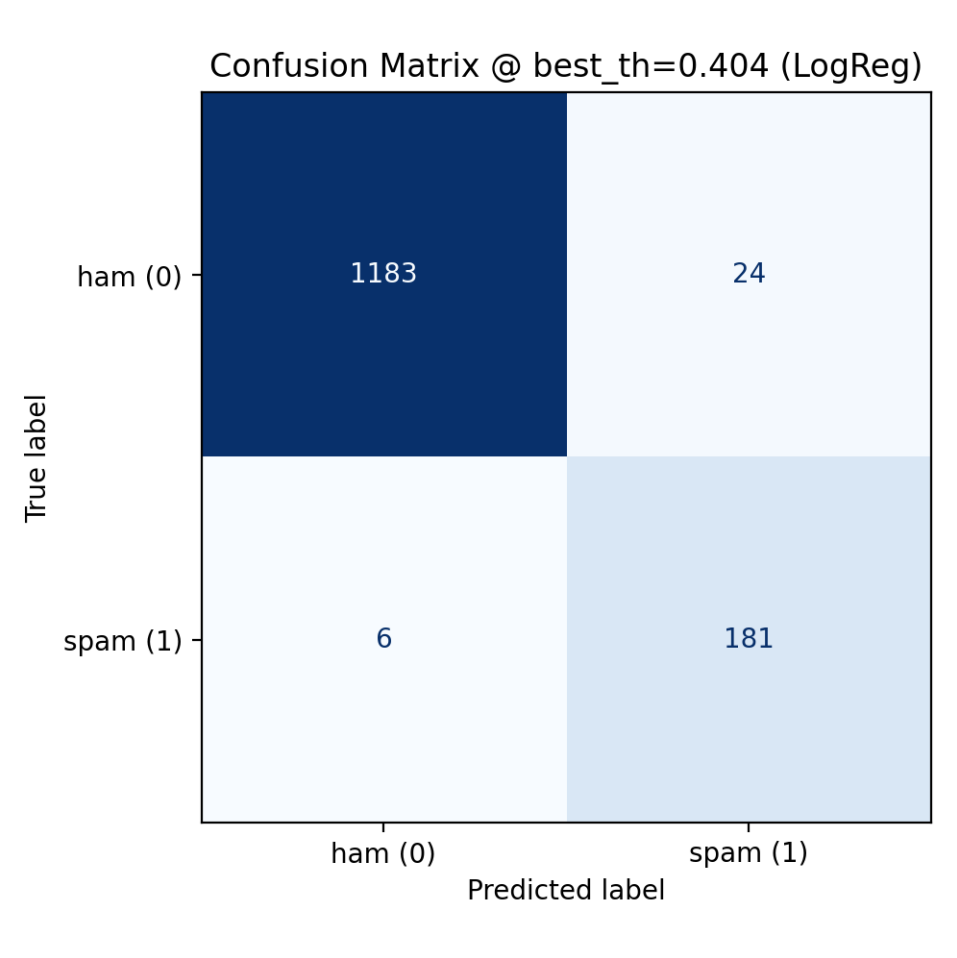
- Text – SMS Spam Classification

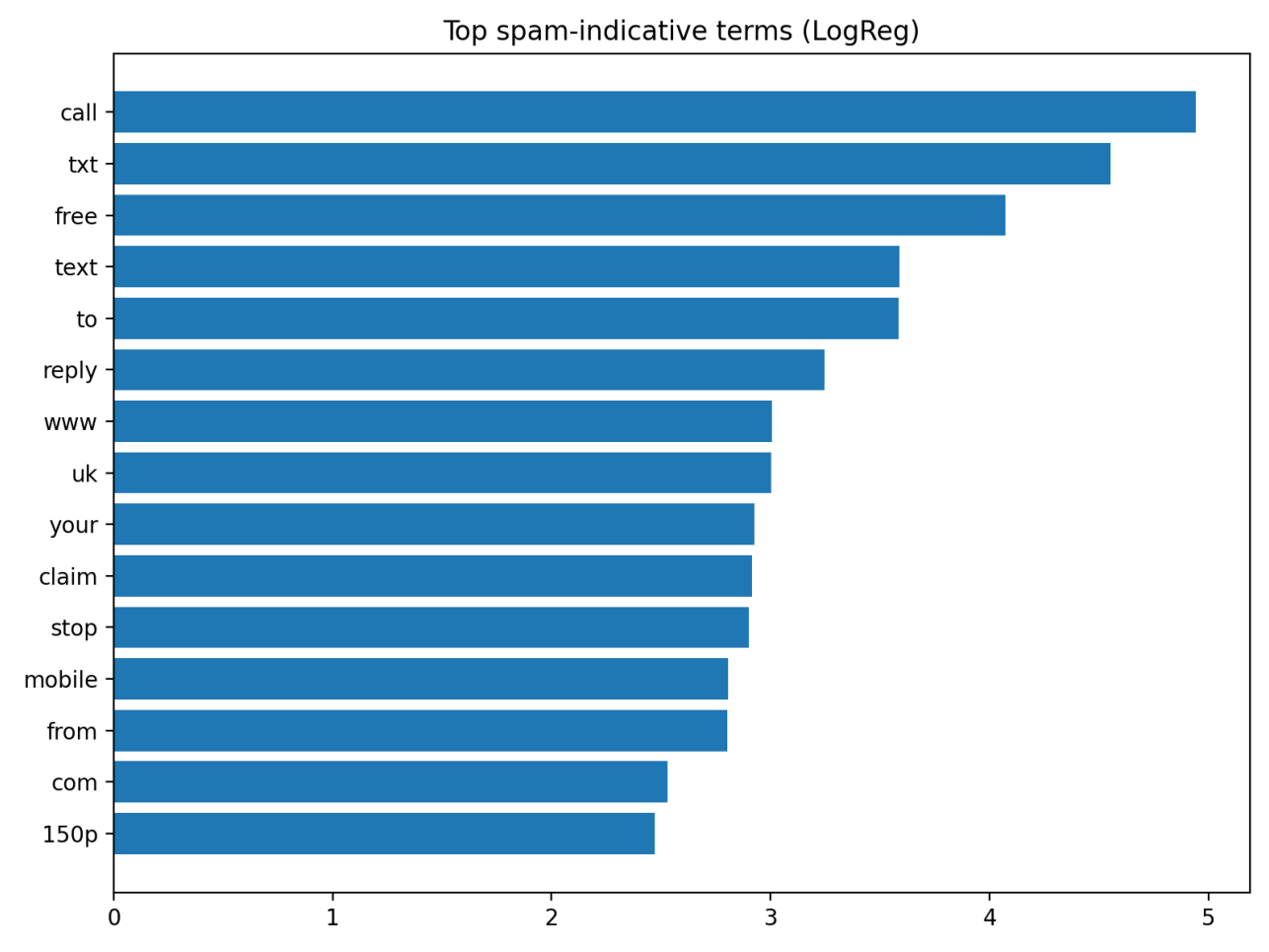
- Dataset: UCI SMS Spam Collection — 5.5k messages; label = ham/spam.
- Features: Word TF-IDF with 1–2-grams, vocab capped for speed.
- Models: Logistic Regression (balanced) and Naive Bayes (calibrated).

	Label	SMS	predicted
0	ham	Later i guess. I needa do mcat study too.	ham
1	ham	But i haf enuff space got like 4 mb...	ham
2	spam	Had your mobile 10 mths? Update to latest Oran...	spam
3	ham	All sounds good. Fingers . Makes it difficult ...	ham
4	ham	All done, all handed in. Don't know if mega sh...	ham

Outline

- Text – SMS Spam Classification

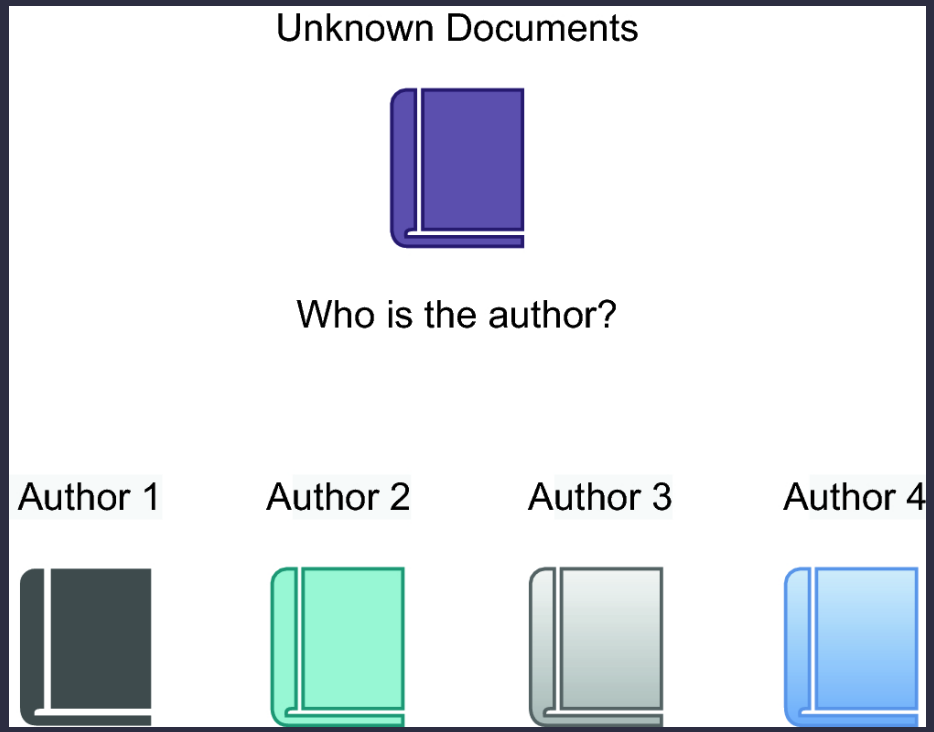




Outline

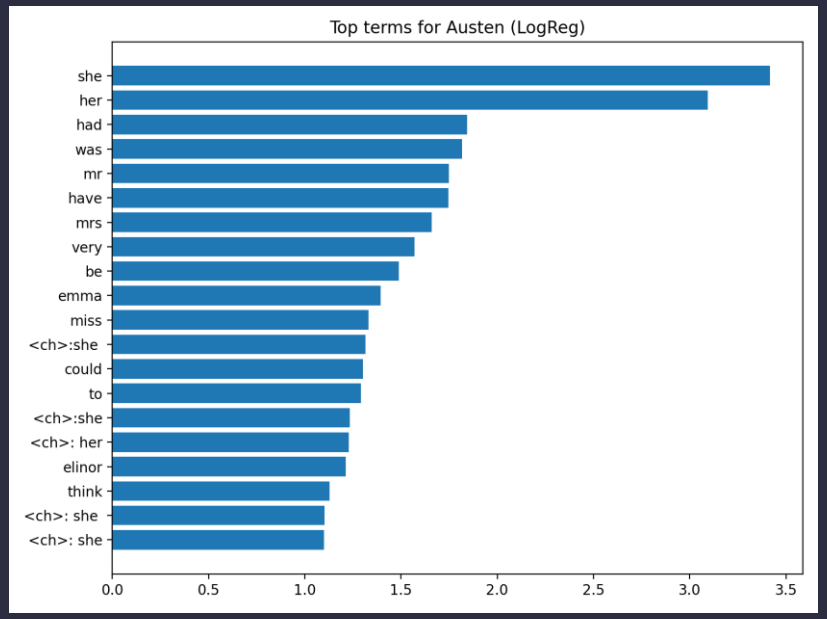
- Text – Author Recognation

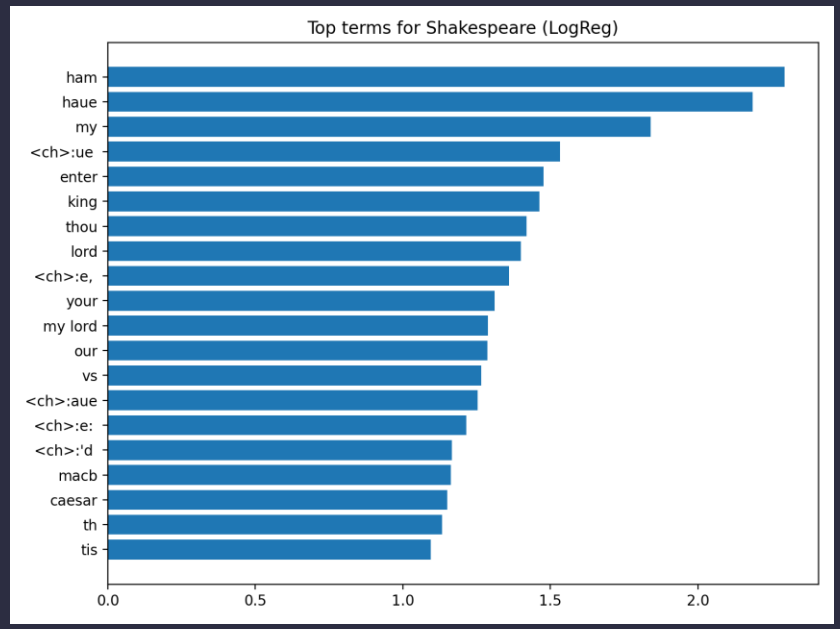
- Dataset: NLTK Gutenberg corpus — paragraph chunks from Jane Austen, William Shakespeare, and Herman Melville (balanced per author).
- Task: Authorship attribution — predict the author of a text snippet.

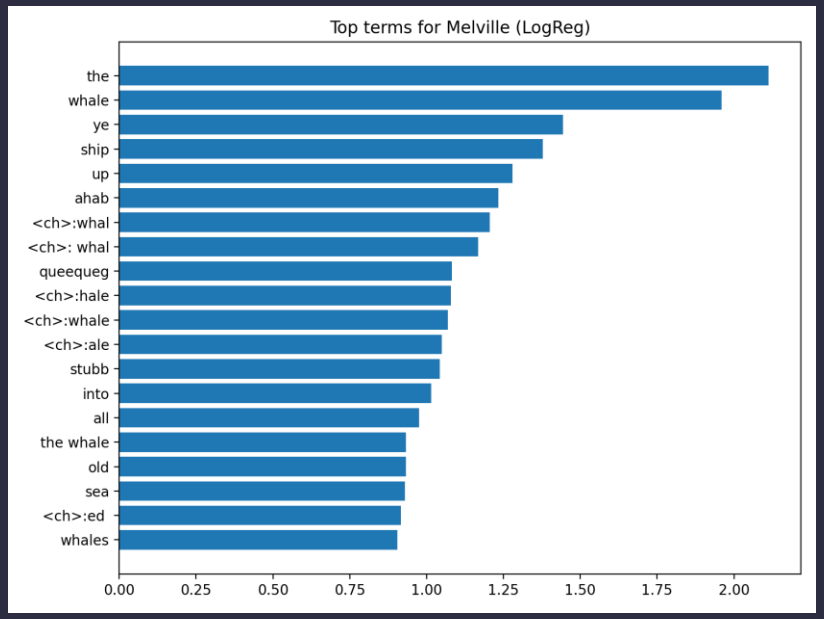


Outline

- Text – Author Recognation

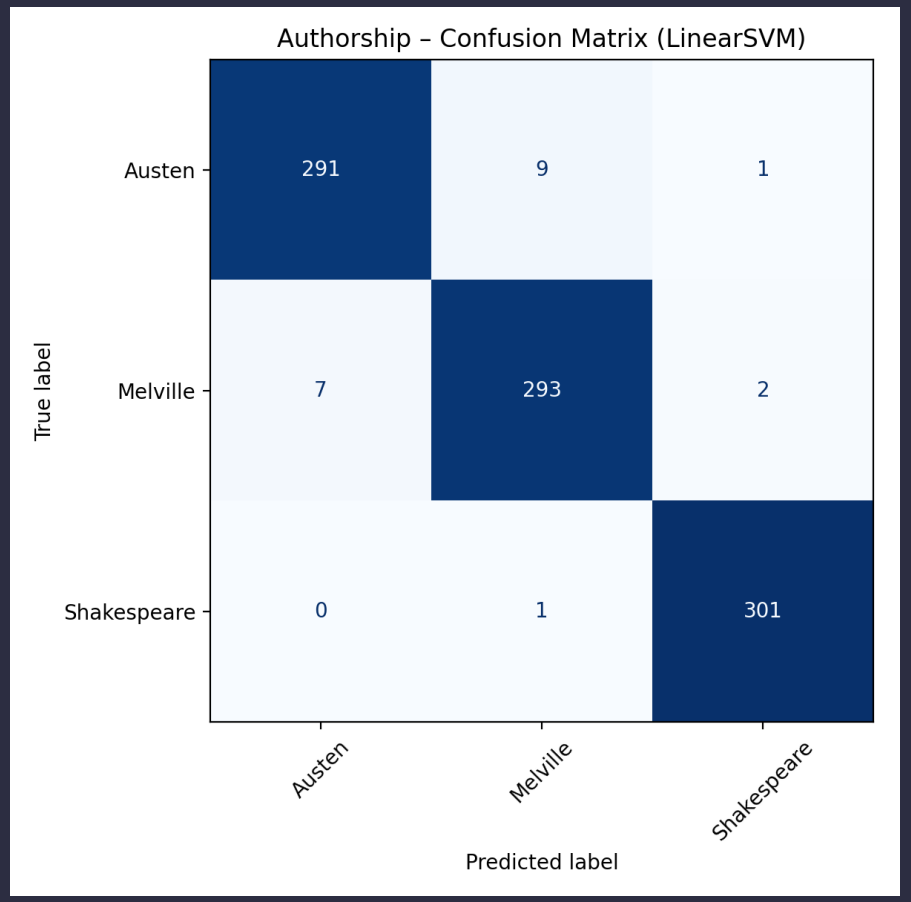


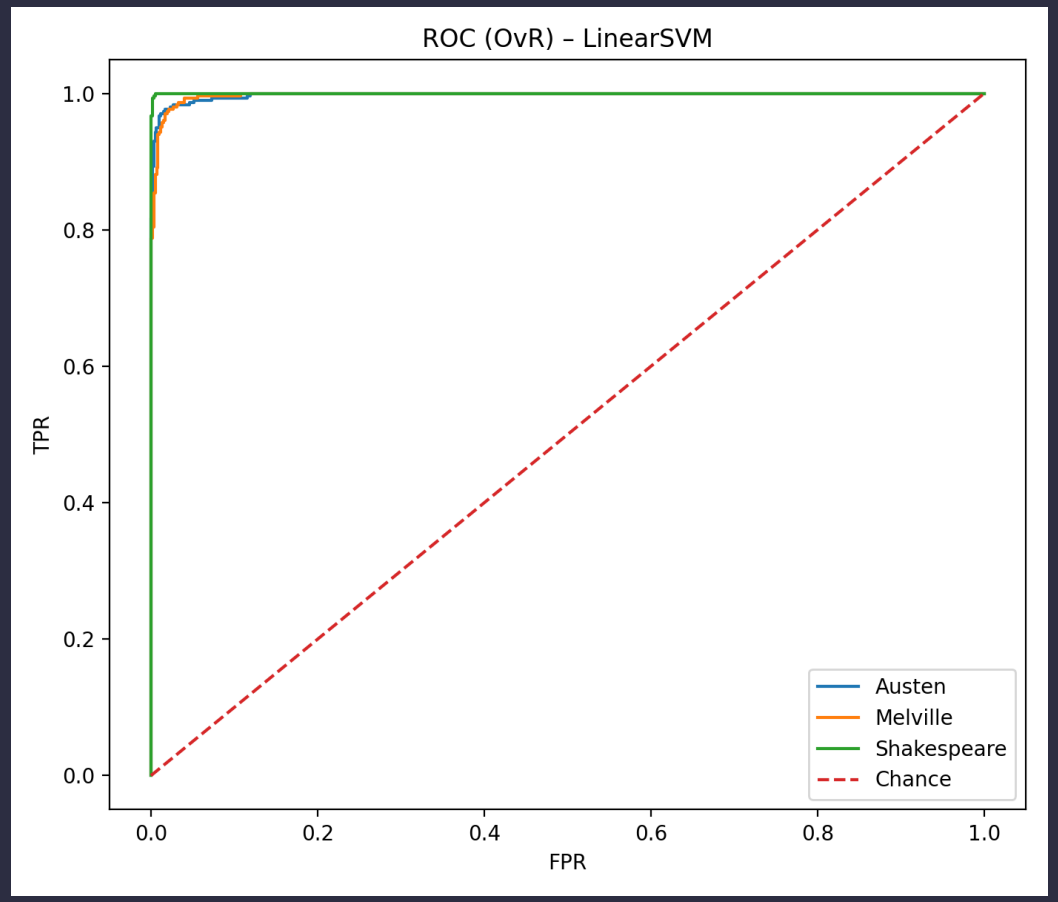




Outline

- Text – Author Recognation





Data Mining—What's in a Name?

Information Harvesting

Data Mining

Knowledge Mining

Knowledge Discovery in Databases

Data Dredging

Data Archaeology

Data Pattern Processing

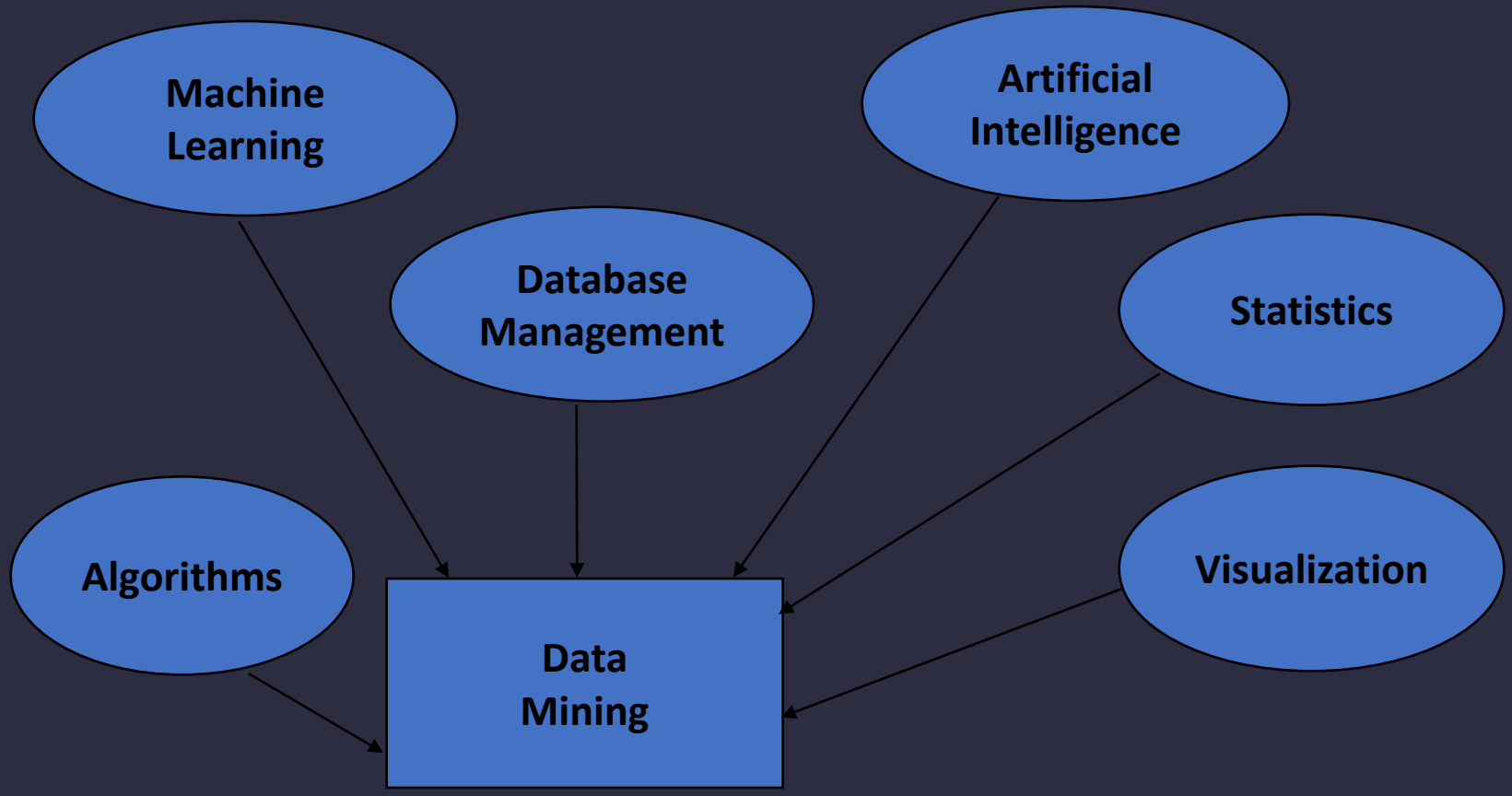
Database Mining

Knowledge Extraction

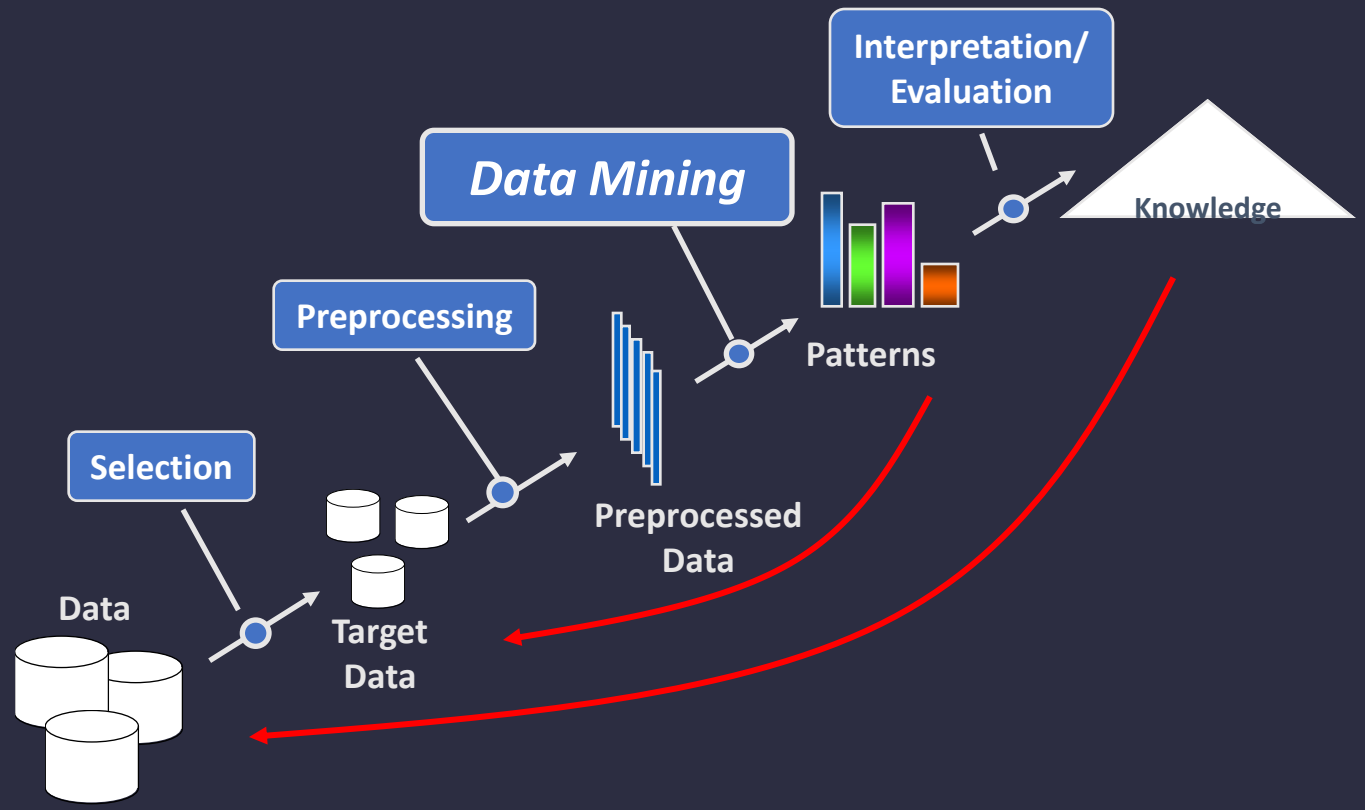
The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of stored data,

- **Using pattern recognition technologies**
- **Statistical and mathematical techniques**
- **Artificial Intelligence**

Integration of Multiple Technologies



Knowledge Discovery in Databases: Process



Multi-Dimensional View of Data Mining

- Data to be mined
  - Relational
  - Text
  - Multi-media
  - Heterogeneous

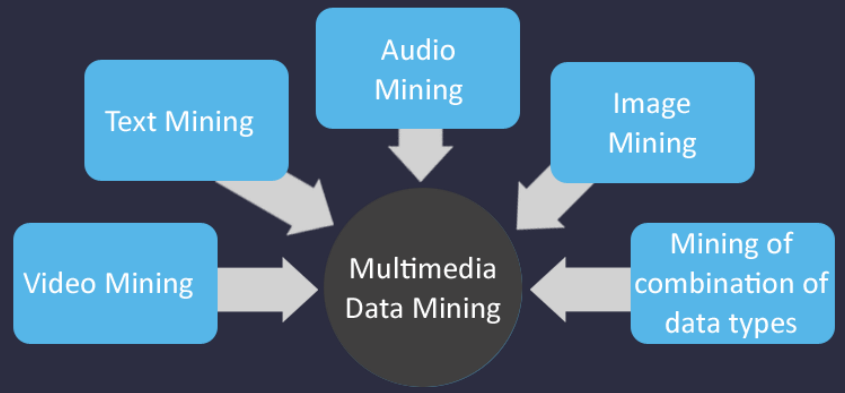
Transaction table

Transaction ID	Customer ID	Product ID	Purche date
53666	24221	389	06-02-2023
50333	24222	789	06-02-2023
54673	24223	879	06-02-2023
58930	24224	975	06-02-2023

Product table

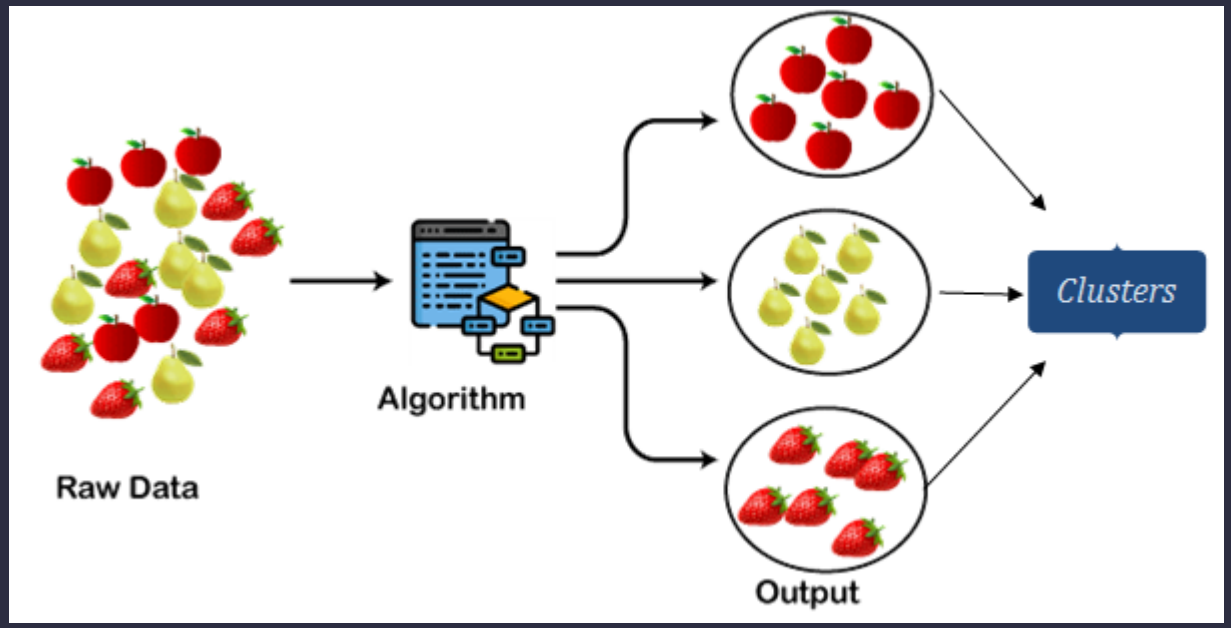
Product ID	Product name	Price per kg
389	Banana	4
789	Apple	5
879	Watermelon	5
975	Mango	7

Customer ID	Last name	First name
24221	Smith	James
24222	Jones	Sam
24223	Taylor	Ann



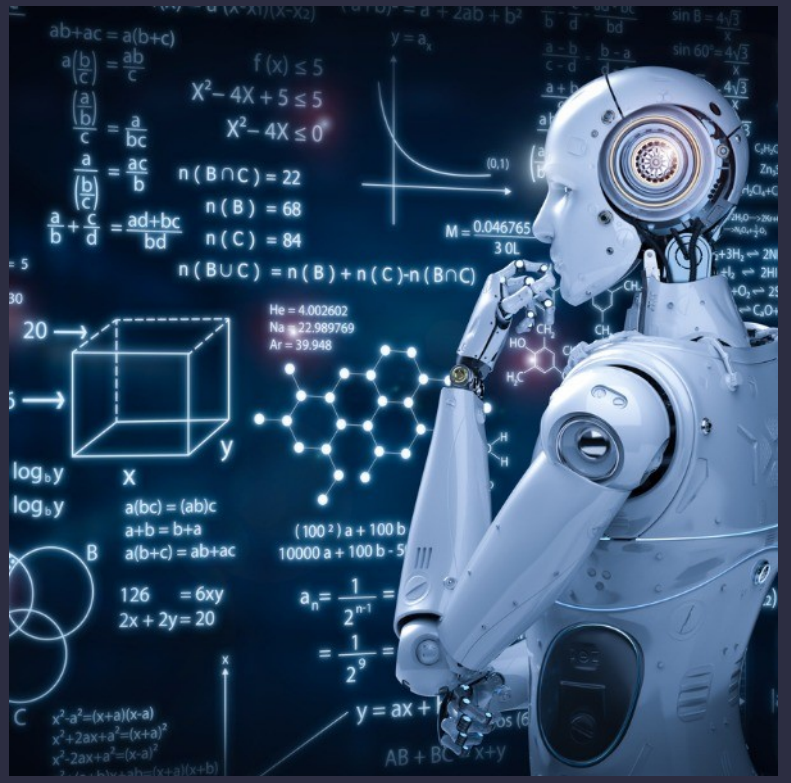
Multi-Dimensional View of Data Mining

- Knowledge to be mined
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels



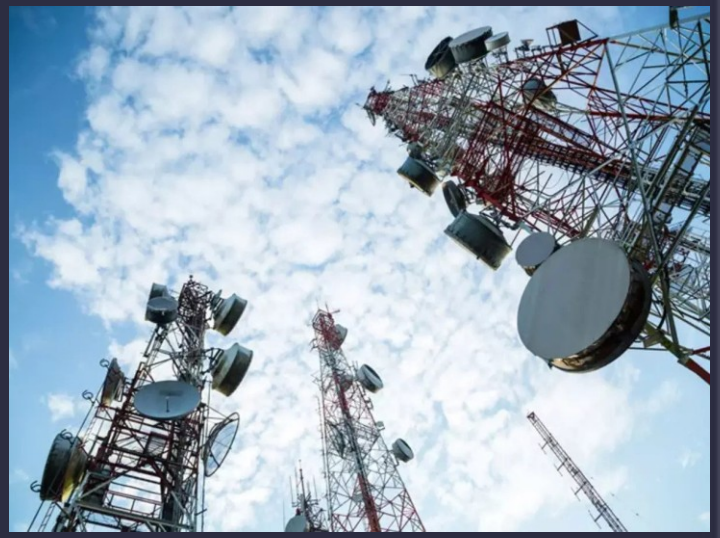
Multi-Dimensional View of Data Mining

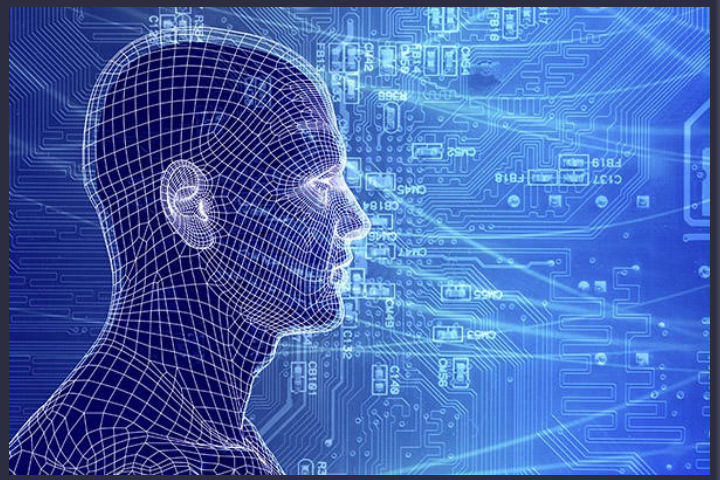
- Techniques Utilized
  - Machine learning
  - Statistics
  - AI
  - Visualization
  - etc…



Multi-Dimensional View of Data Mining

- Applications adapted
  - Retail
  - Telecommunication
  - Banking
  - Fraud analysis
  - Bio-data mining
  - Stock market analysis
  - Web mining
  - etc.





Data Mining: History of the Field

- Knowledge Discovery in Databases workshops started ‘89
  - Now a conference under the auspices of ACM SIGKDD
  - IEEE conference series started 2001
- Key founders / technology contributors:
  - Usama Fayyad, JPL (then Microsoft, now has his own company, Digimine)
  - Gregory Piatetsky-Shapiro (then GTE, now his own data mining consulting company, Knowledge Stream Partners)
  - Rakesh Agrawal (IBM Research)

The term “data mining” has been around since at least 1983 – as a pejorative term in the statistics community

Example: Use in retailing

- Goal: Improved business efficiency
  - Improve marketing (advertise to the most likely buyers)
  - Inventory reduction (stock only needed quantities)
- Information source: Historical business data
  - Example: Supermarket sales records

Date/Time/Register	Fish	Turkey	Cranberries
12/6 13:15 2	N	Y	Y
12/6 13:16 3	Y	N	N

  - Size ranges from 50k records (research studies) to terabytes (years of data from chains)
  - Data is already being warehoused
- Sample question – what products are generally purchased together?
- The answers are in the data, if only we could see them

Data Mining applied to Aviation Safety Records

- Many groups record data regarding aviation safety including the National Transportation Safety Board (NTSB) and the Federal Aviation Administration (FAA)
- Integrating data from different sources as well as mining for patterns from a mix of both structured fields and free text is a difficult task
- Data mining can be used to improve airline safety by finding patterns that predict safety problems

Aircraft Accident Report

- This data mining effort is an extension of the FAA Office of System Safety's Flight Crew Accident and Incident Human Factors Project
- In this previous approach two database-specific human error models were developed based on general research into human factors
  - FAA's Pilot Deviation database (PDS)
  - NTSB's accident and incident database
- These error models check for certain values in specific fields
- Result
  - Classification of some accidents caused by human mistakes and slips.

Data Mining Ideas: Logistics

- Delivery delays
  - Debatable what data mining will do here; best match would be related to "quality analysis": given lots of data about deliveries, try to find common threads in "problem" deliveries
- Predicting item needs
  - Seasonal
    - Looking for cycles, related to similarity search in time series data
    - Look for similar cycles between products, even if not repeated
  - Event-related
    - Sequential association between event and product order (probably weak)

What Can Data Mining Do?

- Cluster
- Classify
  - Categorical
- Association
- Sequence analysis
  - Time-series analysis, Sequential associations

Clustering

- Find groups of similar data items
- Statistical techniques require some definition of “distance” (e.g. between travel profiles) while conceptual techniques use background concepts and logical descriptions

Uses:

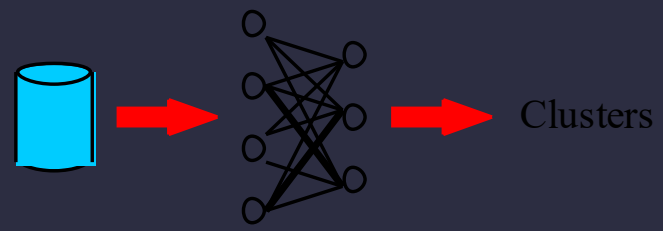
- Demographic analysis

Technologies:

- Self-Organizing Maps
- Probability Densities
- Conceptual Clustering

“Group people with similar travel profiles”

- George, Patricia
- Jeff, Evelyn, Chris
- Rob



Classification

- Find ways to separate data items into pre-defined groups
  - We know X and Y belong together, find other things in same group
- Requires "training data": Data items where group is known

Uses:

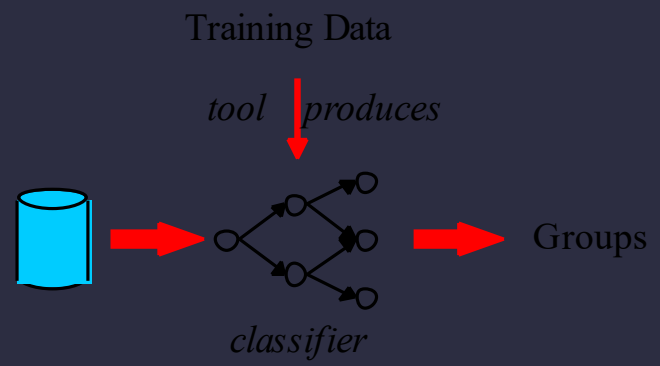
- Profiling

Technologies:

- Generate decision trees (results are human understandable)
- Neural Nets

"Route documents to most likely interested parties"

- English or non-english?
- Domestic or Foreign?



Association

- Identify dependencies in the data:
  - X makes Y likely
- Indicate significance of each dependency

Uses:

- Targeted marketing

"Find groups of items commonly purchased together"

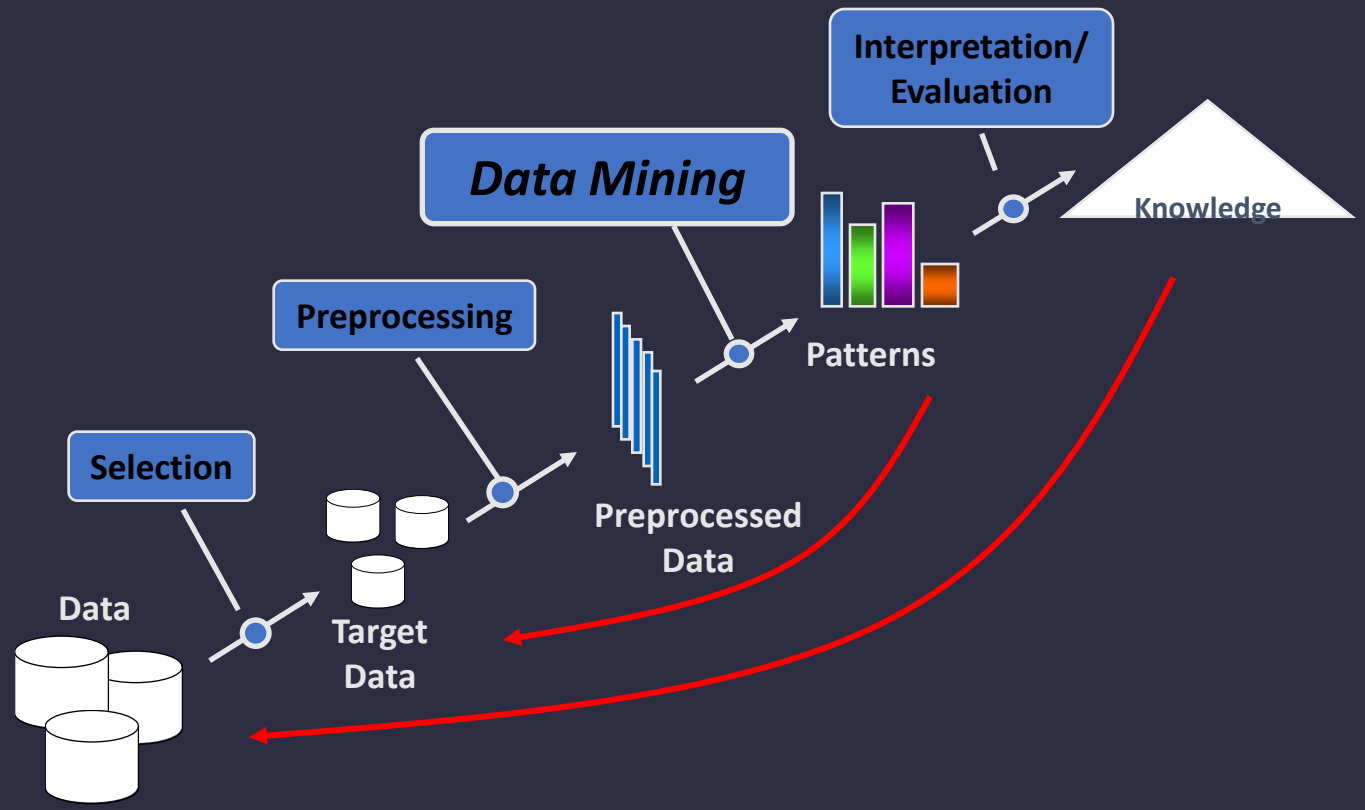
- People who purchase fish are extraordinarily likely to purchase wine
- People who purchase Turkey are extraordinarily likely to purchase cranberries

Date/Time/Register	Fish	Turkey	Cranberries
12/6 13:15 2	N	Y	Y
12/6 13:16 3	Y	N	N

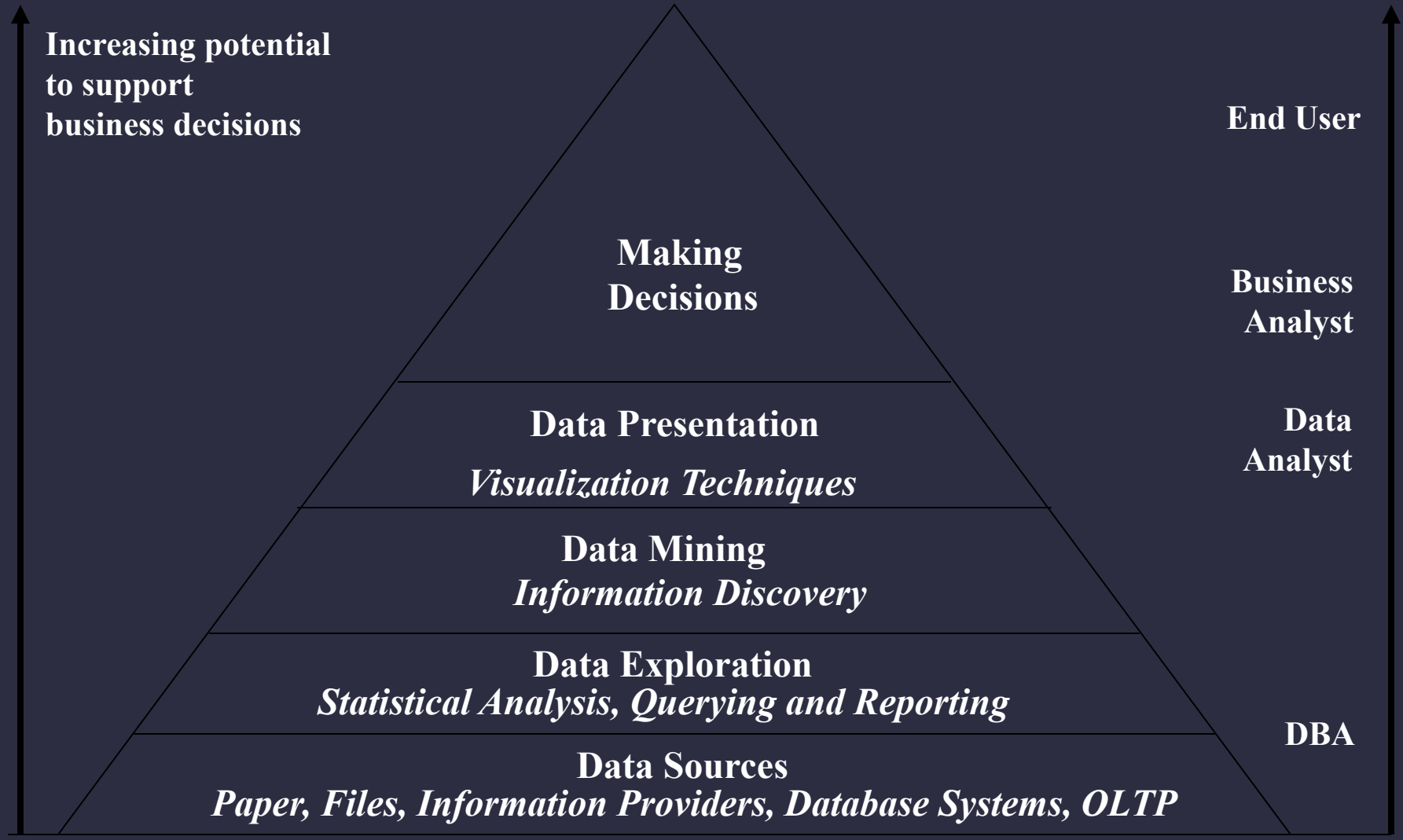
Data Mining Complications

- Volume of Data
  - Clever algorithms needed for reasonable performance
- Interest measures
  - How do we ensure algorithms select “interesting” results?
- “Knowledge Discovery Process” skill required
  - How to select tool, prepare data?
- Data Quality
  - How do we interpret results in light of low quality data?
- Data Source Heterogeneity
  - How do we combine data from multiple sources?

Knowledge Discovery in Databases: Process

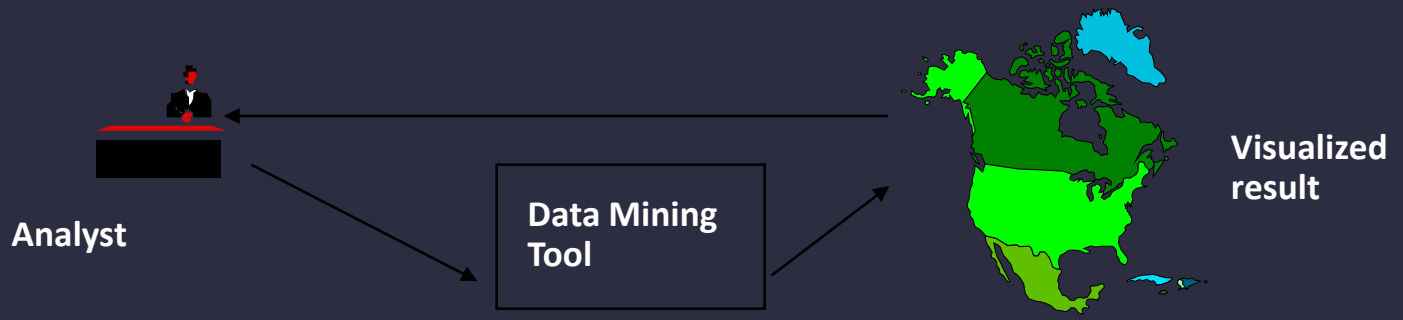


Data Mining


Increasing potential
to support
business decisions
Making
Decisions
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Analysis, Querying and Reporting
Data Sources
Paper, Files, Information Providers, Database Systems, OLTP
End User
Business
Analyst
Data
Analyst
DBA

Data Mining and Visualization

- Approaches
  - Visualization to display results of data mining
    - Help analyst to better understand the results of the data mining tool
  - Visualization to aid the data mining process
    - Interactive control over the data exploration process
    - Interactive steering of analytic approaches (“grand tour”)
- Interactive data mining issues
  - Relationships between the analyst, the data mining tool and the visualization tool



Data Mining and Visualization

Python Background

- https://www.youtube.com/watch?v=dLp9U1goMPM&list=PLFmsF38Rfcg3nOrI3Kc0dofL8pFSi6Do