UNIVERSITY of LOUISIANA LAFAYETTE

CSCE566-DATA MINING
WEEK 1

Introduction to Data Mining

Min Shi
min.shi@louisiana.edu

Aug 26, 2024

Hello

- **Lecture, Dr. Min Shi**
- Assistant Professor, School of Computing & Informatics
- **Research Interests:**
  - Data Mining
  - Deep Learning
  - Graph Neural Networks
  - Medical Image Analysis
- **Office:** James R. Oliver Room 350
- **Email:** min.shi@louisiana.edu
- **Lab Website:** medai-lab.com

Hello

- **Teaching Assistant**
- TBD

Course Information

- **Lecture Times:** Mondays and Wednesdays, 2:30 pm to 3:45 pm
- **Location:** James R. Oliver, Room 119A
- **Office Hours:** Mondays 4 PM – 5 PM
- **Course Website:**

  https://www.medai-lab.com/teaching/2024-fall-CSCE566
- **Prerequisites**: Python, Jupyter, Scikit-learn, Tensorflow Keras

  (*visit course website for learning resources*)
- **Textbook:** This course does not have a required textbook

  (*visit course website for recommendations*)

Course Grading

Course Evaluation Method

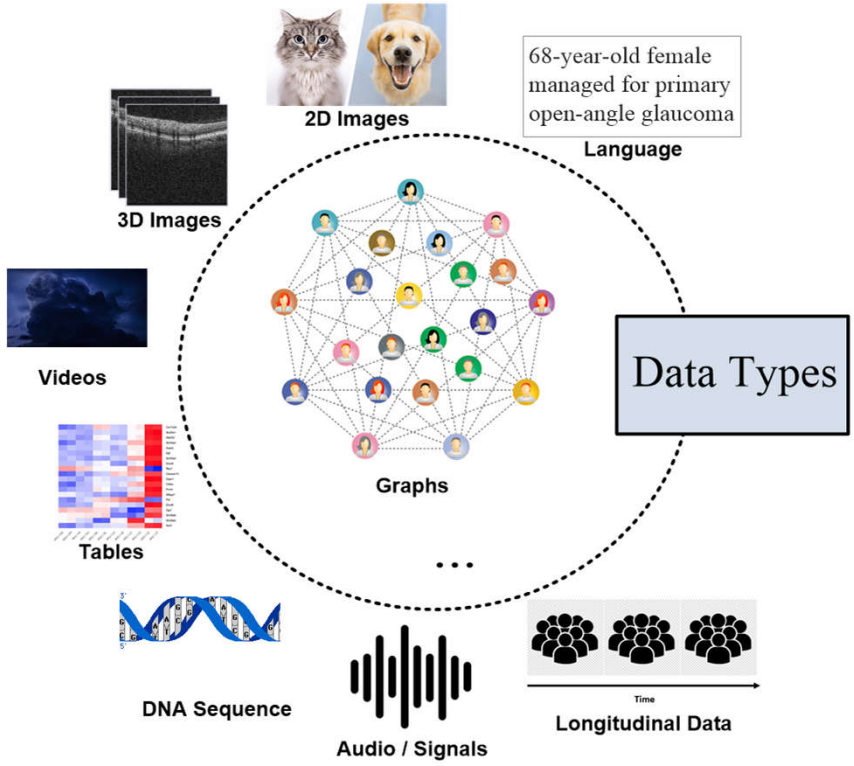
Item	Percentage	Note
Homework	30%	2 HWs
Midterm Exam	20%	
Paper presentation	20%	
Project report and code	30%	

Course Grading Scale

90-100%	A
80-89%	B
70-79%	C
60-69%	D
0-59%	F

Lecture Content

Index	Topics
1	Introduction to data mining
2	Frequent itemset mining
3	Matrix data mining
4	Text data mining
5	Image data mining
6	Graph data mining
7	Time-series data mining
8	Data mining challenges
9	Introduction to deep learning
10	Application: medical image classification
11	Selected paper presentation



Lecture Schedule

Week		Lecture	Topics	Event
Week 1	26-Aug	Introduction to data mining	Intro to class; why data mining; data types; mentoring projects	
	28-Aug		Data mining tasks and evaluations	
Week 2	2-Sep	*Holiday: Labor Day*		HW1 out
	4-Sep	Frequent itemset mining	Frequent pattern analysis; association rules; Apriori algorithm	
Week 3	9-Sep	Matrix data mining	Principle component analysis; clustering algorithms	
	11-Sep		Support vector machine	
Week 4	16-Sep	Text data mining	Vector space model; TF-IDF; naïve bayes	
	18-Sep		Latent Dirichlet Allocation; word embedding	HW1 deadline
Week 5	23-Sep	Image data mining	Linear models and convolutions	
	25-Sep		Unsupervised clustering for images; image categorization, object recognition, segmentation	
Week 6	30-Sep	Graph data mining	Random graph model; PageRank	HW2 out
	2-Oct		Graph embeddings; graph applications	
Week 7	7-Oct	Midterm exam	60 minutes	Paper binding
	9-Oct	Time-series data mining	Vector autoregression (VAR) models; anomaly detection; clustering; classification	
Week 8	14-Oct		Recurrent neural networks; case studies	HW2 deadline
	16-Oct	Data mining challenges	Missing values; noise data; class-imbalanced data; limited data; trustworthiness in data mining	
Week 9	21-Oct	Introduction to deep learning	Feedforward neural networks; backward propagation algorithm	Final projects out
	23-Oct	Application: medical image classification	Hands-on experiences	
Week 10	28-Oct	Selected paper presentation	Presentation: 12 minutes; Question/Answer: 3 minutes	
	30-Oct			
Week 11	4-Oct	Selected paper presentation		
	6-Oct			
Week 12	11-Nov	Selected paper presentation		
	13-Nov			
Week 13	18-Nov	Project work	Feedback on demand	Project report and codes due on 11/30
	20-Nov			

Paper Presentation & Project

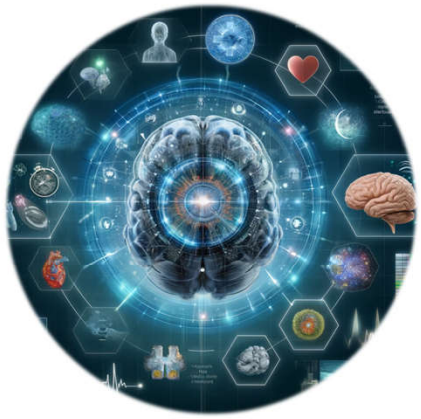
- **Paper Presentation (20%):**
  - Select a research paper to read and present
  - Papers released on October 9
  - Paper presentation begins on October 23
- **Final Project (30%):** regular, proposed, or research projects
  - Regular projects will be released on October 23
  - Independently design and implement codes
  - Project report & codes due on November 30
- **Proposed Projects:** consent must be obtained
  - Project proposal due on October 23
  - Project report & codes due on November 30

Paper Presentation & Project

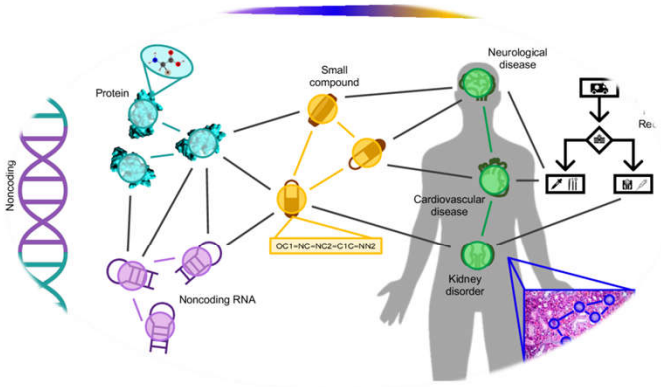
- **Research Project (2-3 students form a team):**
  - Research projects aiming to submit a journal paper (potentially)
  - Tentative project report & codes due on November 30
  - Please send me an email to discuss the details further

Scopes of Research Projects

Multimodal Medical AI



Network Biology and Medicine

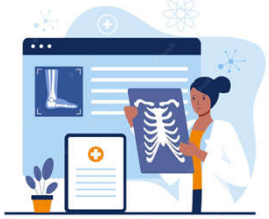


AI Fairness

AI in Medical Education

Ophthalmology

Endoscopy

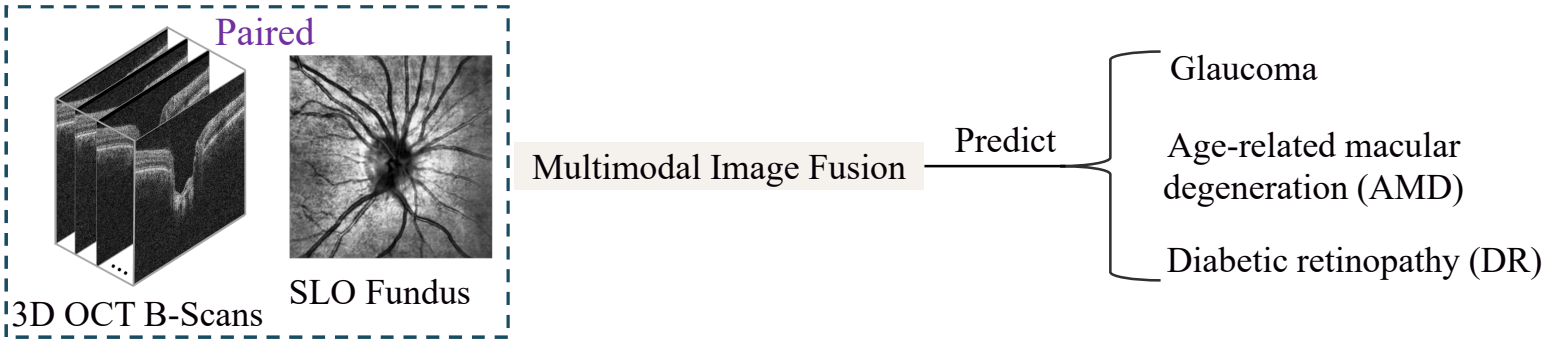


Radiology

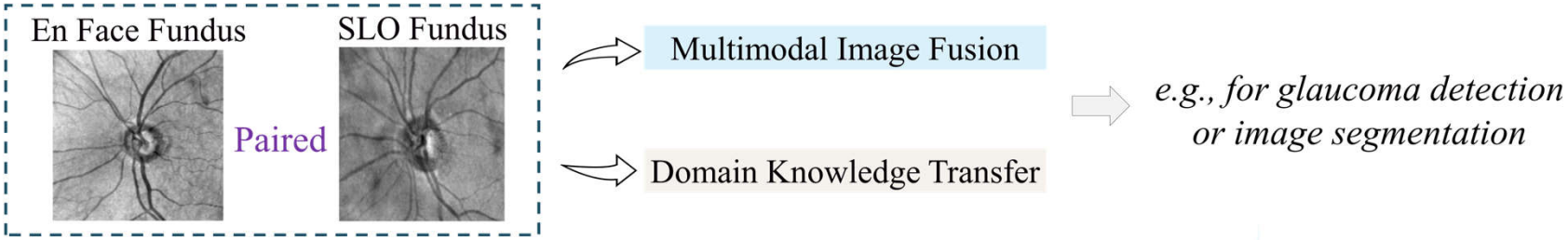
...

Potential Research Projects

1. Multimodal (multitask) medical imaging diagnostics

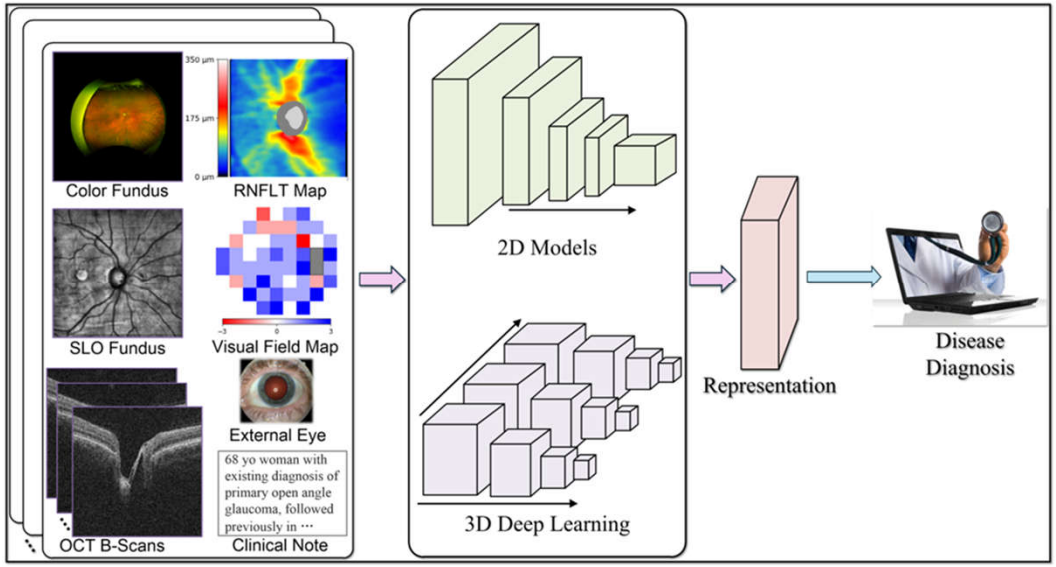


2. Multidomain medical imaging diagnostics



Potential MRP Projects

3. Fairness in medical imaging diagnostics

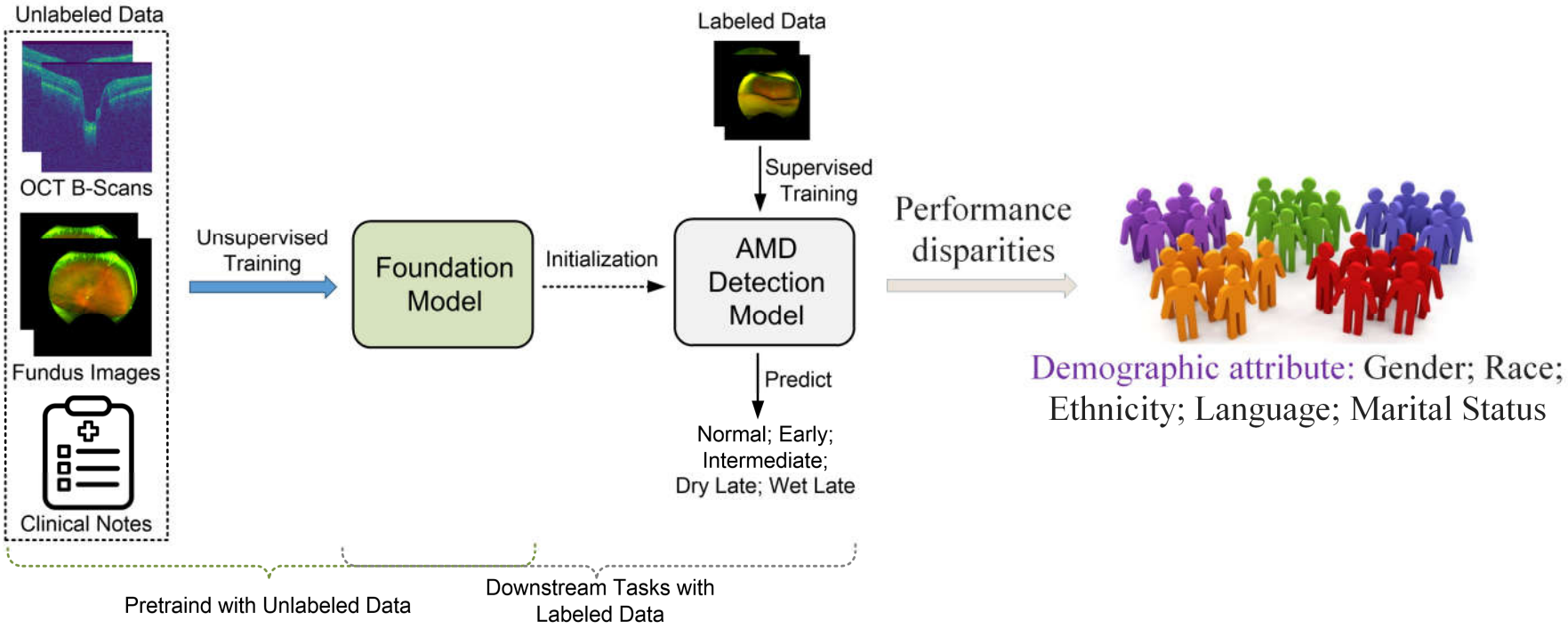


Performance disparities

Demographic attribute: Gender; Race; Ethnicity; Language; Marital Status

Potential MRP Projects

4. Fairness in medical foundation model-based research



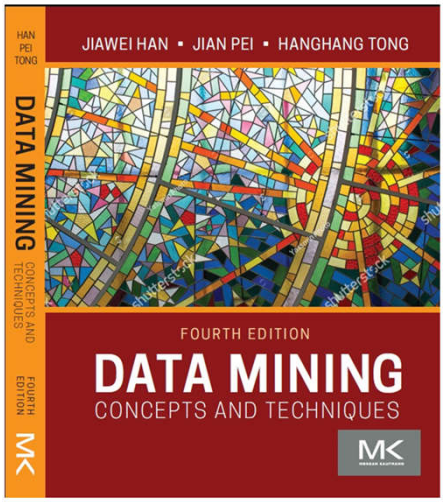
Computing Resources

1. **Personal Laptops/Computers**
2. **Workstations in the Lab**
3. **Google Colab** https://colab.research.google.com/
4. **Environment setup:**
   - Python, numpy, Scikit-learn
   - Jupyter notebook
   - Tensorflow, Keras

Materials / Textbooks (Optional)

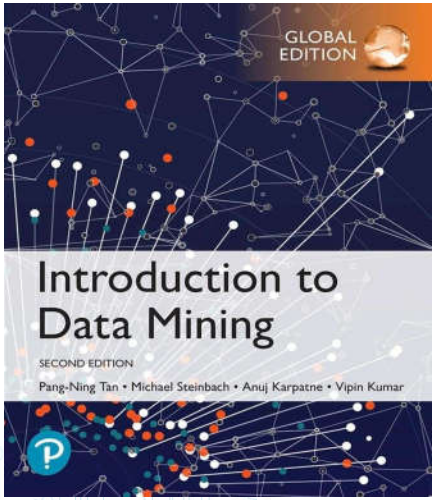
- Jiawei Han, Jian Pei, Hanghang Tong. **Data Mining Concepts and Techniques**. 4th edition. Morgan Kaufmann, 2023.

  The full text for 3rd edition:
  https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf



- Pang-Ning Tan, Michael Steinbach, Vipin Kumar: **Introduction to Data Mining**. 2nd Edition. Pearson / Addison Wesley.

  https://www.ceom.ou.edu/media/docs/upload/Pang-Ning_Tan_Michael_Steinbach_Vipin_Kumar_-_Introduction_to_Data_Mining-Pe_NRDK4fi.pdf



Online Learning Resources

1. **Python**
   - Python for Data Analysis (link: https://wesmckinney.com/book/)
   - 30-Days-Of-Python (link: https://github.com/Asabeneh/30-Days-Of-Python?tab=readme-ov-file#-30-days-of-python)
   - Top 10 GitHub Repos to learn Python (link: https://www.kaggle.com/discussions/getting-started/235259)
2. **Jupyter:**
   - Arnawesome-jupyter (link: https://github.com/markusschanta/awesome-jupyter)
   - Python Data Science Handbook (link: https://github.com/jakevdp/PythonDataScienceHandbook?tab=readme-ov-file)

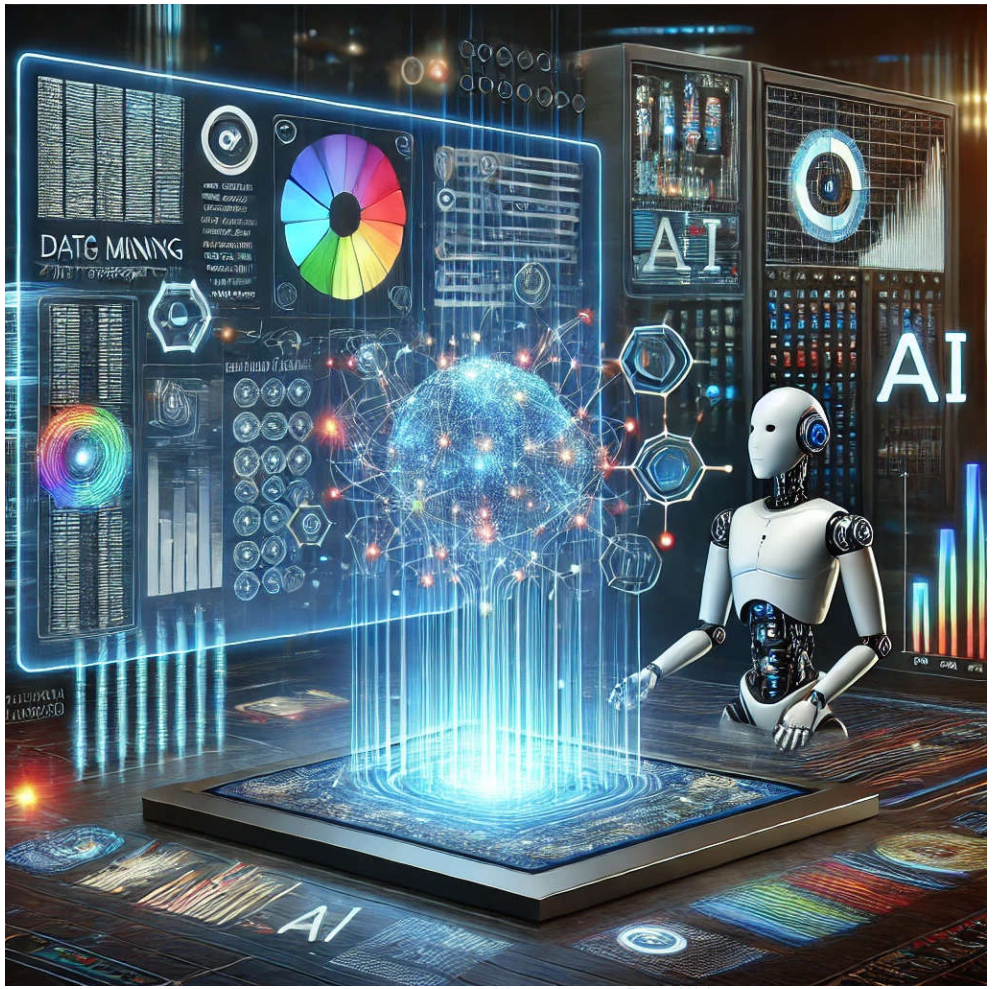
Outline: Introduction to Data Mining

1. What is Data Mining?
2. Data Types
3. Data Mining Tasks
4. Data Mining Process
5. Evaluation Setting and Metrics

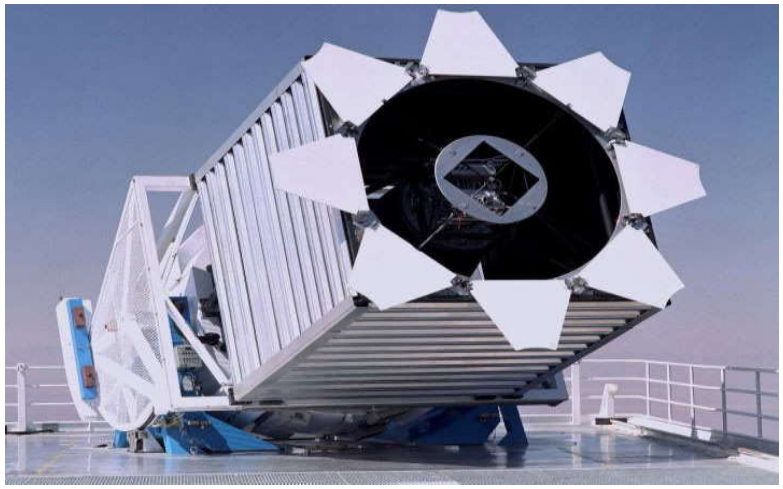
Why data mining?

- Large quantities of data are collected about all aspects of our lives
- This data contains interesting patterns
- Data Mining helps us to
  1. discover these patterns and
  2. use them for decision making

  across all areas of society, including
  - Business and industry
  - Science and engineering
  - Medicine and biotech
  - Government
  - Individuals



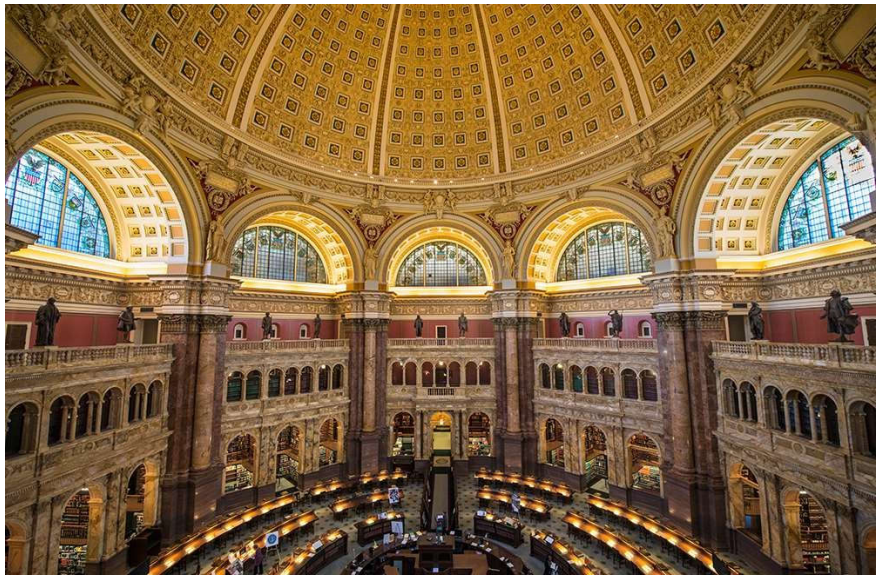
“We are drowning in data…”



Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year

Predict
- Type of sky object:
  Star or galaxy?

“We are drowning in data…”



US Library of Congress
≈ 235 TB archived
≈ 40 Wikipedias

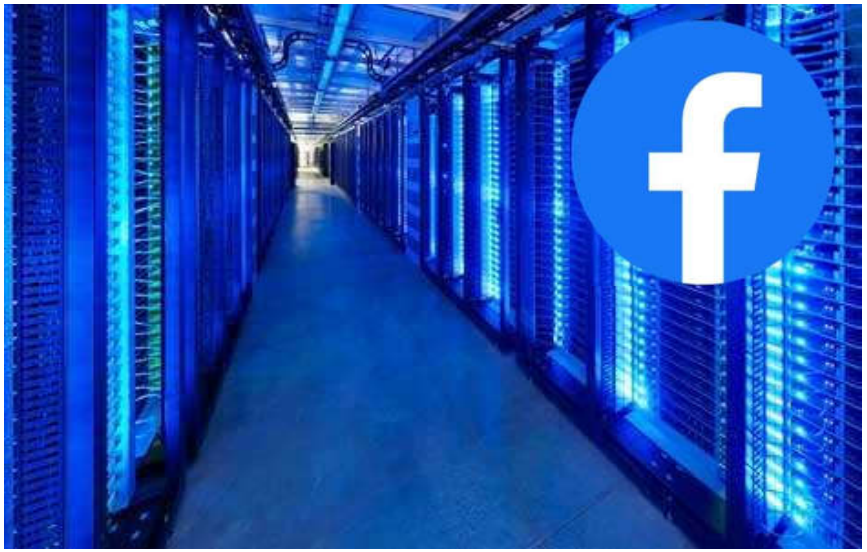
arXiv Preprint Server
> 2 million papers

Discover
- Topic distributions*
- Citation networks

Train
- Large Language Models

* Lansdall-Welfare, et al.: Content analysis of 150 years of British periodicals. PNSA, 2017.

“We are drowning in data…”



Facebook

- 4 Petabyte of new data generated every day
- over 300 Petabyte in Facebook‘s data warehouse

Predict

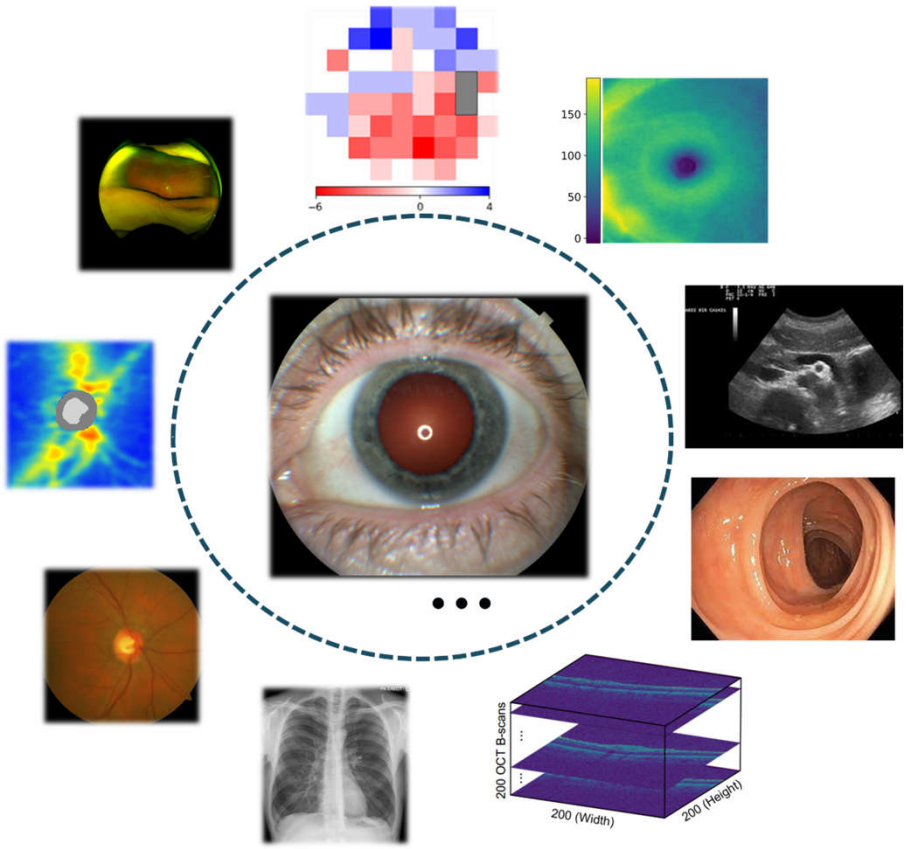
- Interests and behavior of over one billion people

“We are drowning in data…”

Predict

- Interests and behavior of mankind

“We are drowning in data…”

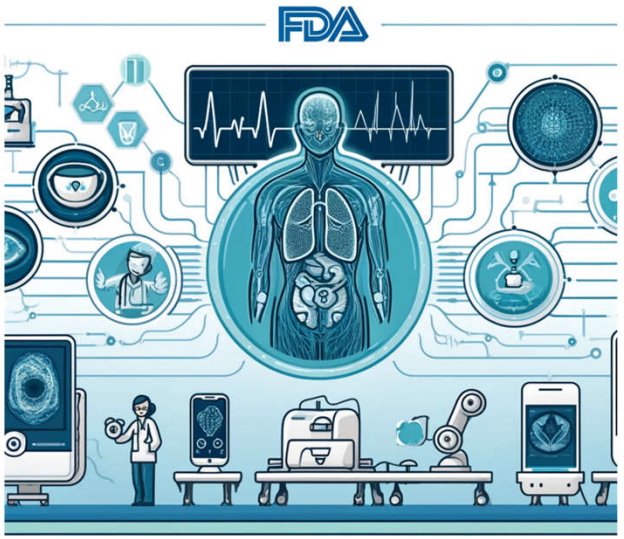


Medical data generated by hospitals and clinics

Predict

- Onset of diseases
- Progression of diseases

300+ FDA-Cleared AI Technologies from 2019 to 2022



“We are drowning in data…”

Law enforcement agencies
collect unknown amounts of data from various sources

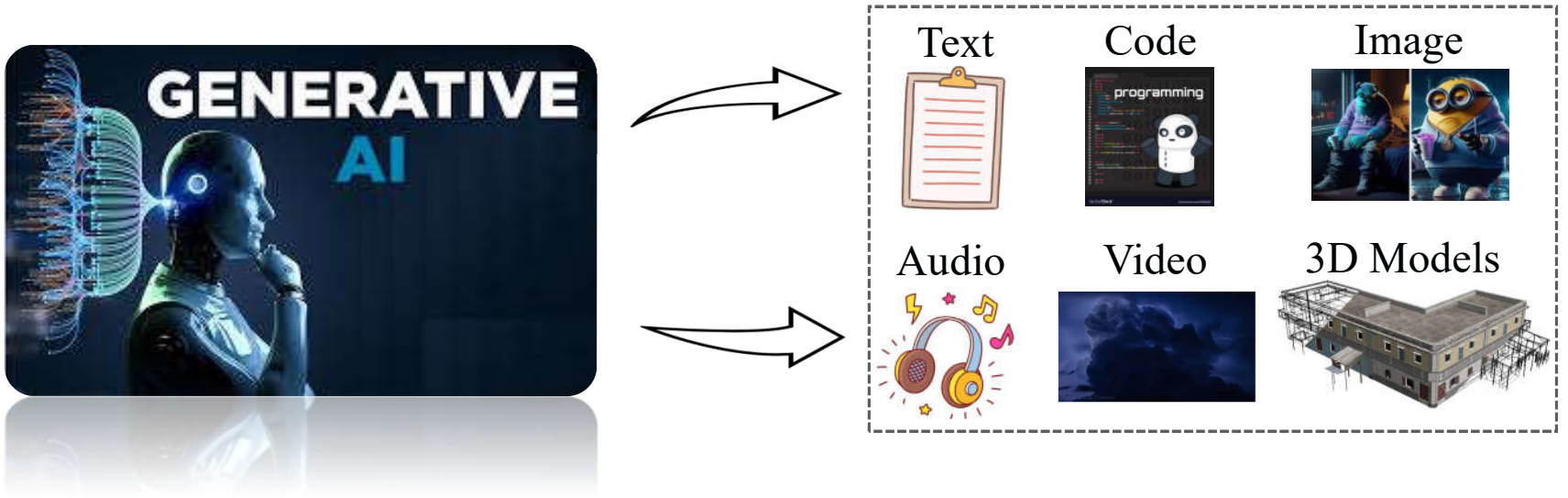
- Cell phone calls
- Location data
- Web browsing behavior
- Credit card transactions
- Online profiles (Facebook)
- …

Predict

- Terrorist or not?
- Social credit

“We are drowning in data…”

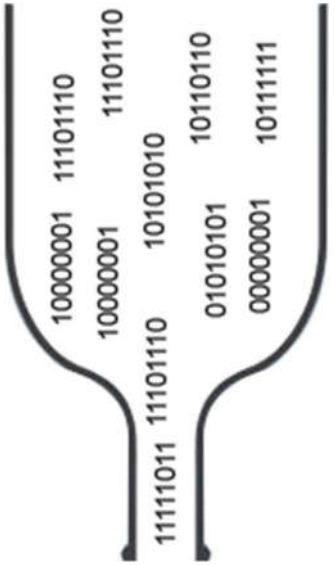
The advent of generative AI accelerates the production of data



Train

- Deep learning models

“We are drowning in data, but starving for knowledge”



← Amount of data that is collected

← Amount of data that can be looked at by humans

We are interested in the patterns, not the data itself!

Data Mining methods help us to

- discover interesting patterns in large quantities of data
- take decisions based on the patterns

The definition of data mining

"Data mining is the process of discovering knowledge or patterns from massive amounts of data."

Rock	Gold	Tools	Miner
			
Data	Knowledge	Algorithm	Analysts

Estimated $300 billion industry around big data analytics.

The definition of data mining

- More definitions

Exploration & analysis, of large quantities of data in order to discover meaningful patterns.

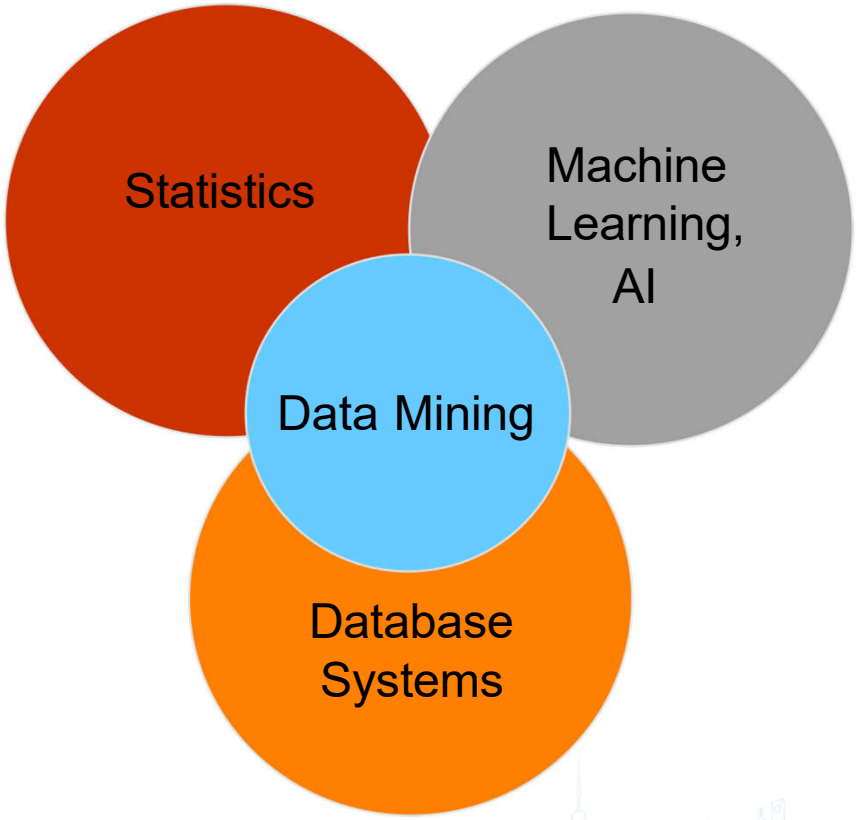
Non-trivial extraction of
- **implicit,**
- **previously unknown, and**
- **potentially useful**

information from data.

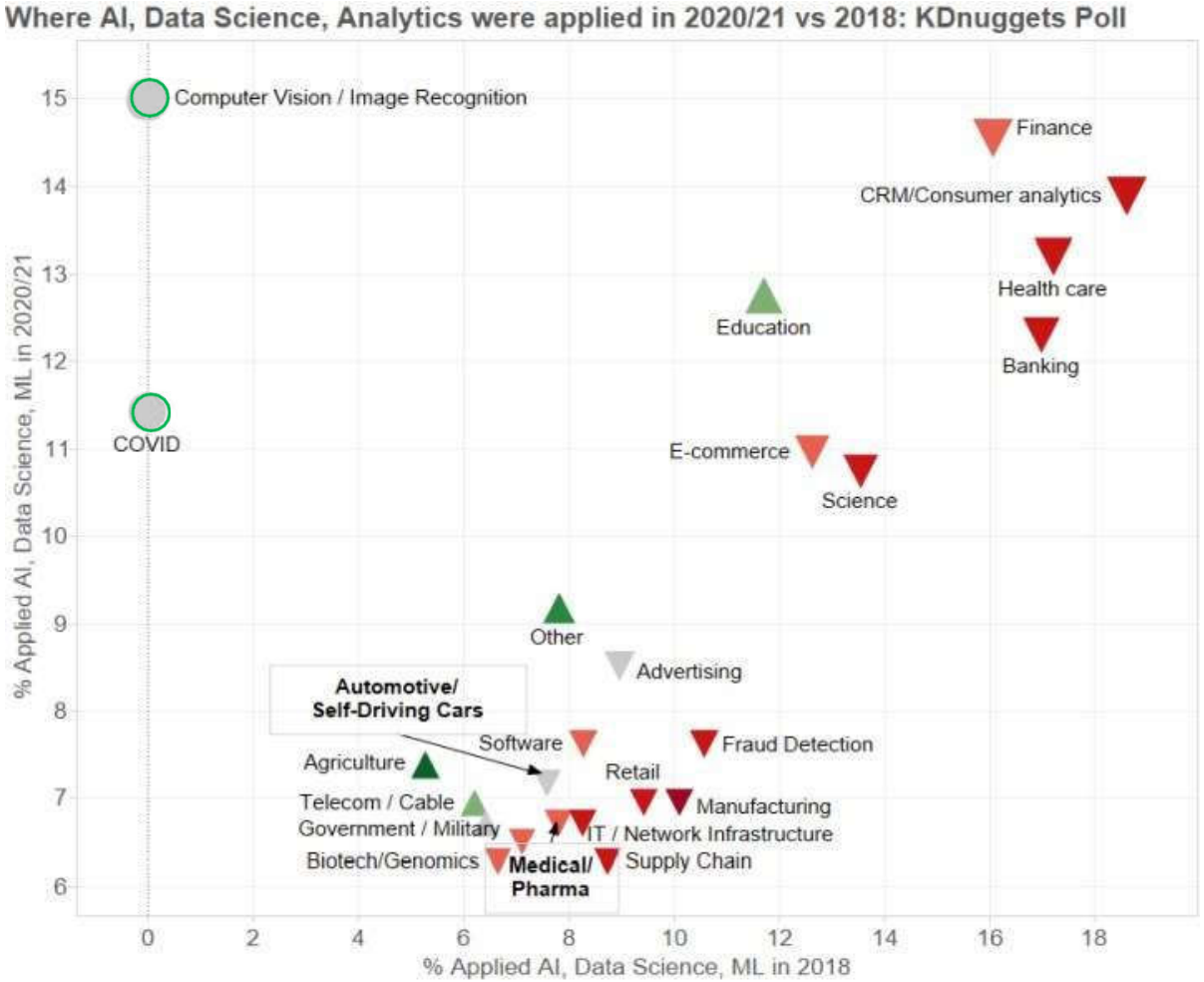
- Data mining methods
  1. detect interesting patterns in large quantities of data
  2. **support** human decision making by providing such patterns
  3. **predict** the outcome of a future observation based on the patterns

The origin of data mining

- Data Mining combines ideas from statistics, machine learning, artificial intelligence, and database systems
- Tries to overcome short-comings of traditional techniques concerning
  - large amount of data
  - high dimensionality of data
  - heterogeneous and complex nature of data
  - explorative analysis beyond hypothesize-and-test paradigm



Statistics of data mining applications



Here are the top application areas with more than 10% share:

- **Computer Vision / Image Recognition**, 15.0%
- **Finance**, 14.5%
- **CRM/Consumer analytics**, 13.9%
- **Health care**, 13.2%
- **Education**, 12.8%
- **Banking**, 12.3%
- **COVID**, 11.4%
- **E-commerce**, 11.0%
- **Science**, 10.7%

Source: KDnuggets online poll, 447 (2021) and 435 (2018) participants
https://www.kdnuggets.com/2021/06/poll-where-analytics-data-science-ml-applied.html

What is data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Tabular

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Size: Number of objects
Dimensionality: Number of attributes
Sparsity: Number of populated object-attribute pairs

Types of attributes

There are different types of attributes

- **Categorical**
  - Examples: eye color, zip codes, words, rankings (e.g., good, fair, bad), height in {tall, medium, short}
  - Nominal (no order or comparison, e.g., ID numbers, eye color, zip codes) vs Ordinal (order but not comparable, e.g., taste of potato chips on a scale from 1-10, grades, height {tall, medium, short})

- **Numerical**
  - Examples: dates, temperature, time, length, value, count
  - Discrete (counts) vs Continuous (temperature)
  - Special case: Binary attributes (yes/no, exists/not exists)

Example: categorical + numerical

Attributes

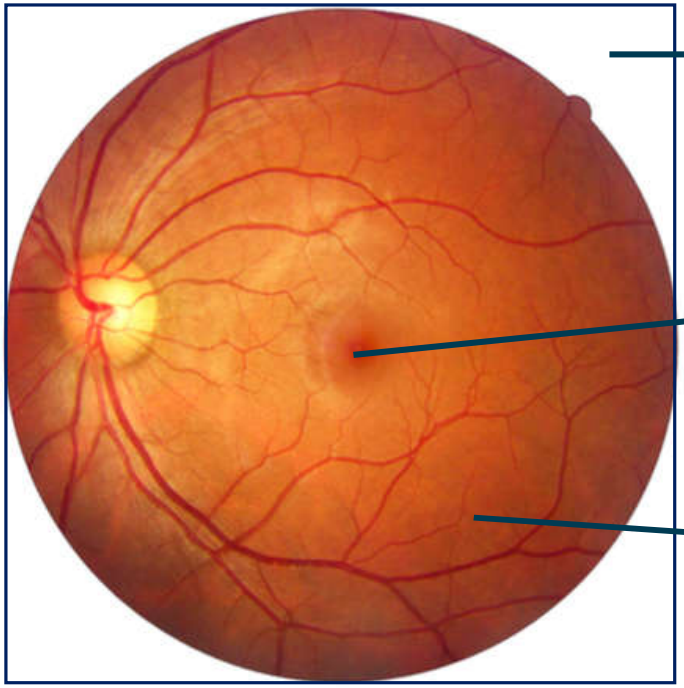
Age	Gender	Blood Pressure (mmHg)	Cholesterol (mg/dL)	Smoking Status	Exercise Frequency	Disease Progression	Outcome
58	Male	141	277	Never	High	Severe	Negative
48	Female	80	259	Never	Low	None	Positive
34	Male	106	231	Never	Moderate	Severe	Negative
62	Female	141	203	Current	High	Moderate	Positive
27	Female	156	217	Former	None	None	Negative

Labels

- **Age** (Numerical): Age of the individuals.
- **Gender** (Nominal): Male or Female.
- **Blood Pressure (mmHg)** (Numerical): Blood pressure measurement in millimeters of mercury.
- **Cholesterol (mg/dL)** (Numerical): Cholesterol level in milligrams per deciliter.
- **Smoking Status** (Nominal): Categories include 'Never', 'Former', and 'Current'.
- **Exercise Frequency** (Ordinal): Ranked as 'None', 'Low', 'Moderate', and 'High'.
- **Disease Progression** (Ordinal): Stages include 'None', 'Mild', 'Moderate', and 'Severe'.
- **Outcome** (Nominal): Disease outcome, either 'Negative' or 'Positive'.

Example: numerical

An image can be represented as a matrix composed of numerical pixel values, displaying the intensity levels ranging from 0 (black) to 255 (white).



(255, 255, 255)

(218, 110, 47)

(199, 81, 34)

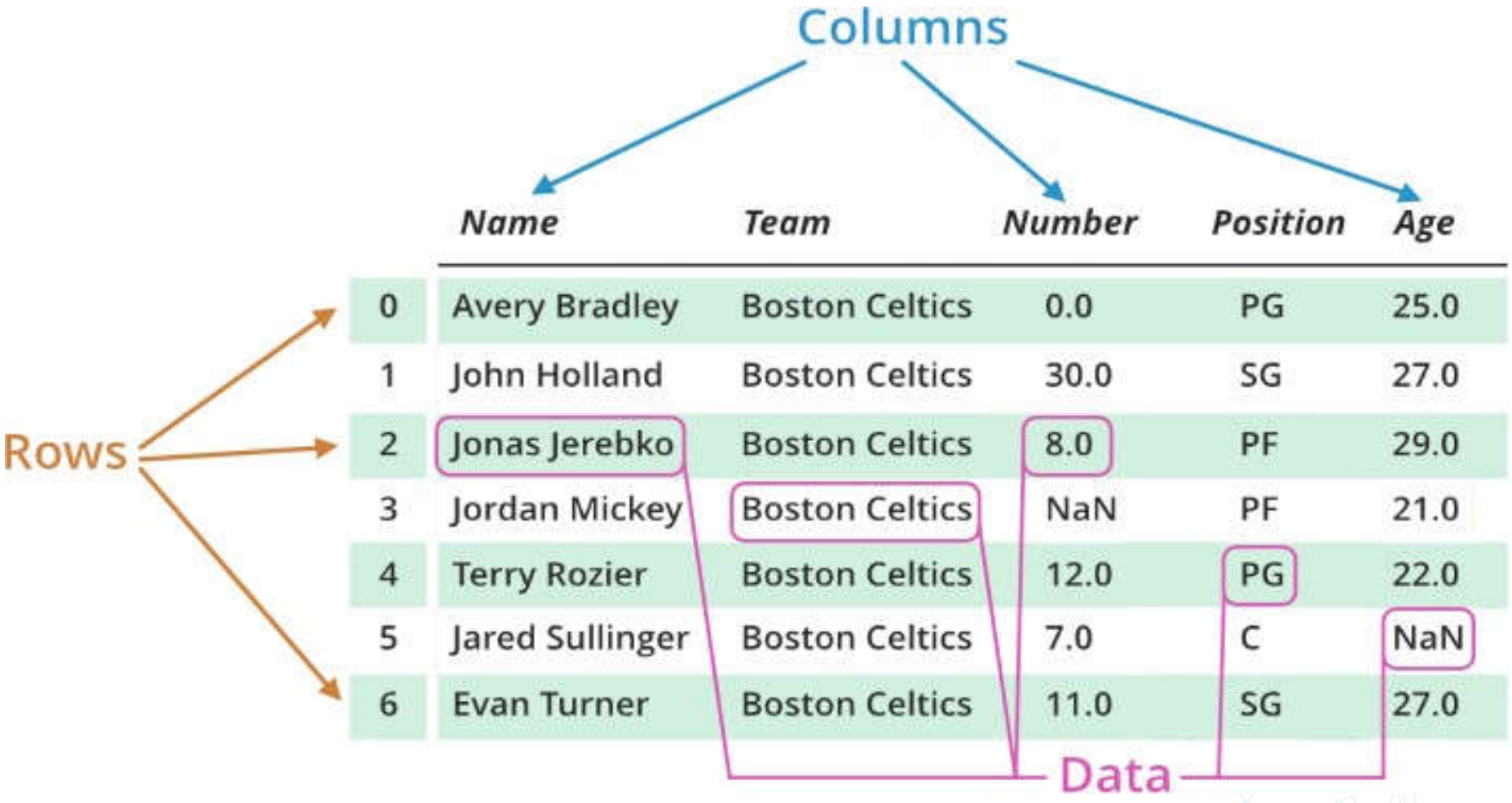
Color Fundus Image

3-channel pixel values

Outline: Introduction to Data Mining

1. What is Data Mining?
2. Data Types
3. Data Mining Tasks
4. Data Mining Process
5. Evaluation Setting and Metrics

Data type – tabular data



	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

Data type – transaction data

- Each record (transaction) is a set of items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

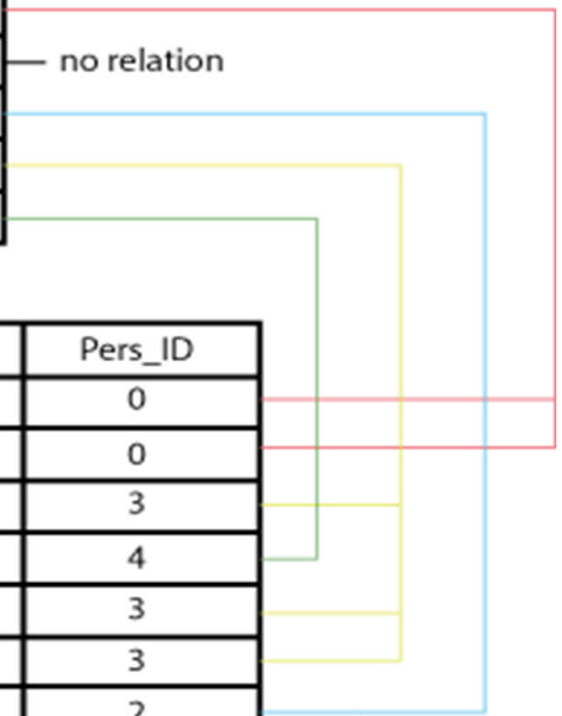
- A set of items can also be represented as a binary vector, where each attribute is an item.

Data type – relational data

- Relational tables.

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom



Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Data type – text data

- Texts in various domains and languages.

News

Social Media

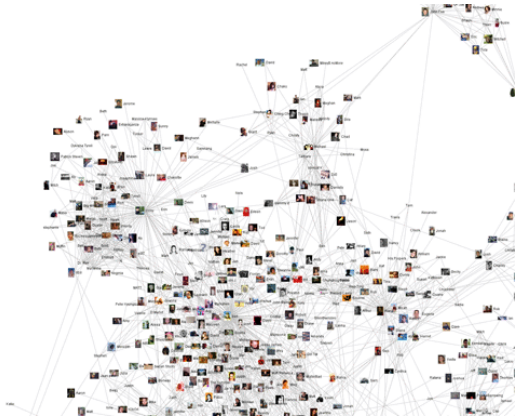
Business & Finance

Scientific Papers

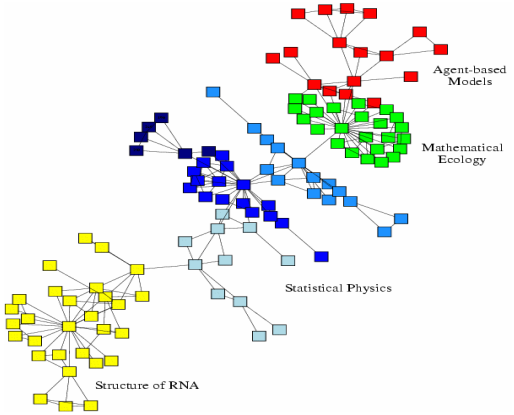
Medical Records

.....

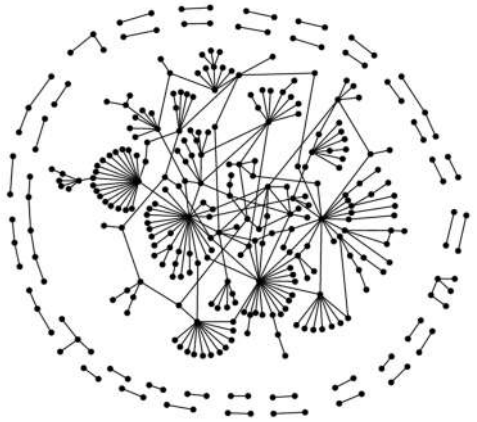
Data type – graph or network data



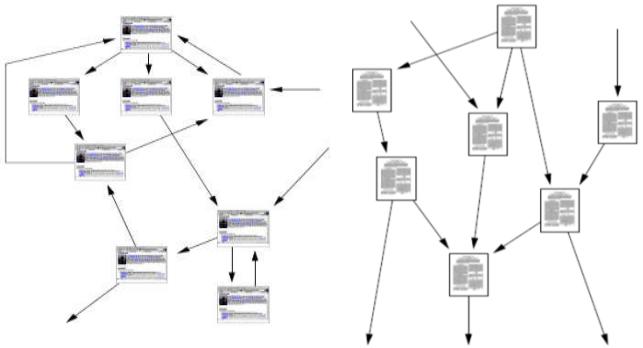
Social Networks



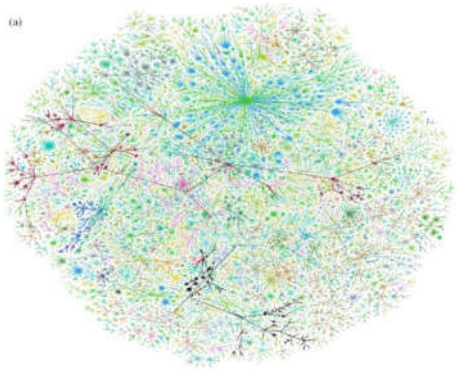
Economic Networks



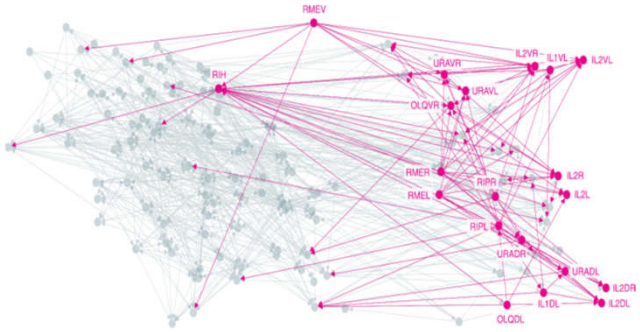
Biomedical Networks



Information Networks: Web & Citations



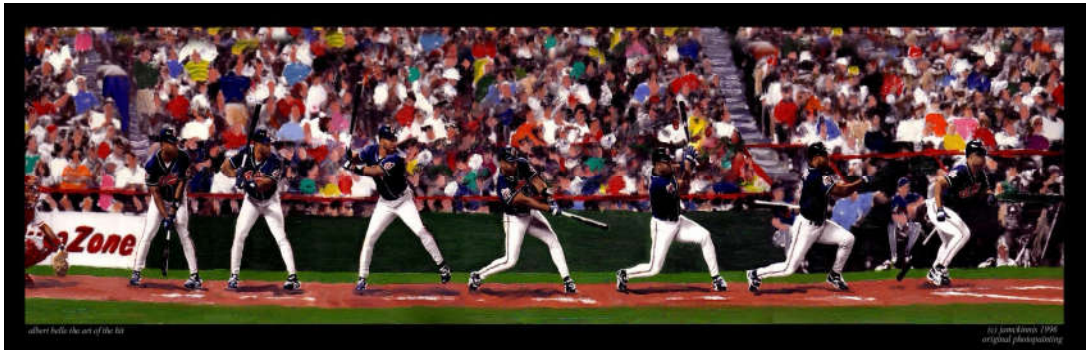
Internet



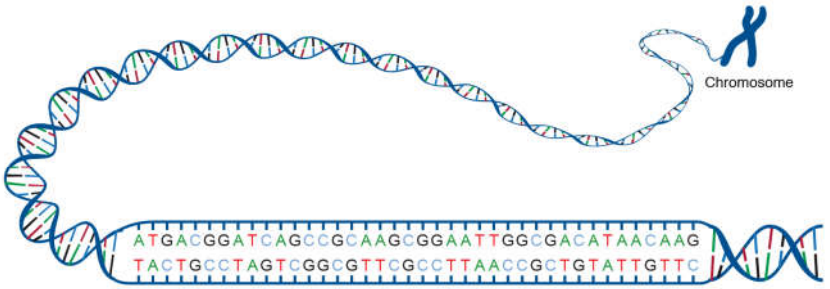
Network of Neurons

Data type – sequential data

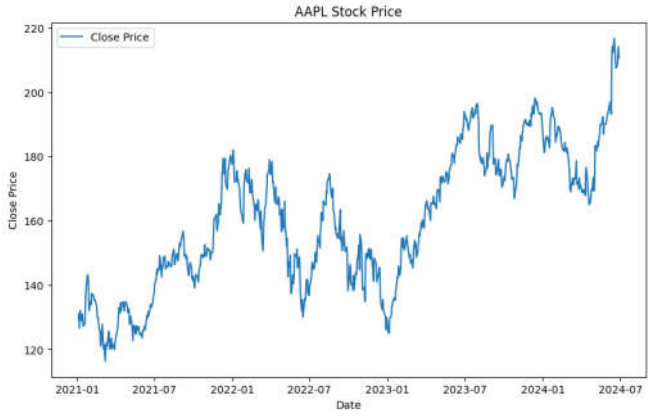
- Video
  - sequence of images



- Genetic sequence
  - sequence of nucleotides or bases

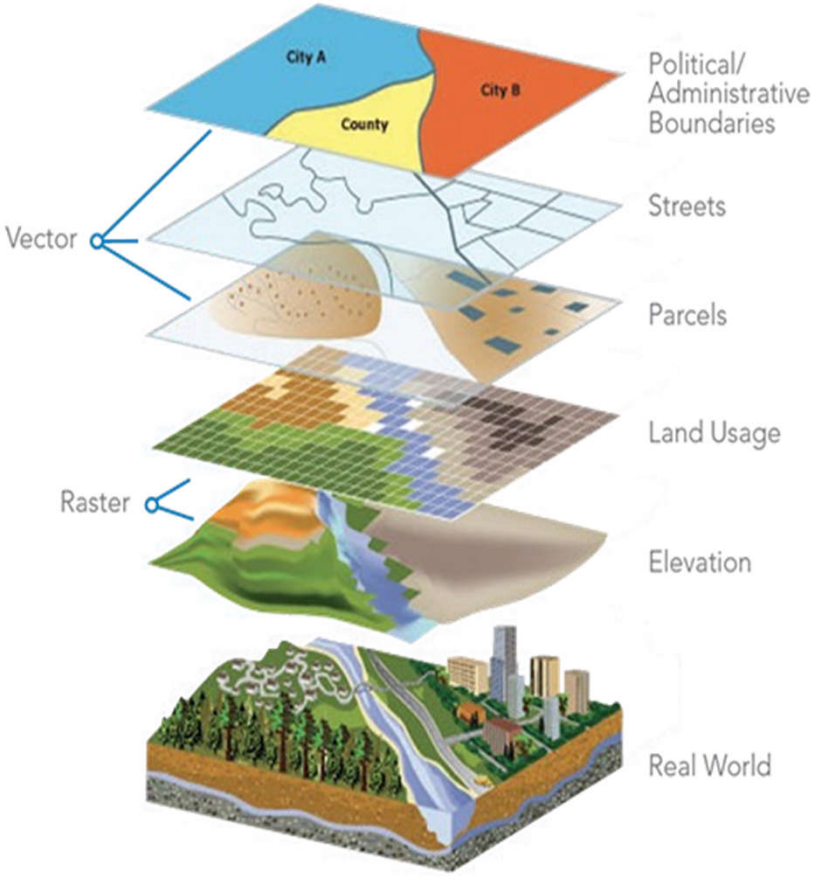


- Time-series data

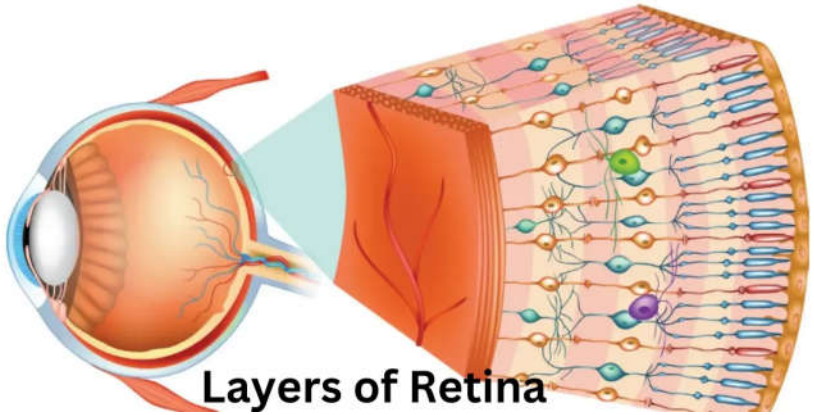


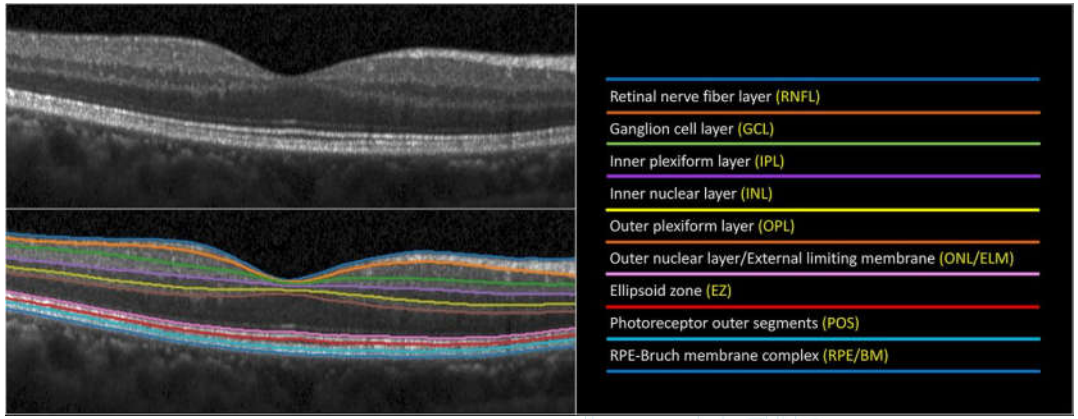
Data type – spatial / layer data

- Maps



- OCT b-scans

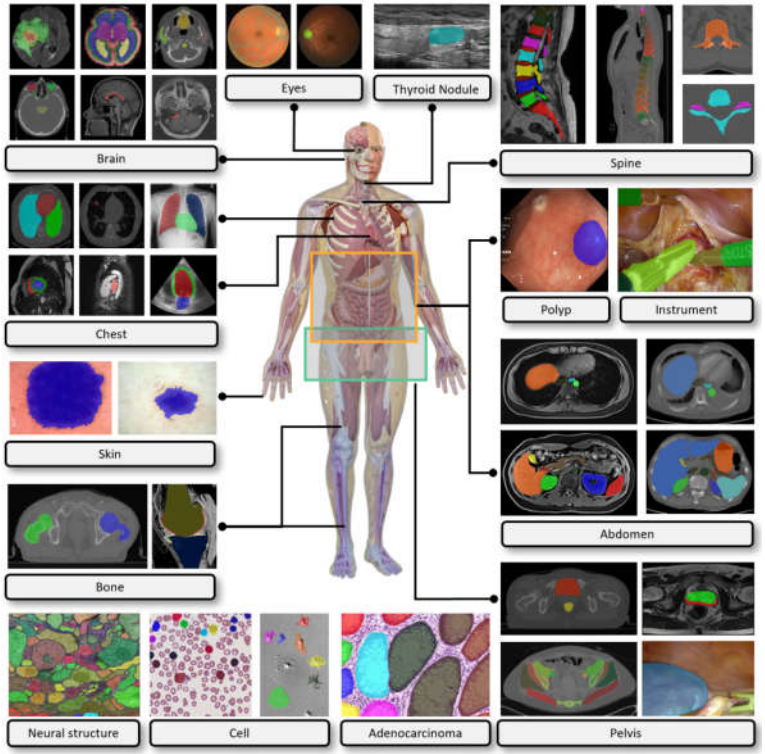




Data type – image data

- Natural images

- Medical images

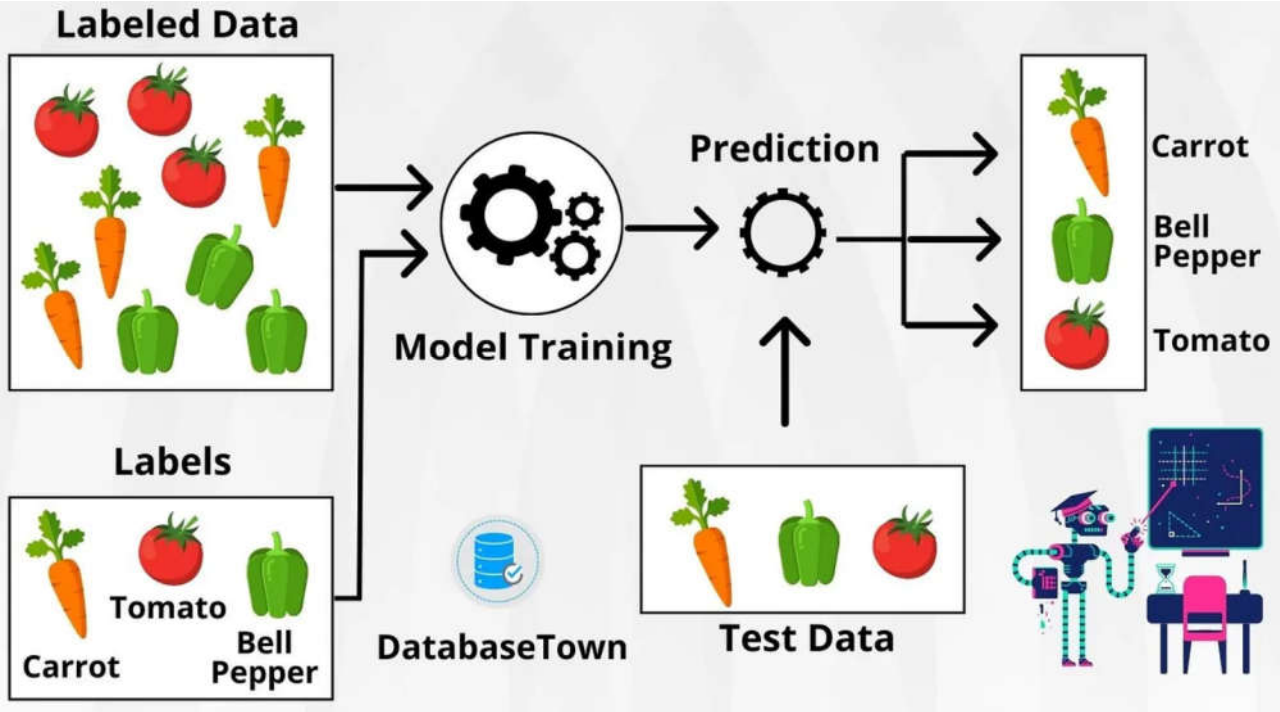


Outline: Introduction to Data Mining

1. What is Data Mining?
2. Data Types
3. Data Mining Tasks
4. Data Mining Process
5. Evaluation Setting and Metrics

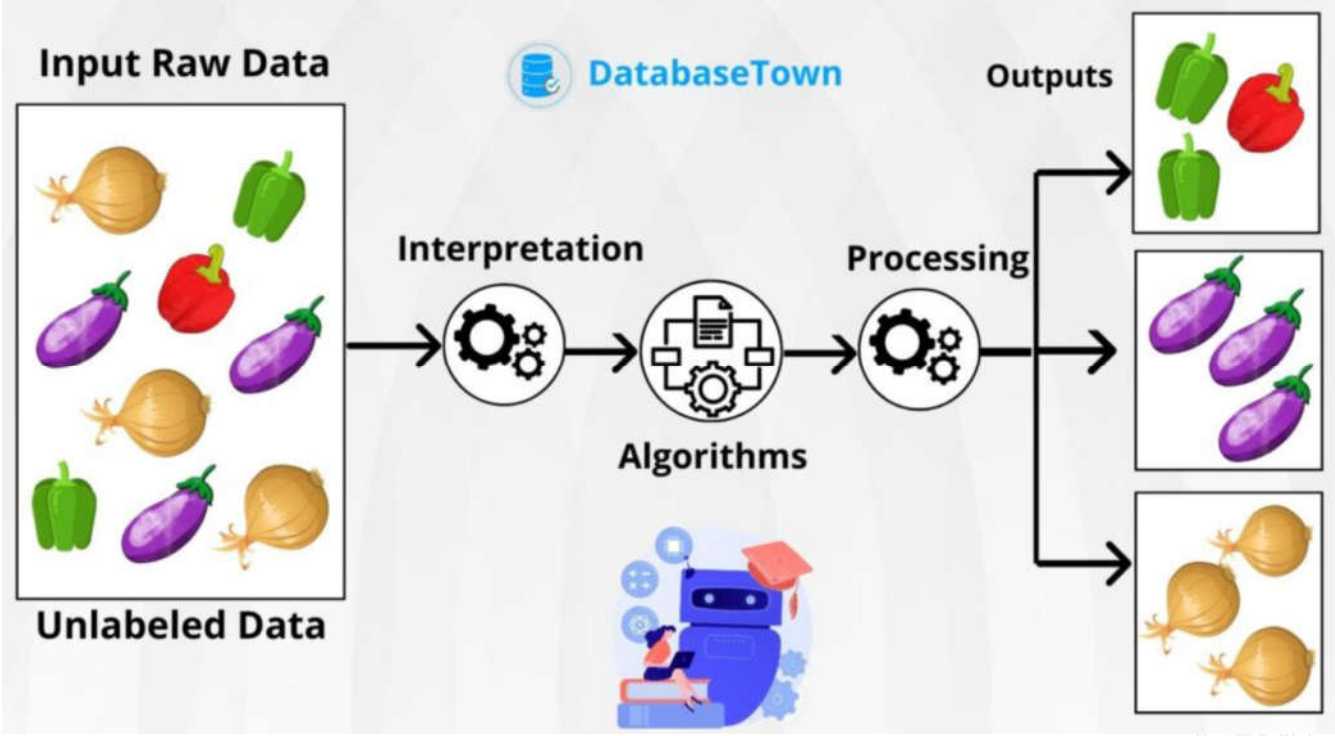
Supervised vs. unsupervised

- Supervised tasks have one or more target variables
  - Learning with **data + labels**
  - Classification, regression, segmentation, …

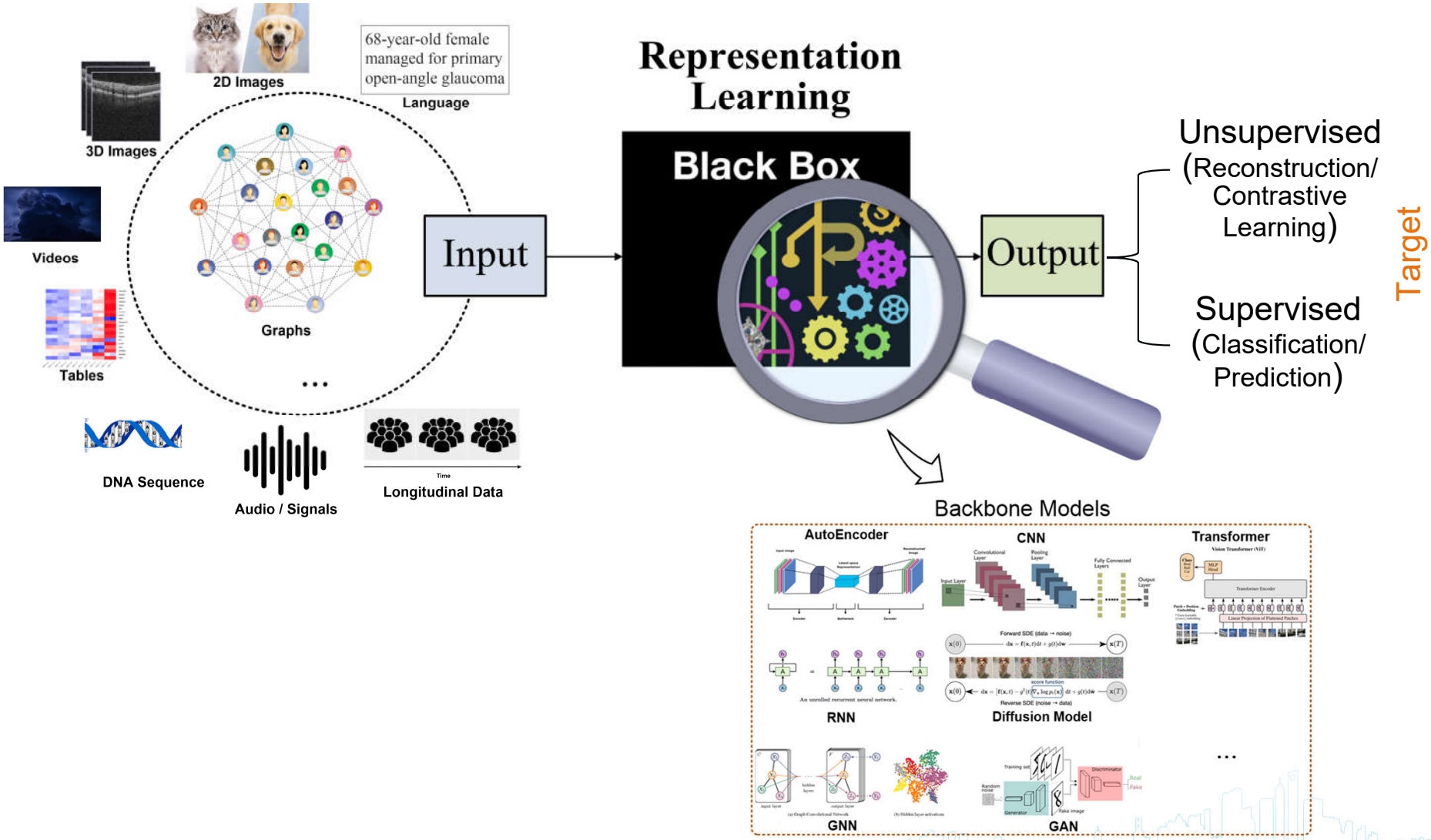


Supervised vs. unsupervised

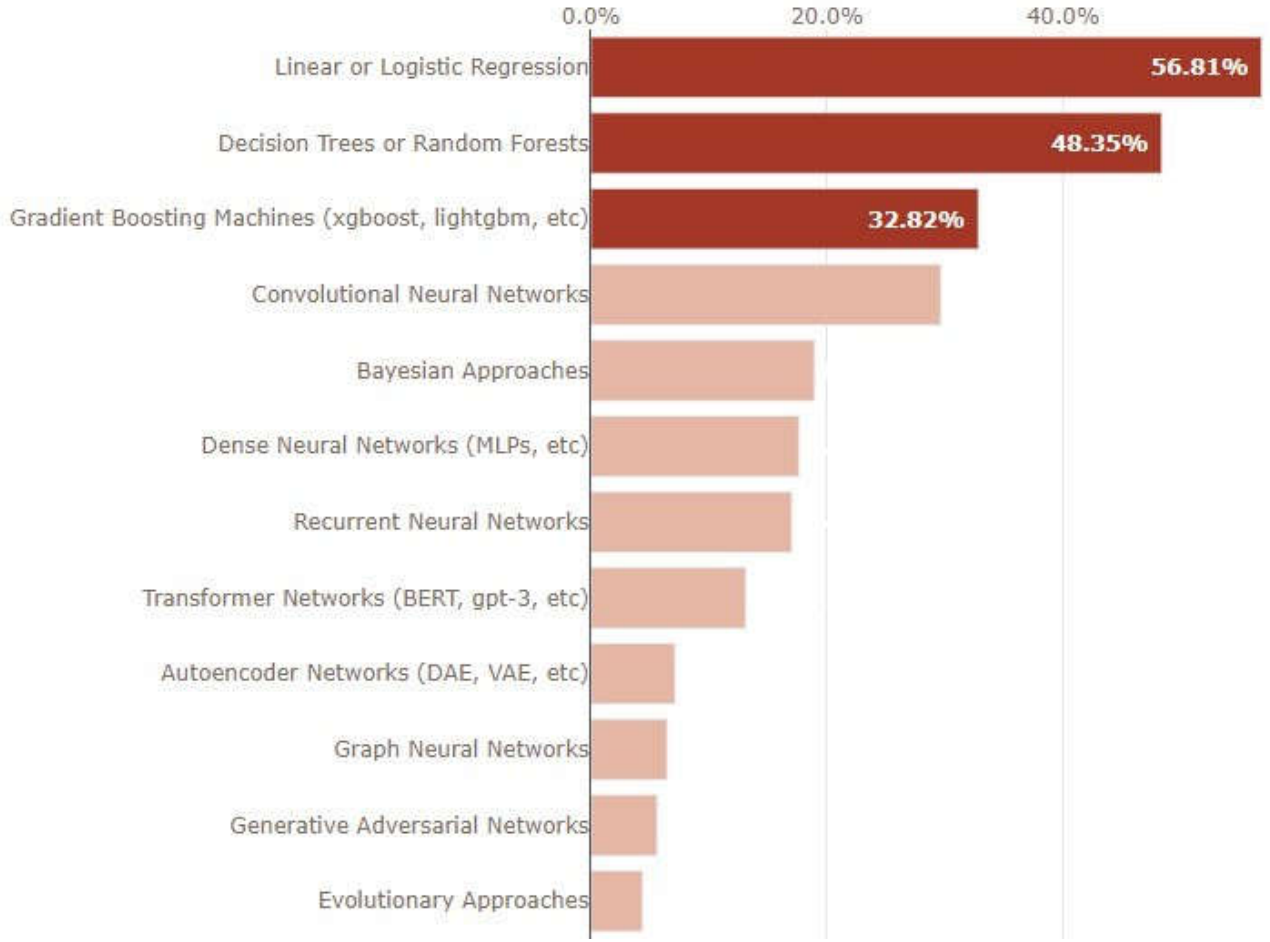
- Unsupervised tasks do not have any predefined outputs or targets
  - Learning with **data only**
  - Clustering, frequent patterns, dimension reduction, …



Supervised vs. unsupervised

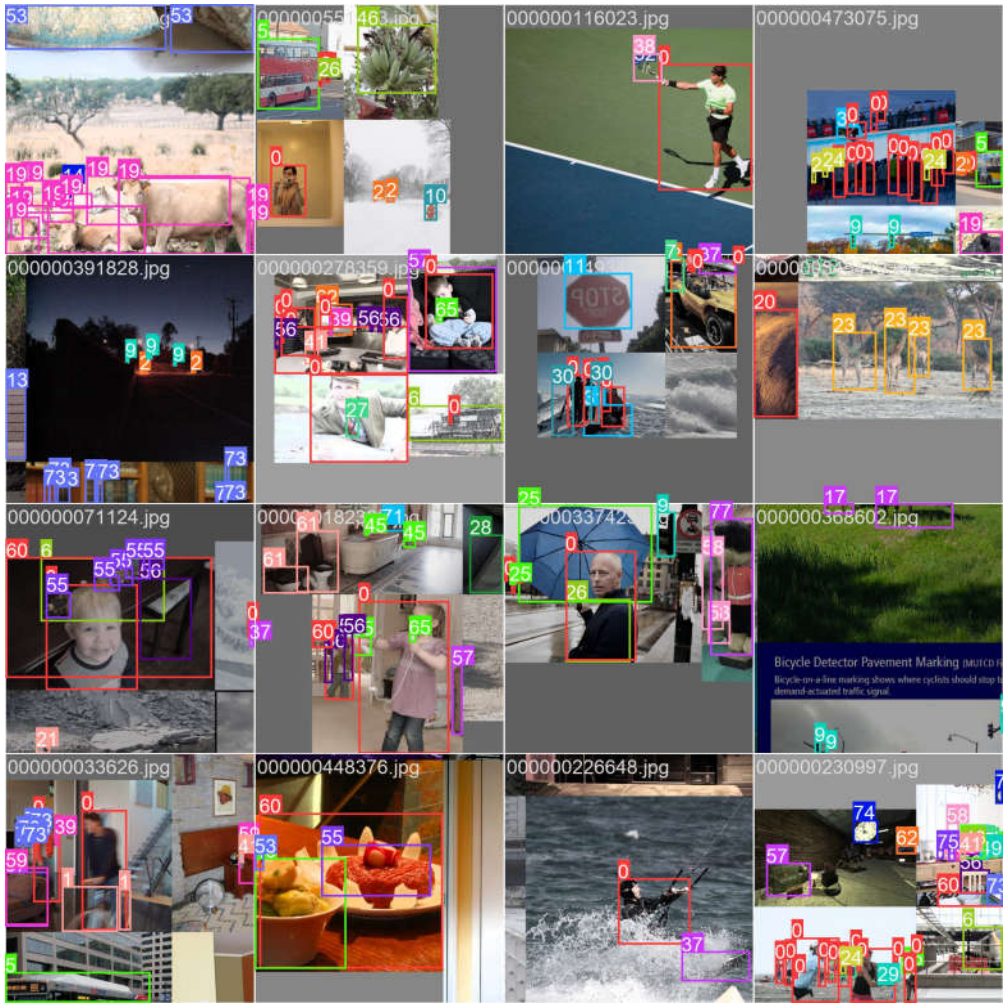


Which methods are used in practice?

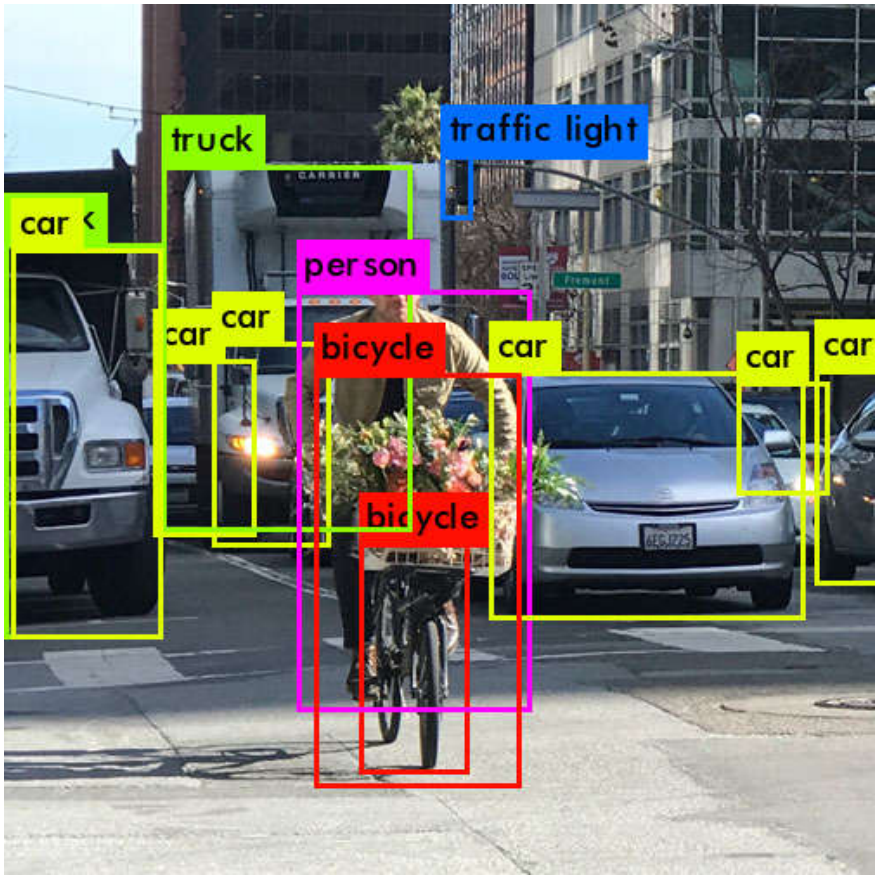


Source: Kaggle online poll 2022, 23,997 respondents,
https://www.kaggle.com/code/eraikako/data-science-and-mlops-landscape-in-industry

Supervised example: object detection

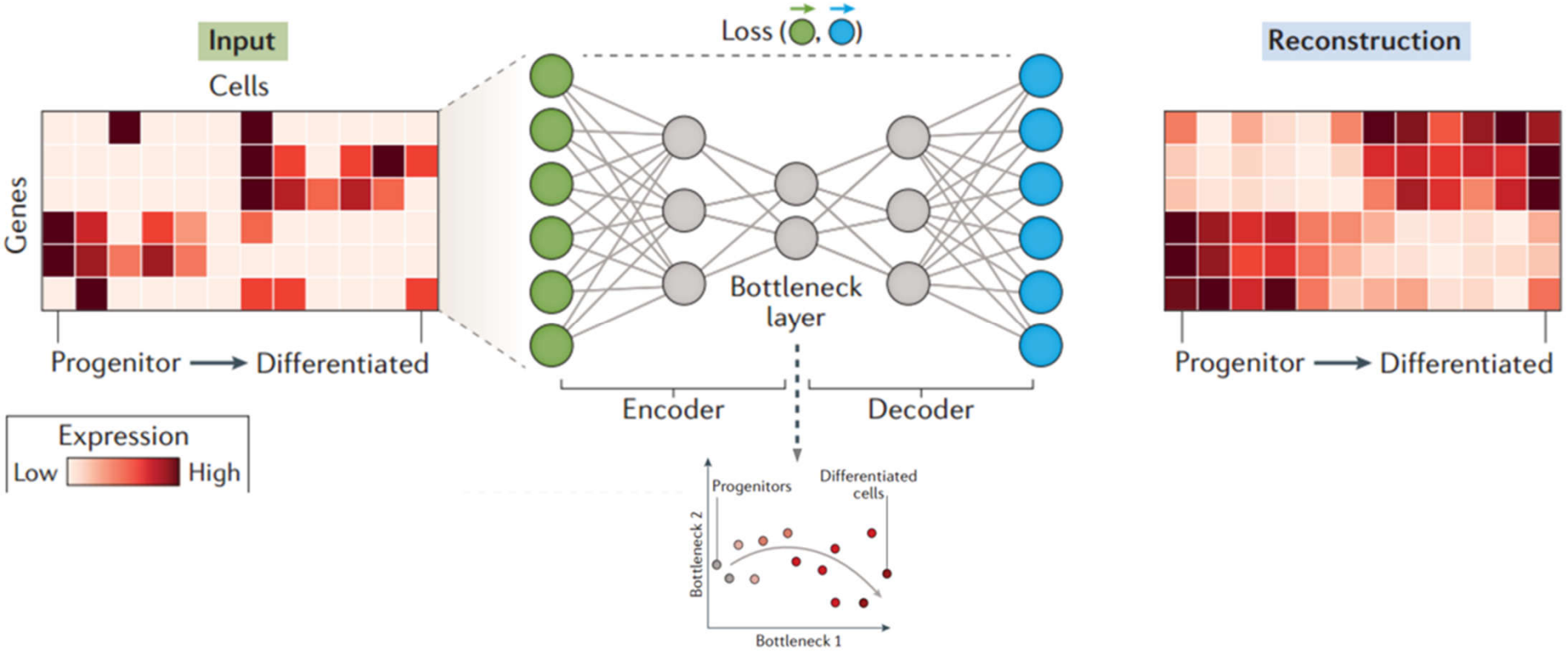


The COCO dataset contains 200K images with annotations



Real-world object detection

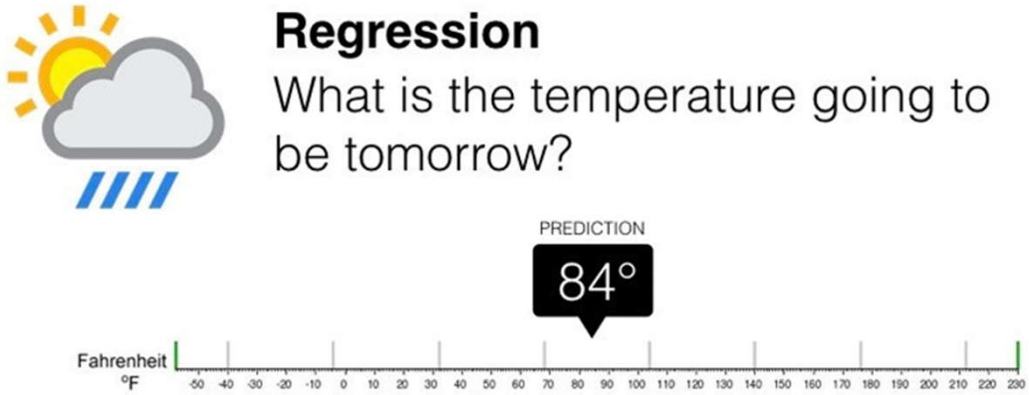
Unsupervised example: dimension reduction

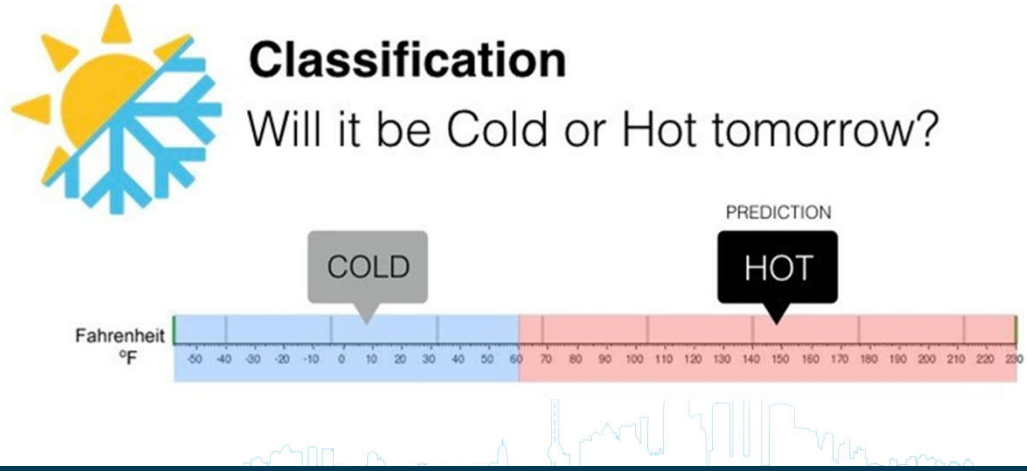


The low-dimensional representation revealing the cell differentiation process

Classification vs. regression

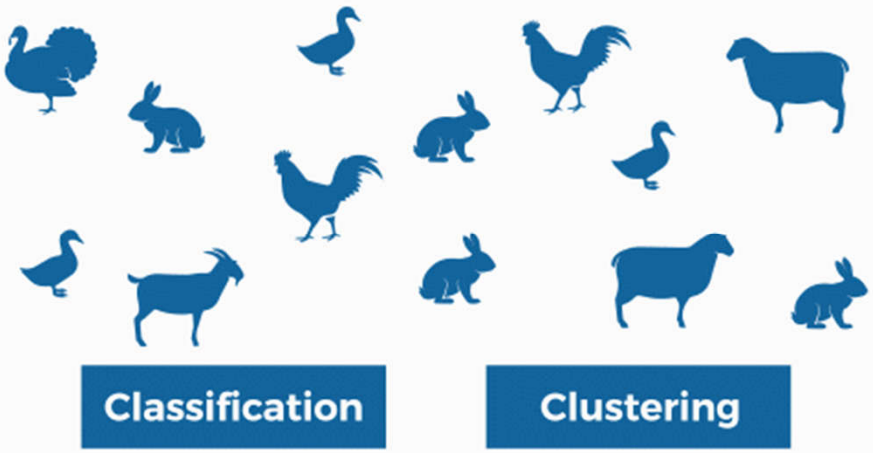
- Classification and Regression are both supervised tasks
  - The target variables in classification are **discrete**
  - The target variables in regression are **continuous**





Classification vs. clustering

- Classification is to accurately predict the target class for each instance in the data.
- Clustering is to discover the inherent groupings in the data.



Example: classification

Find a model for predicting credit worthiness

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

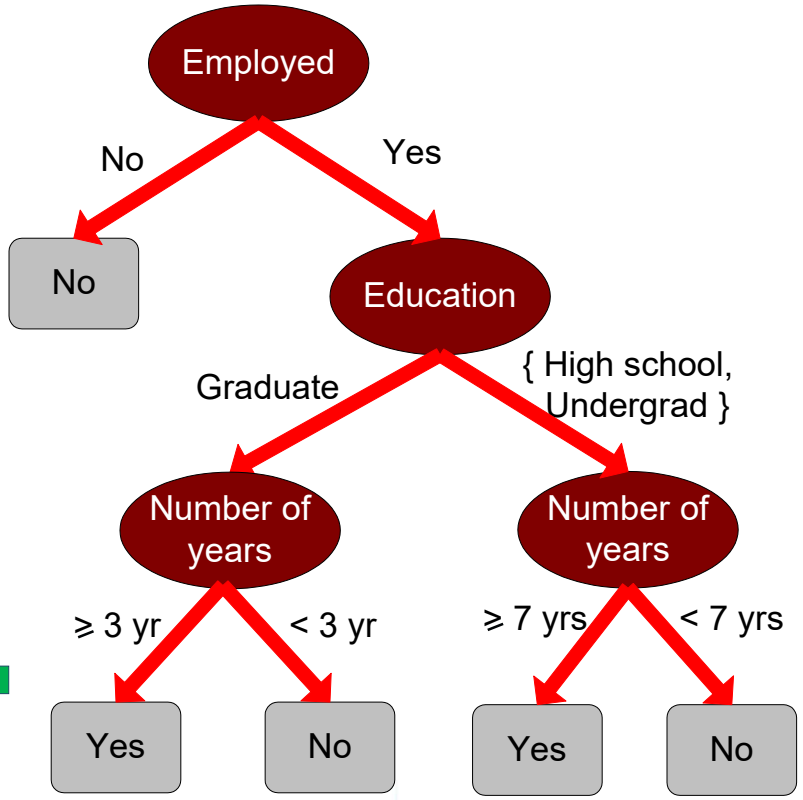
Training set

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

Testing set

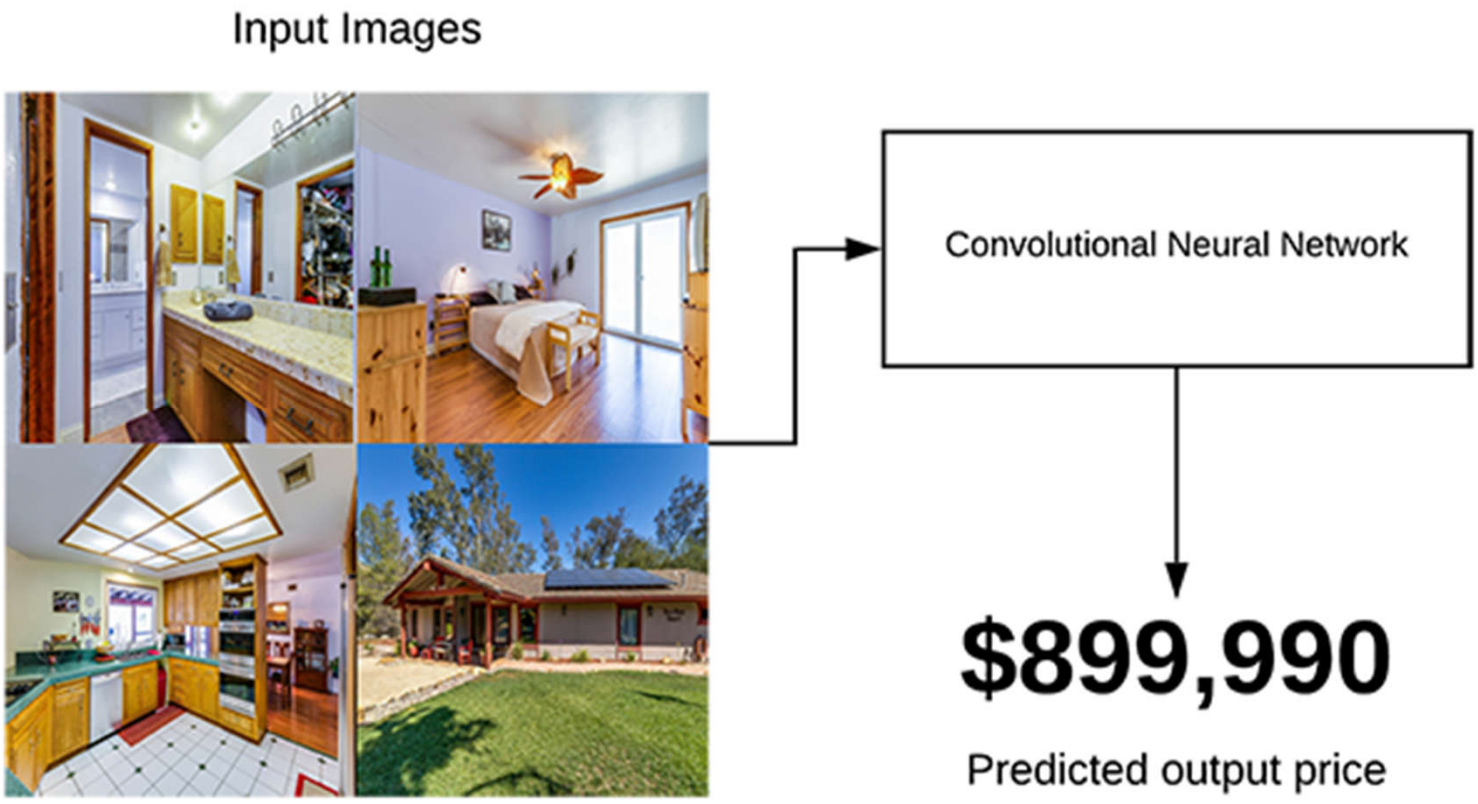
Train

Test



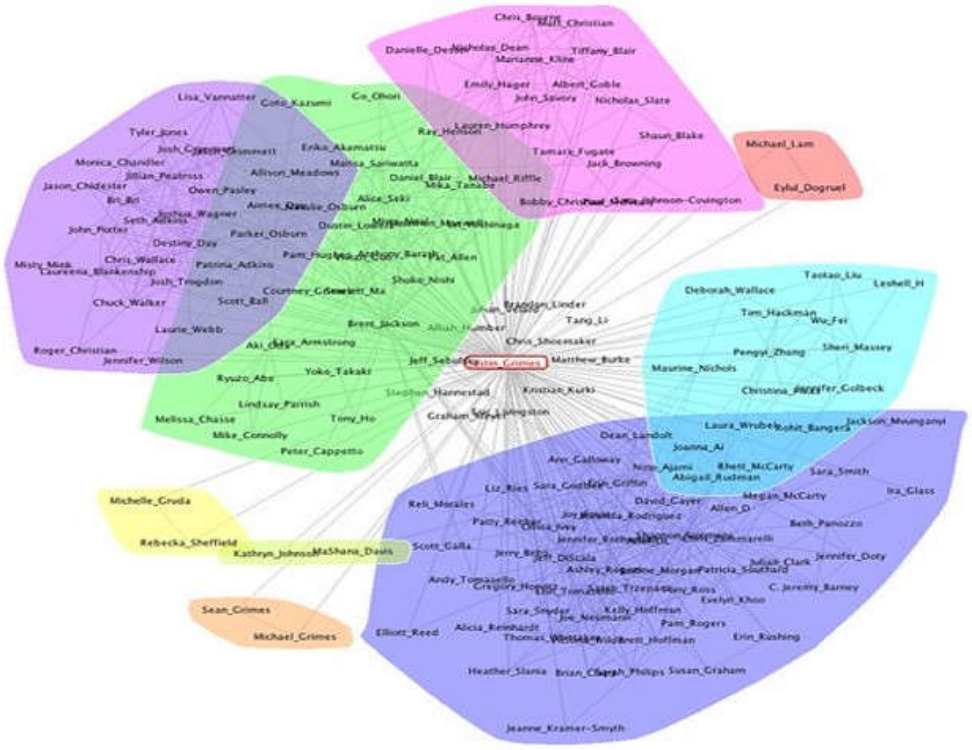
Example: regression

Find a model for predicting property value

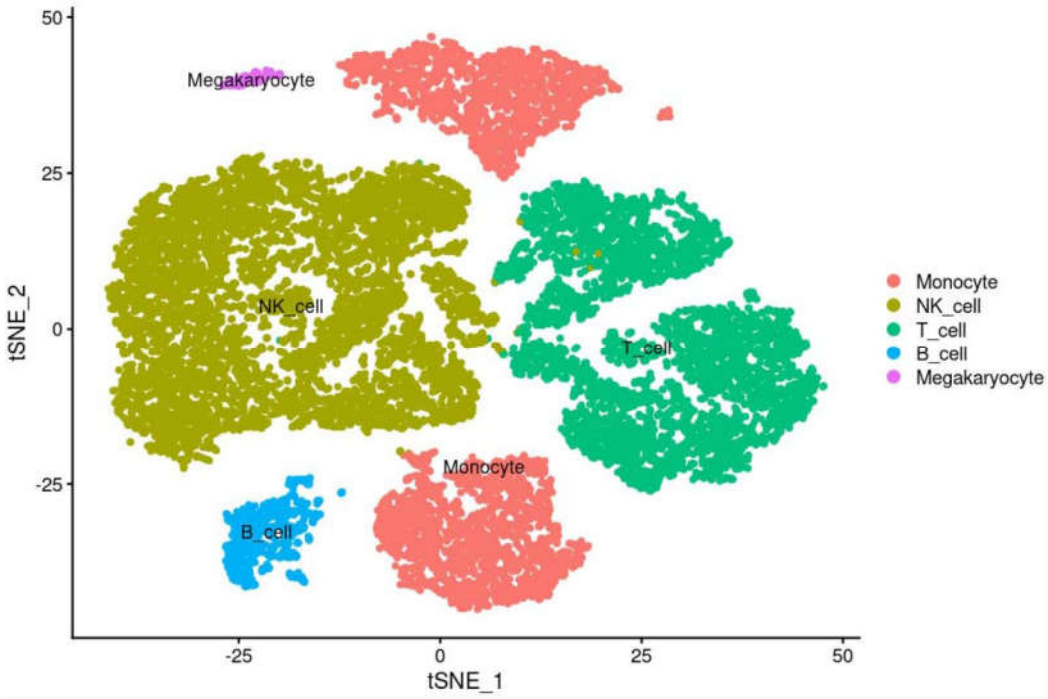


Example: clustering

Find a model for grouping samples with similar properties or functions



Social network community detection



Cell clustering

Association analysis

- Given a set of records each of which contain some number of items from a given collection
- Discover frequent itemsets and produce association rules which will predict occurrence of an item based on occurrences of other items

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Frequent Itemsets
{Diaper, Milk, Beer}
{Milk, Coke}

Association Rules
{Diaper, Milk} --> {Beer}
{Milk} --> {Coke}

Example: association analysis

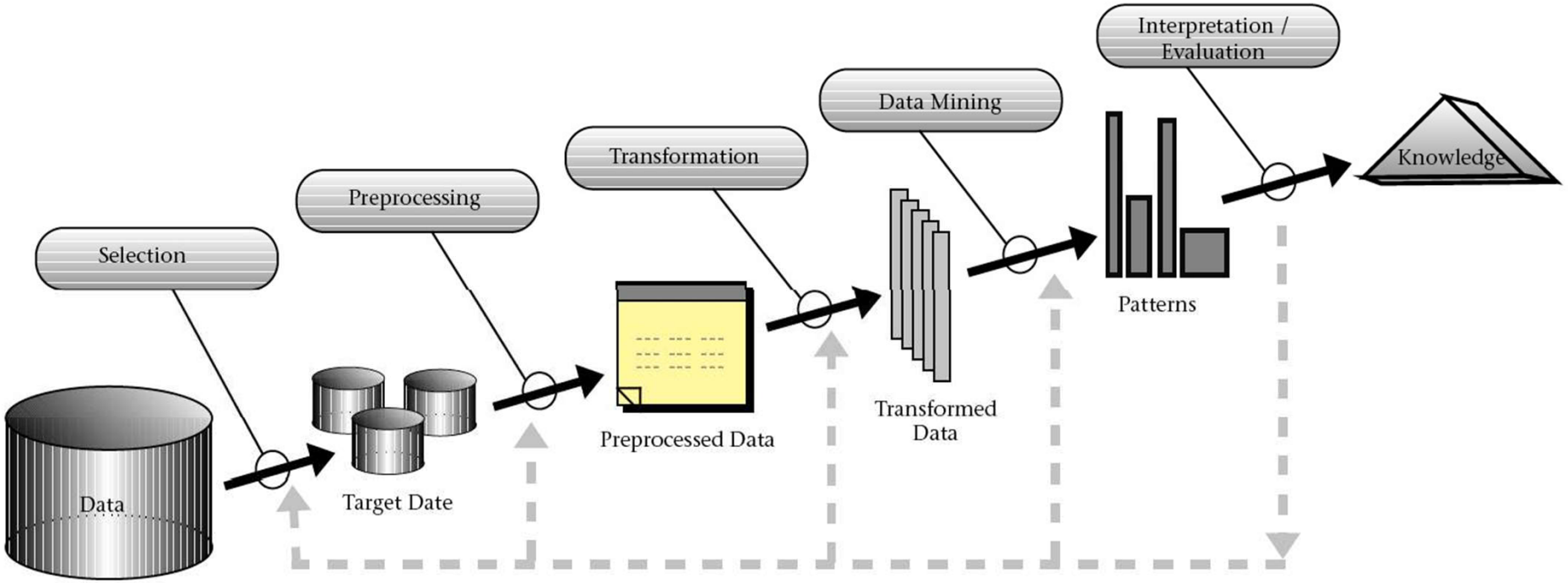
- Application area: supermarket shelf management
  - Goal: To identify items that are bought together by sufficiently many customers
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items
  - A classic rule and its implications:
    - if a customer buys **diapers** and **milk**, he is likely to buy beer too
    - so, don't be surprised if you find six-packs stacked next to diapers!
    - promote diapers to boost beer sales
    - if selling diapers is discontinued, this will affect beer sales as well

- Application area: sales promotion

Outline: Introduction to Data Mining

1. What is Data Mining?
2. Data Types
3. Data Mining Tasks
4. Data Mining Process
5. Evaluation Setting and Metrics

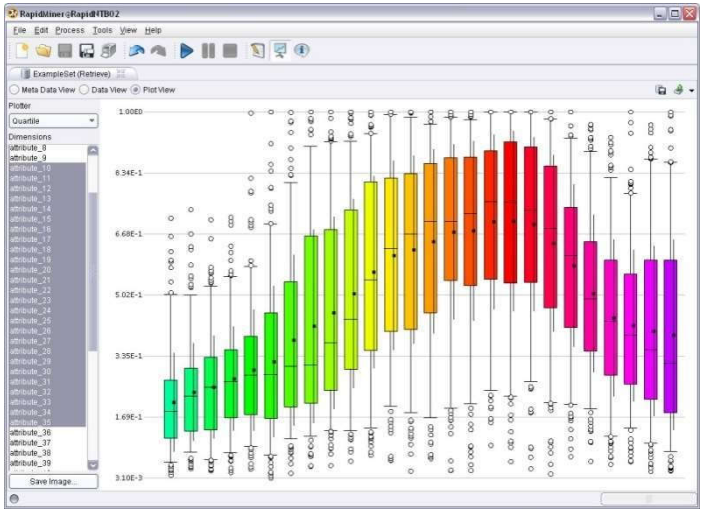
The data mining process

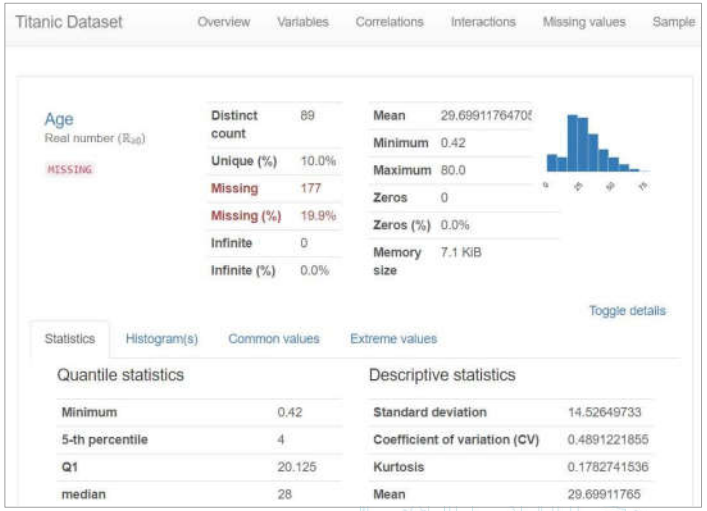


Source: Fayyad et al. (1996)

Selection and exploration

- Selection
  - What data is potentially useful for the task at hand?
  - What data is available?
  - What do I know about the quality of the data?
- Exploration / Profiling
  - Get an initial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records





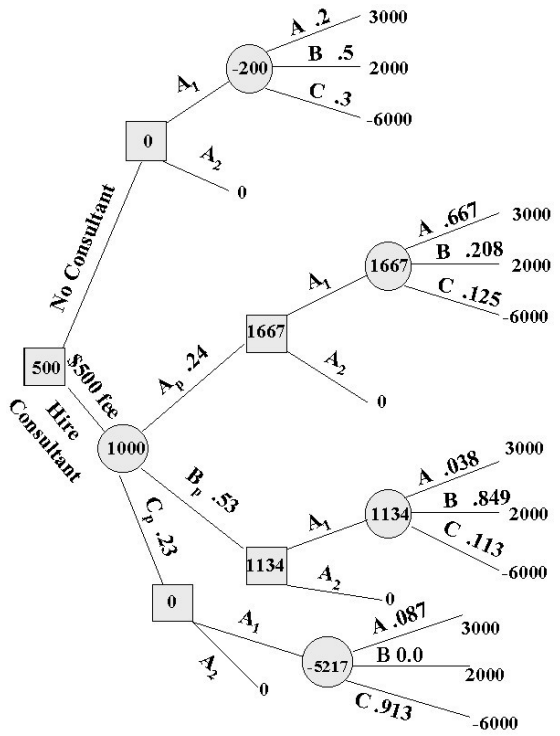
Preprocessing and transformation

- Transform data into a representation that is suitable for the chosen data mining methods
  - scales of attributes (nominal, ordinal, numeric)
  - number of dimensions (represent relevant information using less attributes)
  - amount of data (determines hardware requirements)
- Methods
  - discretization and binarization
  - feature subset selection / dimensionality reduction
  - attribute transformation / text to term vector / embeddings
  - aggregation, sampling
  - integrate data from multiple sources
- Good data preparation is key to producing valid and reliable models
- Data integration and preparation is estimated to take 70-80% of the time and effort of a data mining project

Data mining

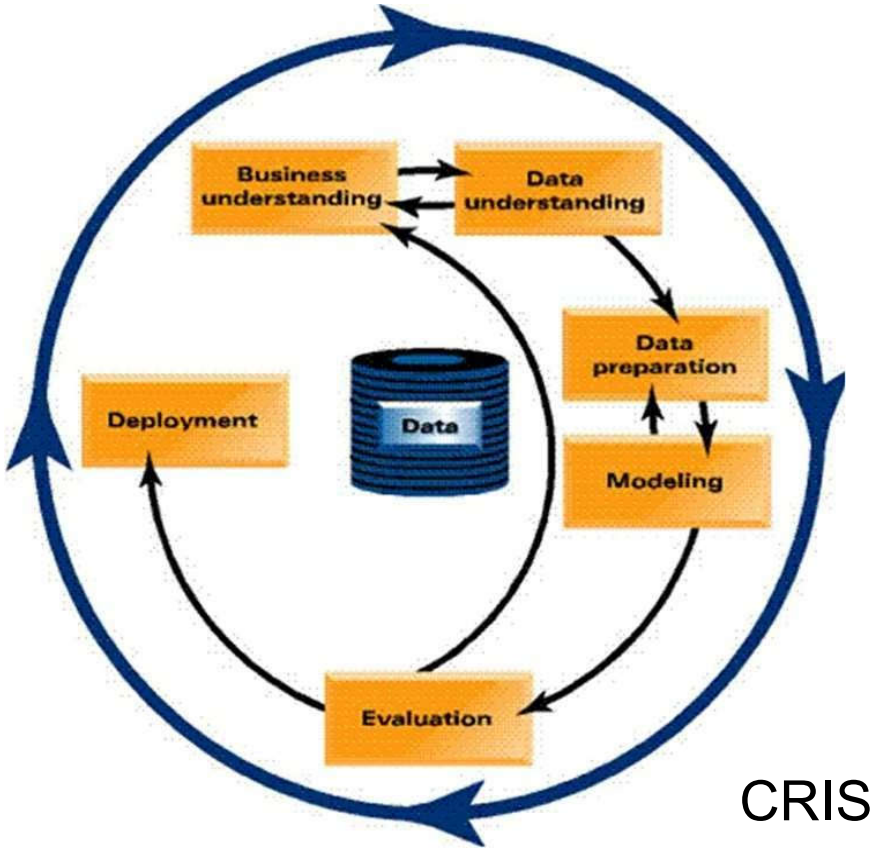
- Input: Preprocessed Data
- Output: Model / Patterns

1. Apply data mining method
2. Evaluate resulting model / patterns
3. **Iterate**
   - experiment with different hyperparameter settings
   - experiment with multiple alternative methods
   - improve preprocessing and feature generation
   - increase amount or quality of training data



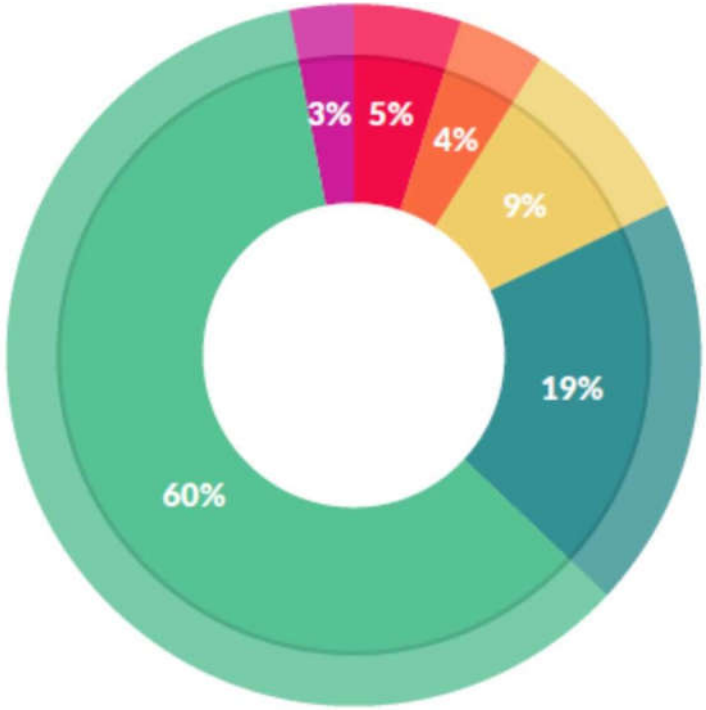
Deployment

- Use model in the business context
- Keep iterating in order to maintain and improve model



CRISP-DM Process Model

How do data scientists spend their days?



What data scientists spend the most time doing

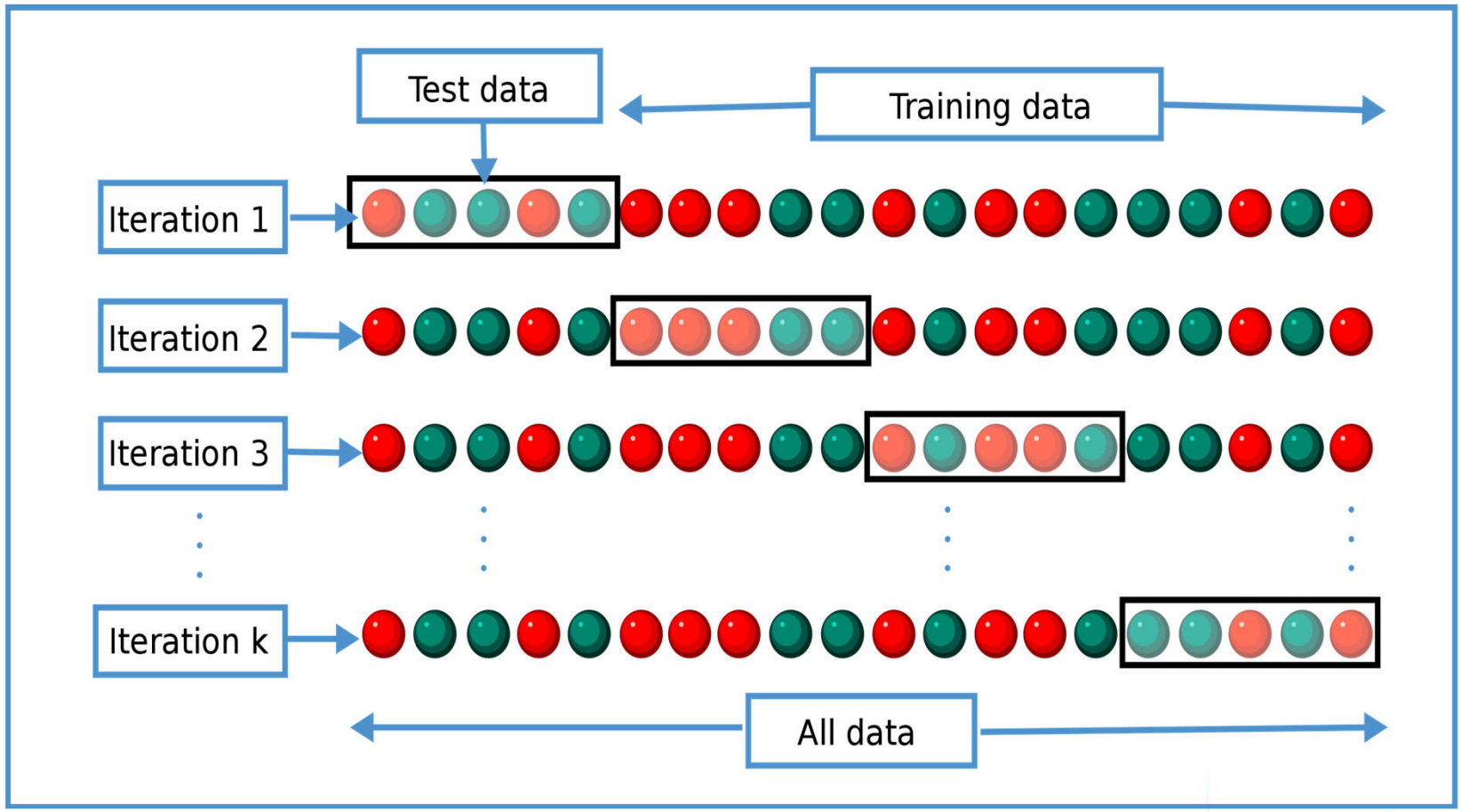
- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: CrowdFlower Data Science Report 2016: http://visit.crowdflower.com/data-science-report.html

Outline: Introduction to Data Mining

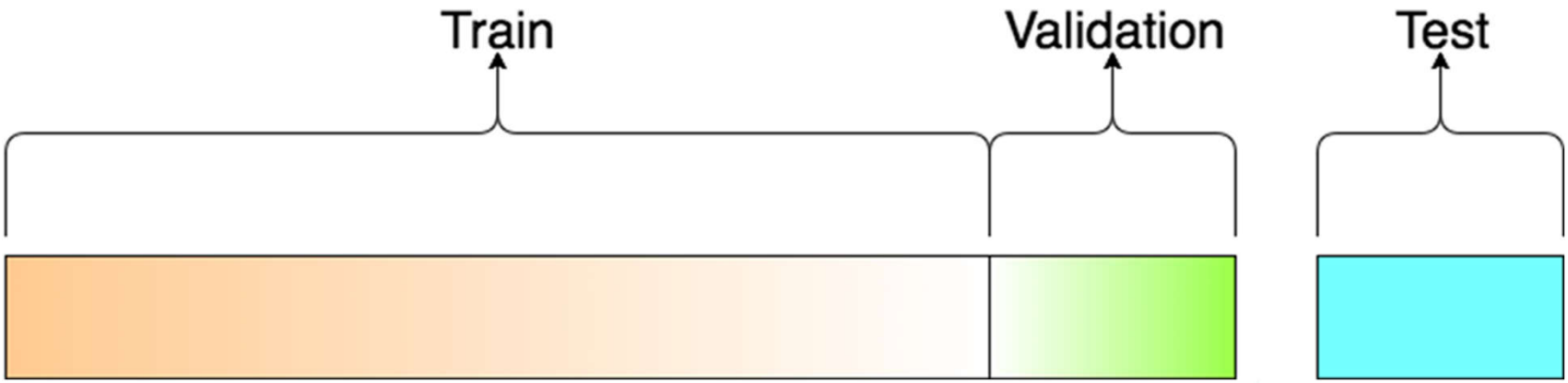
1. What is Data Mining?
2. Data Types
3. Data Mining Tasks
4. Data Mining Process
5. Evaluation Setting and Metrics

Cross validation – k folds

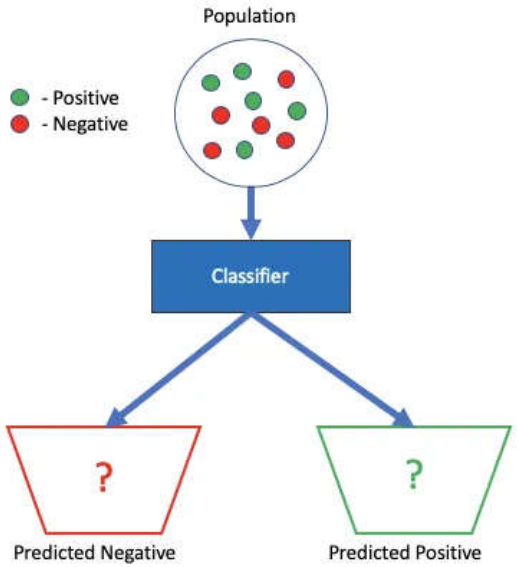


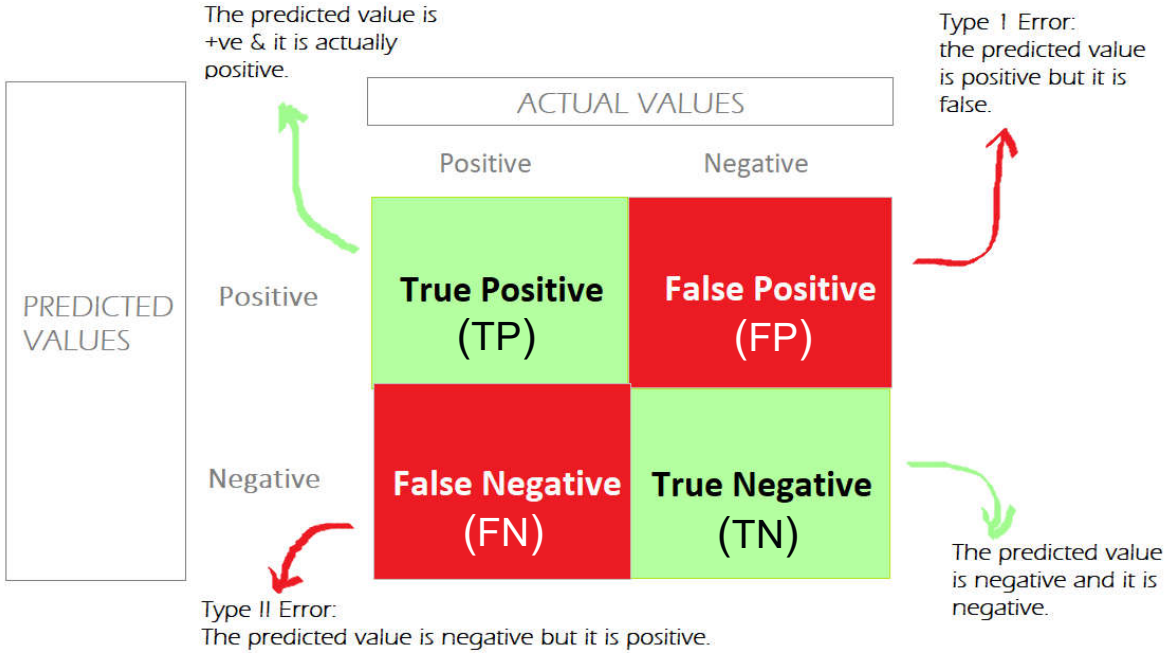
Train, validation and test sets

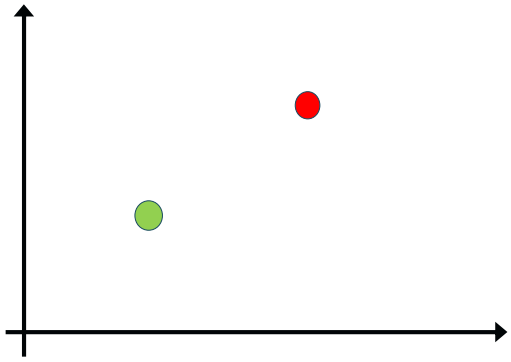
- The data samples are separated in three sets without overlaps
  - Train set is used to train/build the model
  - Validation set is used to fine-tune the model hyper-parameters
  - Test set is used for evaluating the model performance/accuracy



Accuracy, precision, recall, F1, sensitivity,…







$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

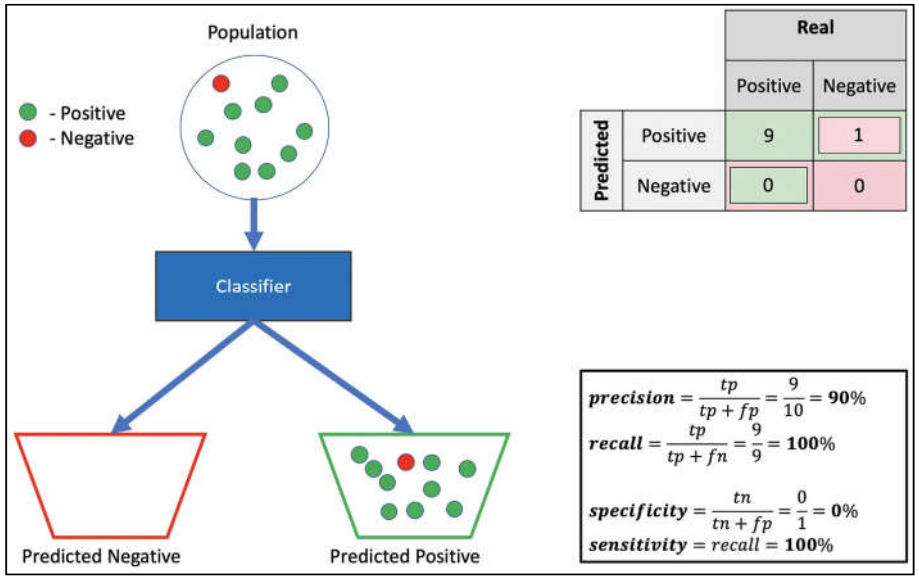
$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

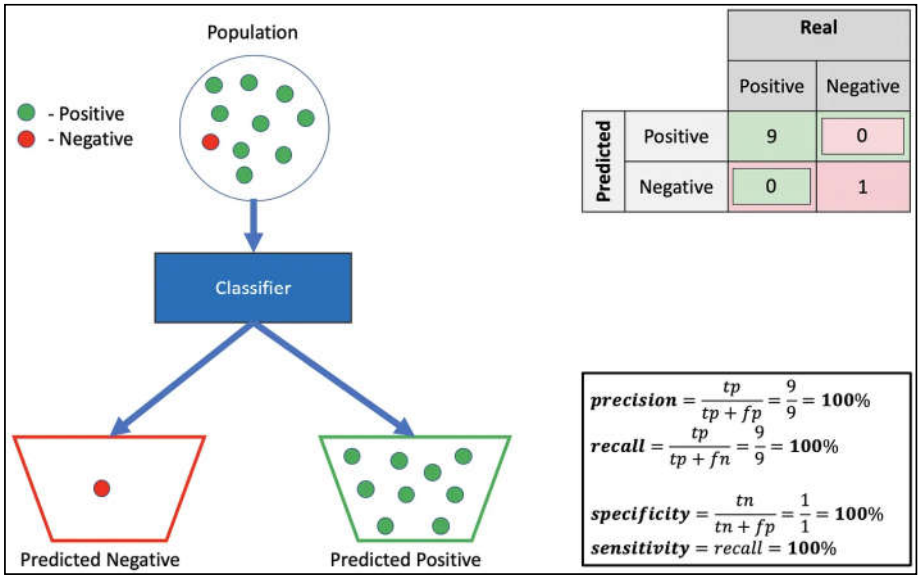
$$\text{Specificity} = \frac{TN}{TN + FP}$$

Accuracy, precision, recall, F1, sensitivity,…

		Actual class (ground truth)		
Total (n)=100		Dog (Positive)	Not a Dog (Negative)	
Predicted class	Dog (Positive)	15 (TP)	20 (FP, Type I Error)	**Precision** =TP/(TP+FP) =0.42
	Not a Dog (Negative)	5 (FN, Type II Error)	60 (TN)	
	Accuracy =(TP+TN)/Total =0.75	**Sensitivity,Recall, TPR** =TP/(TP+FN) = 0.75	**FPR** = FP/(FP+TN) =0.25	**F1 Score** =2*(Precision*Recall) / (Precision+Recall) =0.53
	Error Rate =(FP+FN)/Total =0.25	**Miss Rate, FNR** =FN/(TP+FN) =0.25	**Specificity, TNR** = TN/(FP+TN) =0.75	

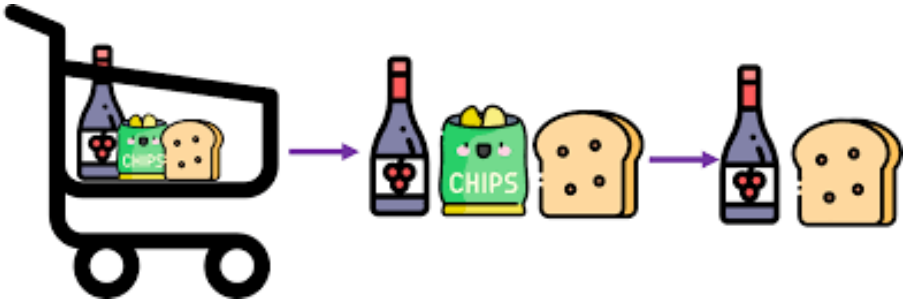
Selection of metrics





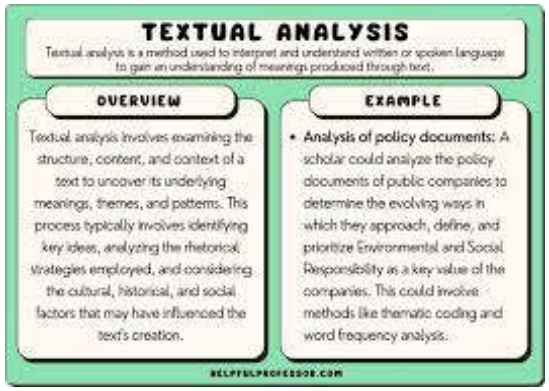
- **Accuracy** is useful when the class distributions are similar.
- **Precision and Recall** are especially useful in cases where classes are imbalanced. For instance, in medical testing, recall might be more important as missing a positive (sick patient) could be more detrimental than falsely identifying someone as sick (high precision but lower recall).
- **F1-Score** is used when you need a balance between precision and recall and there's an uneven class distribution.
- **Specificity** is important in cases where you want to be sure of a negative result, for example, ensuring patients not having a disease are not treated unnecessarily.

The course introduces …



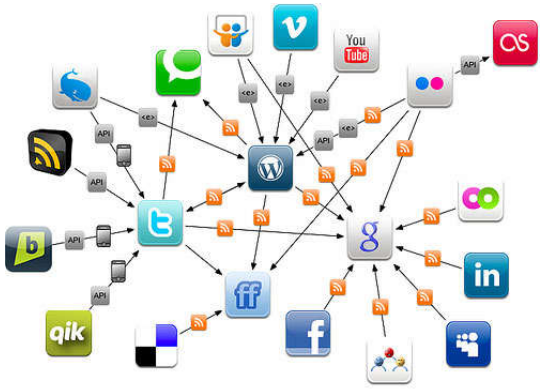
Frequent itemset mining

Matrix data mining



Text data mining

Image data mining



Network data mining

Time-series data mining

Email: min.shi@louisiana.edu