

Introduction to Data Mining

Chapter 1 Introduction

by Michael Hahsler

Based on slides by Tan,
Steinbach, Karpatne, Kumar

 This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).



Agenda

- What is Data Mining?
- Data Mining Tasks
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues



What is Data Mining?

One of many definitions:

*"Data mining is the science **of extracting useful knowledge** from huge data repositories."*

ACM SIGKDD, Data Mining Curriculum: A Proposal

<http://www.kdd.org/curriculum>

Why Data Mining? Commercial Viewpoint

- Businesses collect and warehouse lots of data.
 - Purchases at department/grocery stores
 - Bank/credit card transactions
 - Web and social media data
 - Mobile and IOT
- Computers are cheaper and more powerful.
- Competition to provide better services.
 - Mass customization and recommendation systems
 - Targeted advertising
 - Improved logistics



Agenda

- What is Data Mining?
- Data Mining Tasks**
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues



Data Mining Tasks

Data Preparation

Data Wrangling: Data acquisition, understanding, cleaning, and preprocessing.

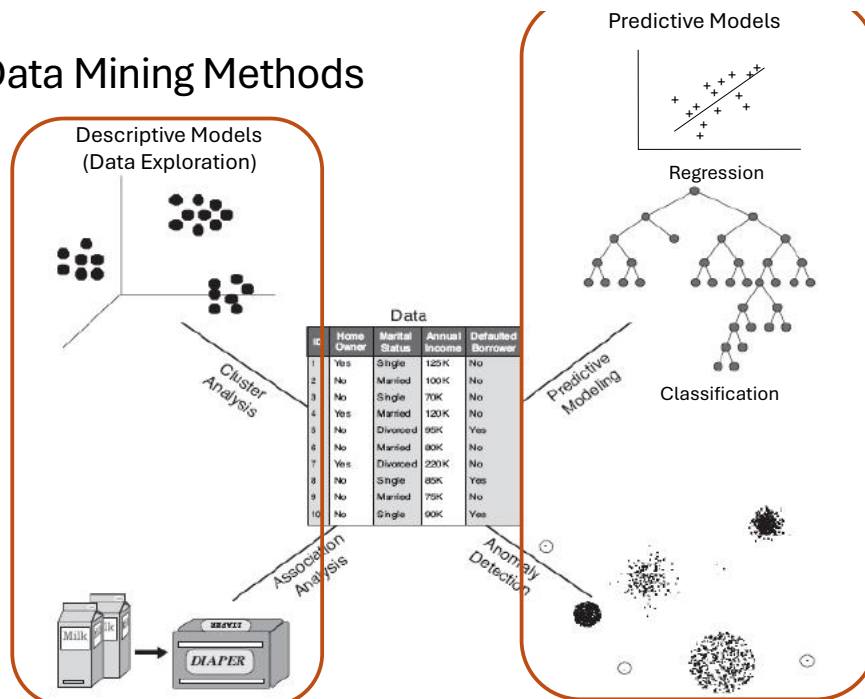
Descriptive Methods

Data Exploration: Find human-interpretable patterns that describe the data. Visualize patterns.

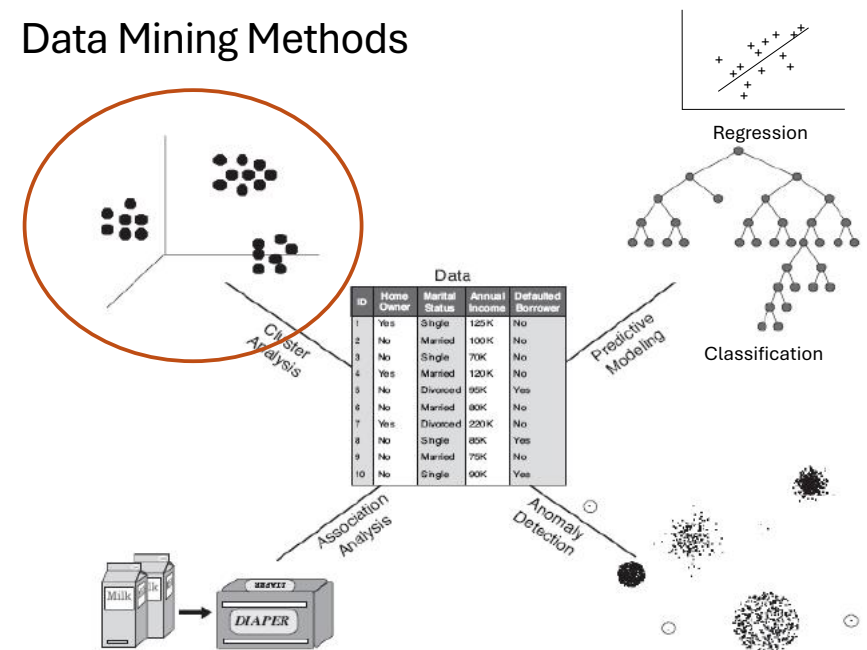
Predictive Methods

Modeling: Use some features (variables) to predict unknown or future value of other variable.

Data Mining Methods



Data Mining Methods

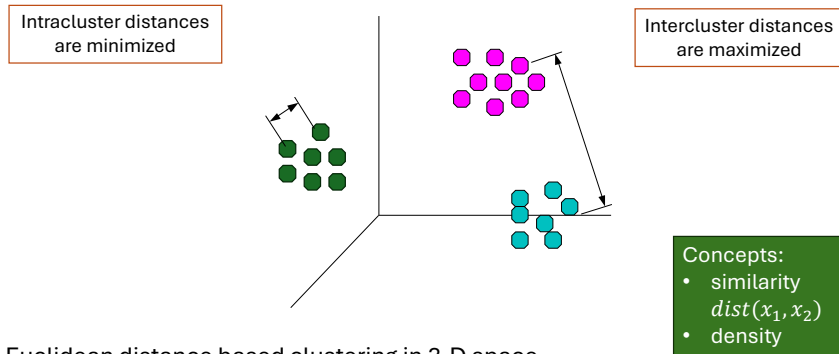


Clustering

Group points such that

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Ideal grouping is not known → Unsupervised Learning



Euclidean distance based clustering in 3-D space.

Clustering Documents



Goal: Find groups of documents that are similar to each.



Approach: Identify frequently occurring terms in each document. Define a similarity measure based on term co-occurrences. Use it to cluster.



Gain: Can be used to organize documents or to create recommendations.

Clustering: Market Segmentation



Goal: subdivide a market into distinct subsets of customers. Use a different marketing mix for each segment.



Approach:

1. Collect different attributes of customers based on their geographical and lifestyle related information and observed buying patterns.
2. Find clusters of similar customers.

Clustering: Data Reduction

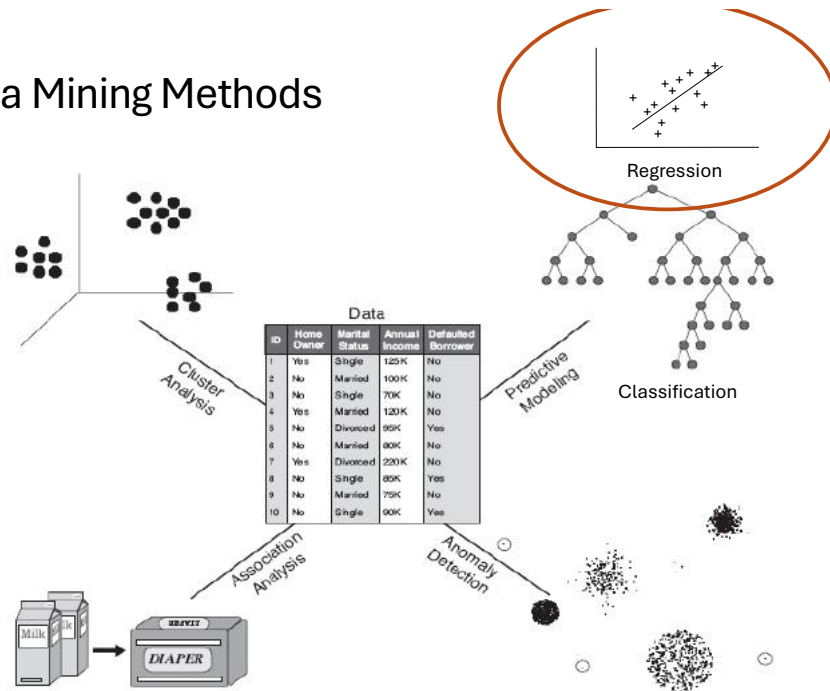


Goal: Reduce the data size for predictive models.



Approach: Group data given a subset of the available information and then use the group label instead of the original data as input for predictive models.

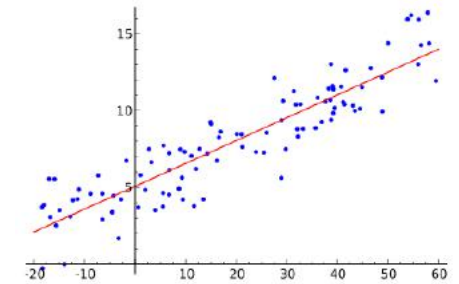
Data Mining Methods



Regression

Concept:
• Function fitting:
 $y = f(X)$

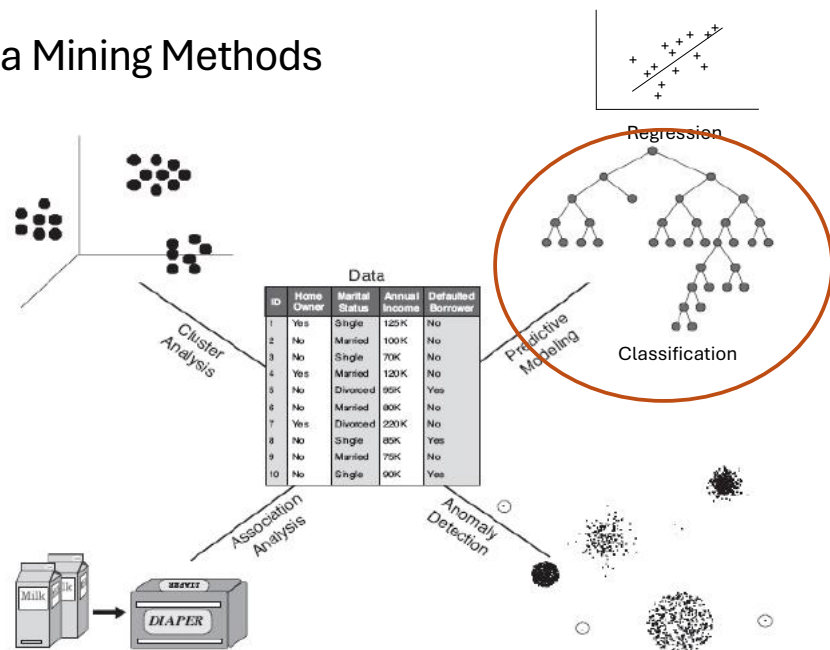
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Studied in statistics and econometrics.



Applications:

- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices (autoregressive models).

Data Mining Methods



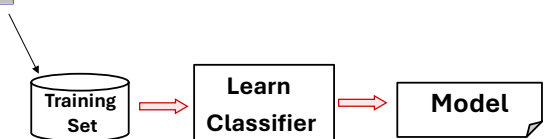
Classification

Find a **model** for the class attribute as a function of the values of other attributes/features.

Class information is available → **Supervised Learning**

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

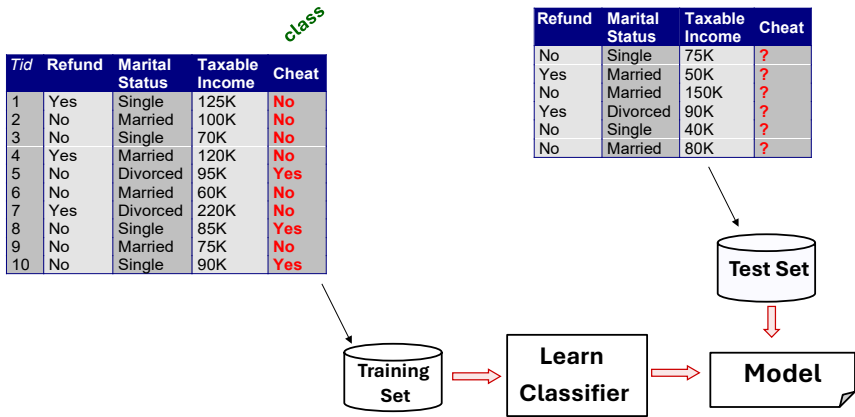
Concept:
• Probability estimation
 $P(y | X)$



Classification

Find a **model** for the class attribute as a function of the values of other attributes/features.

Goal: assign new records to a class as accurately as possible.



Classification: Customer Attrition/Churn

- Goal: To predict whether a customer is likely to be lost to a competitor.
- Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes (frequency, recency, complaints, demographics, etc.).
 - Label the customers as loyal or disloyal.
 - Find a model for disloyalty.
 - Rank each customer on a loyal/disloyal scale (e.g., churn probability).

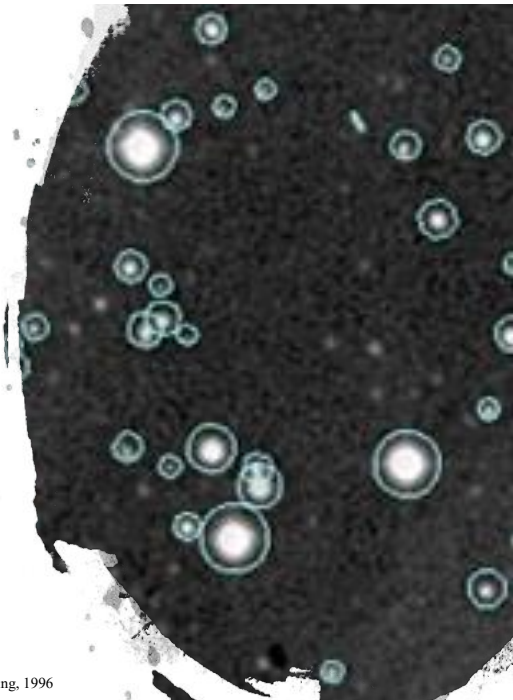


Classification: Direct Marketing

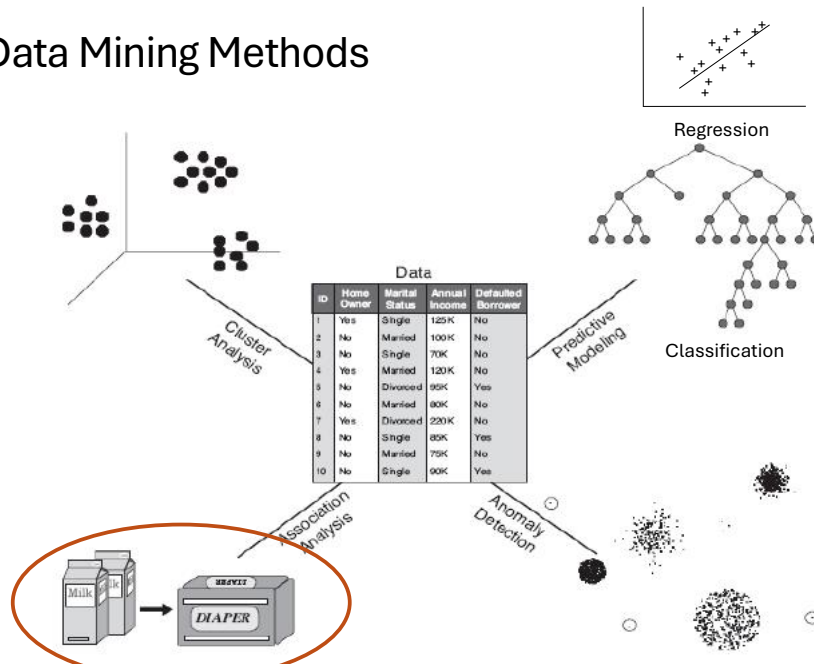
- Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new product.
- Approach:
 - Use the data for a similar product introduced before or from a focus group. We have customer information (e.g., demographics, lifestyle, previous purchases) and know which customers decided to buy and which decided otherwise. This buy/don't buy decision forms the class attribute.
 - Use this information as input attributes to learn a classifier model.
 - Apply the model to new customers to predict if they will buy the product.

Classification: Sky Survey Cataloging

- Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
- Approach:
 - Segment the image to identify objects.
 - Derive features per object (40).
 - Use known objects to model the class based on these features.
- Result: Found 16 new high red-shift quasars.



Data Mining Methods



Association Rule Discovery



- Given is a set of transactions. Each contains a number of items.
- Produce dependency rules of the form $LHS \rightarrow RHS$
- which indicate that if the set of items in the LHS are in a transaction, then the transaction likely will also contain the RHS item.

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

Transaction data

Discovered Rules

- $\{Milk\} \rightarrow \{Coke\}$
- $\{Diaper, Milk\} \rightarrow \{Beer\}$

Concept:
 • Probability estimation
 $P(X \rightarrow y)$

Association Rule Discovery Marketing and Sales Promotion

- Let the rule discovered be $\{Potato\ Chips, \dots\} \rightarrow \{Soft\ drink\}$
- Soft drink as RHS:** What should be done to boost sales? Discount Potato Chips?
- Potato Chips in LHS:** Shows which products would be affected if the store discontinues selling Potato Chips.
- Potato Chips in LHS and Soft drink in RHS:** What products should be sold with Potato Chips to promote sales of Soft drinks!



Association Rule Discovery Supermarket shelf management

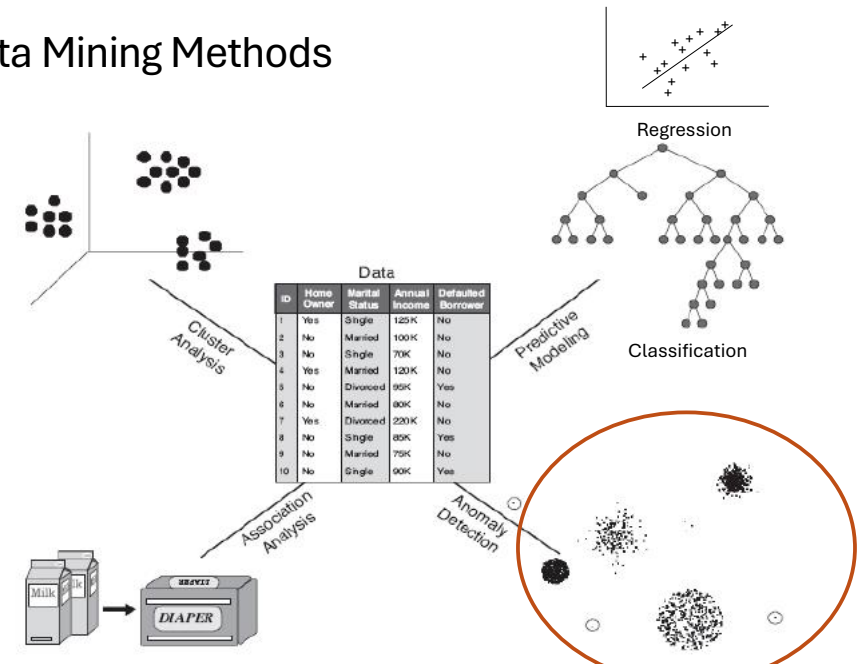
- Goal:** To identify items that are bought together by sufficiently many customers.
- Approach:**
 - Process the point-of-sale data to find dependencies among items.
 - Place dependent items
 - close to each other (convenience).
 - far from each other to expose the customer to the maximum number of products in the store.



Association Rule Discovery Inventory Management

- **Goal:** Anticipate the nature of repairs to keep the service vehicles equipped with right parts to speed up repair time.
- **Approach:** Process the data on tools and parts required in previous repairs at different consumer locations and discover co-occurrence patterns.

Data Mining Methods



Deviation/Anomaly Detection

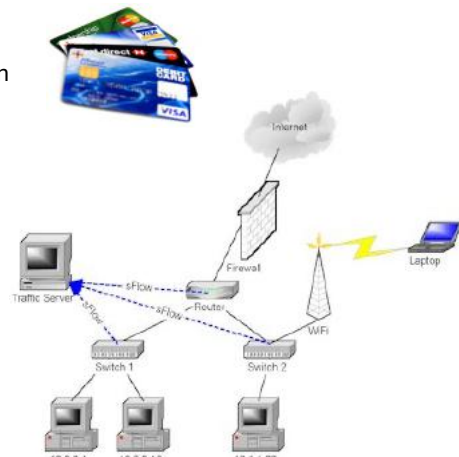
Concept:
• Probability estimation
 $P(y | X)$

- Detect significant deviations from normal behavior.

- Applications:
— Credit Card Fraud Detection

- Network Intrusion Detection

Typical network traffic at University level may reach over 100 million connections per day



Other Data Mining Tasks

Text mining – document clustering, topic models

Graph mining – social networks

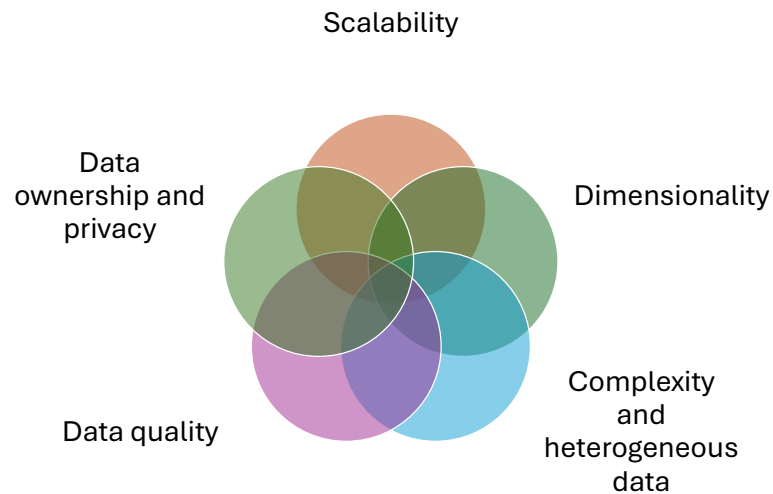
Data stream mining/real time data mining

Mining spatiotemporal data (e.g., moving objects)

Visual data mining

Distributed data mining

Challenges of Data Mining



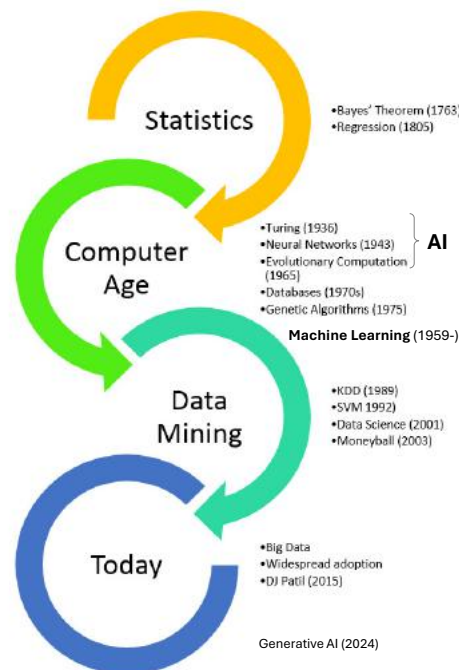
Agenda

- What is Data Mining?
- Data Mining Tasks
- **Relationship to Statistics, Optimization, Machine Learning and AI**
- Tools
- Data
- Legal, Privacy and Security Issues

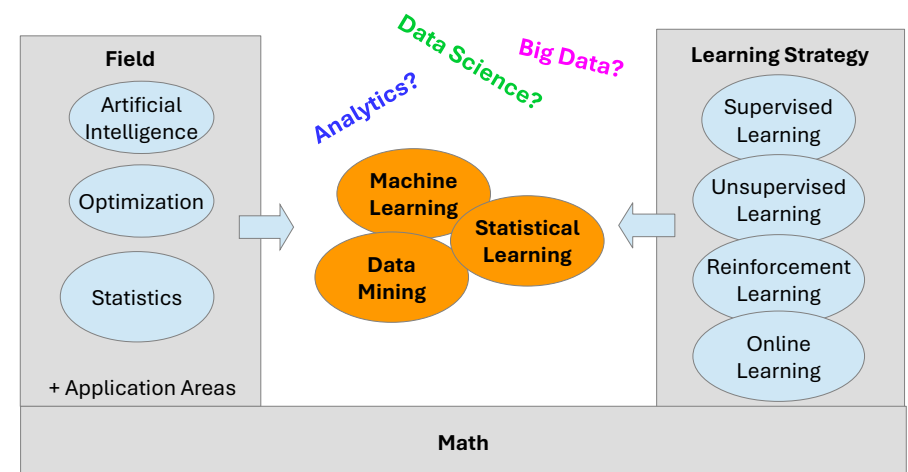


Origins of Data Mining

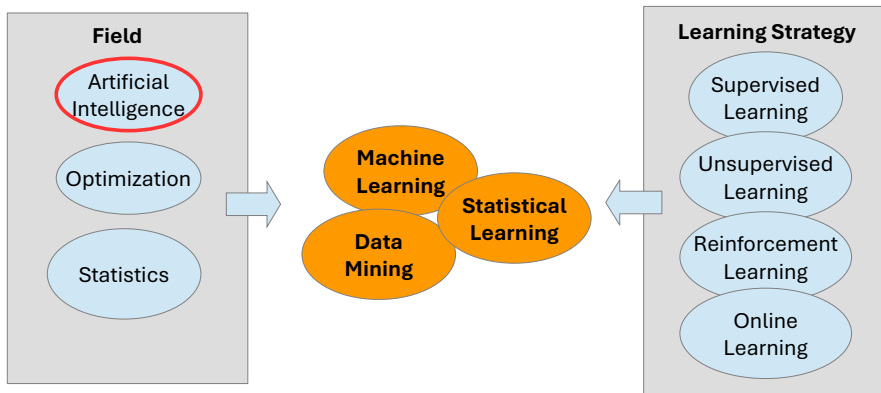
- Draws ideas from AI, machine learning, pattern recognition, statistics, and database systems.
- There are differences in terms of
 - used data and
 - the goals.



Relationship to other Fields

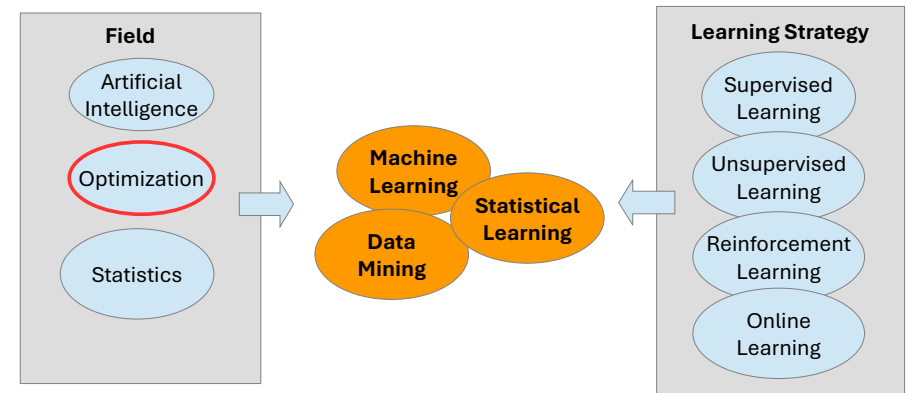


Relationship to other Fields



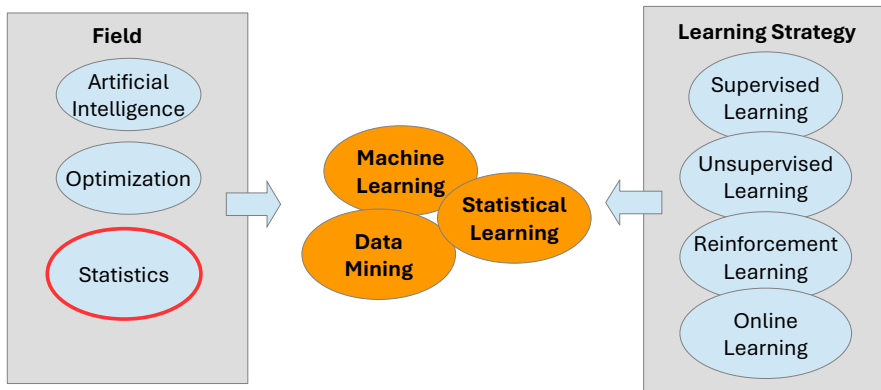
Artificial Intelligence: Create an **autonomous agent** that perceives its environment and takes actions that maximize its chance of reaching some goal.
Areas: reasoning, knowledge representation, planning, learning, natural language processing, and vision.

Relationship to other Fields



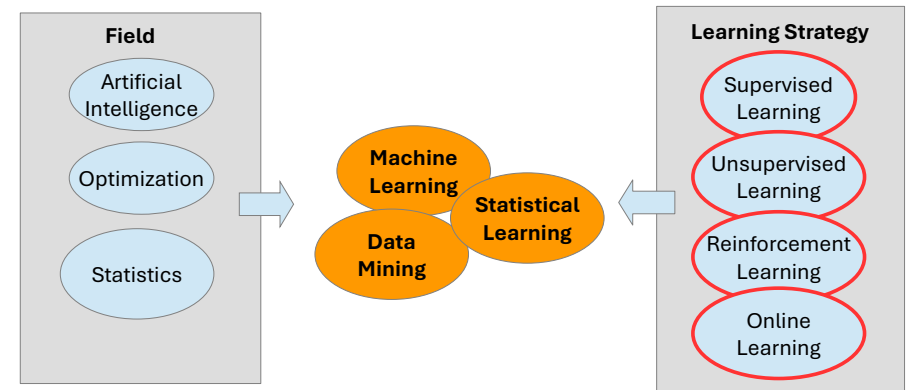
Optimization: Selection of a best alternative from some set of available alternatives using an objective criterion.
Techniques: Linear programming, integer programming, nonlinear programming, stochastic and robust optimization, heuristics, etc.

Relationship to other Fields



Statistics: Study of the collection, analysis, interpretation, presentation, and organization of data.
Techniques: Descriptive statistics, statistical inference (estimation, testing), design of experiments.

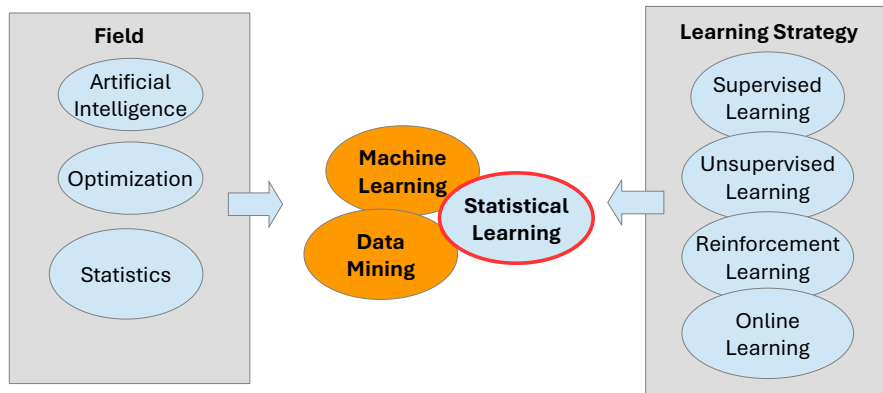
Relationship to other Fields



Learning Strategy: From what data do we learn?

- Is a training set with correct answers available? → Supervised learning
- Long-term structure of rewards? → Reinforcement learning
- No answer and no reward structure available? → Unsupervised learning
- Do we have to update the model regularly? → Online learning

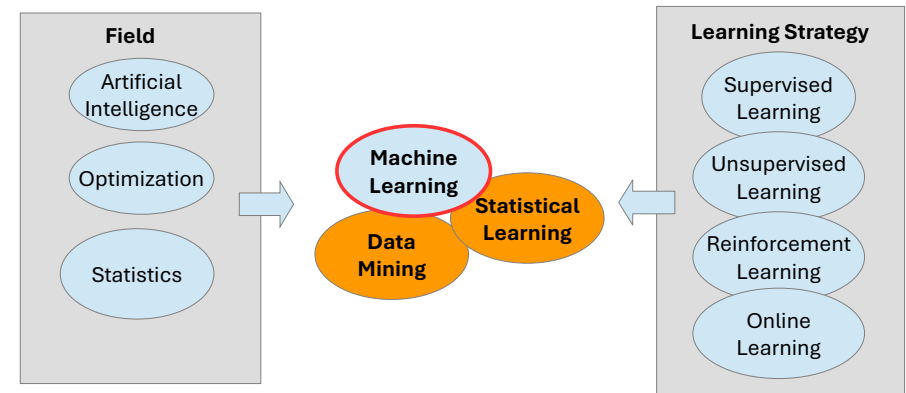
Relationship to other Fields



Statistical learning: deals with the problem of finding a **predictive function** based on data.

Techniques: (Linear) classifiers, regression and regularization.

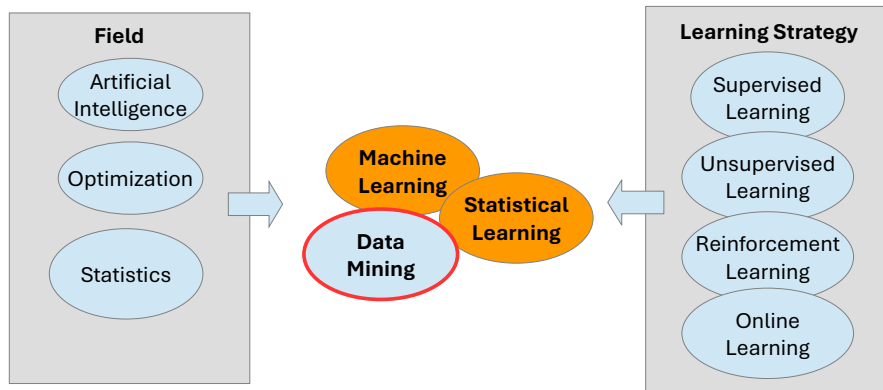
Relationship to other Fields



Machine Learning create algorithms that can learn from data and generalize to unseen data to perform a task without explicit instructions.

Techniques: Focus on supervised learning (e.g., classification).

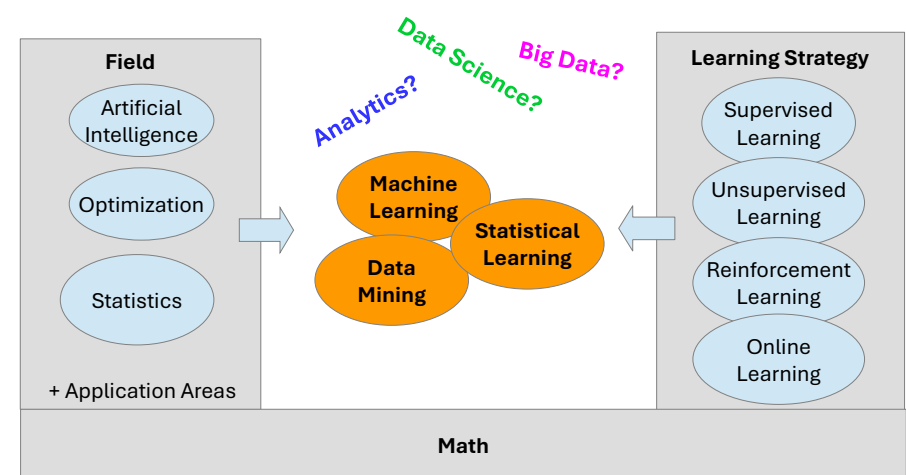
Relationship to other Fields



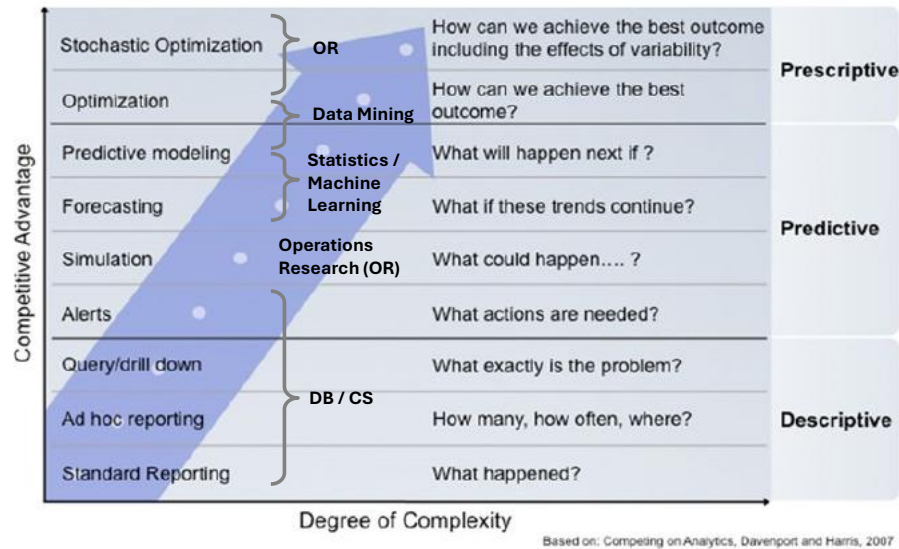
Data Mining: Analyze a given dataset to gain insights (knowledge) that can be used to improve outcomes.

Techniques: Any applicable technique from databases, statistics, machine/statistical learning. New methods were developed by the data mining community.

Analytics, Data Science and Big Data

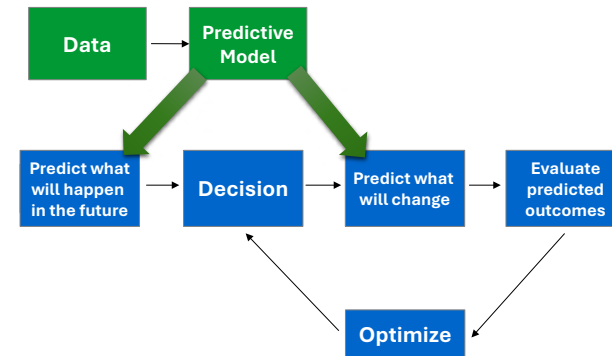


Analytics covers Data Mining



The Prescriptive Analytics Approach

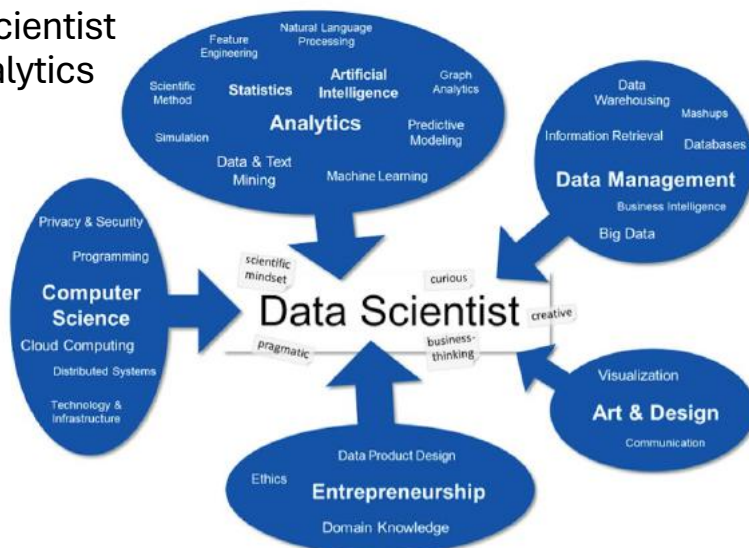
What decisions should we make now to achieve the best future outcome?
This is also the objective of Data Mining.



Issues:

- What are the decision variables? Causality?
- Relationship can be non-linear. Convex? Optimization may be challenging.
- Uncertainty about quality and reliability of the predictive model.

A Data Scientist does Analytics



Source: T. Stadelmann, et al., Applied Data Science in Europe

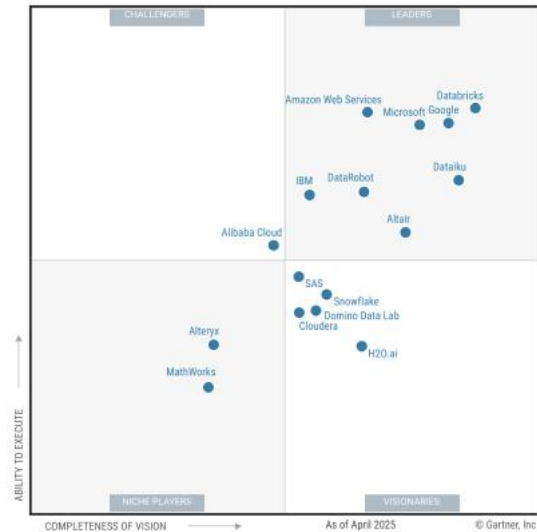
Good luck finding this person!
Probably a team effort!

Agenda

- What is Data Mining?
- Data Mining Tasks
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools**
- Data
- Legal, Privacy and Security Issues



Tools: Commercial Players

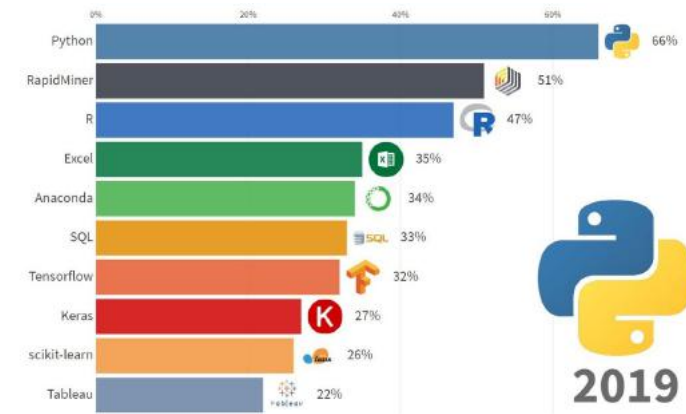


Gartner®

2025 Gartner MQ Data Science and Machine Learning.

Only covers companies, not open-source tools like Python and R.

Tools: Popularity – Data Science Tools Popularity Poll



<https://www.kdnuggets.com/polls/>

Question: What tools do you use? (multiple answers possible) N = 1800

Tools: Types

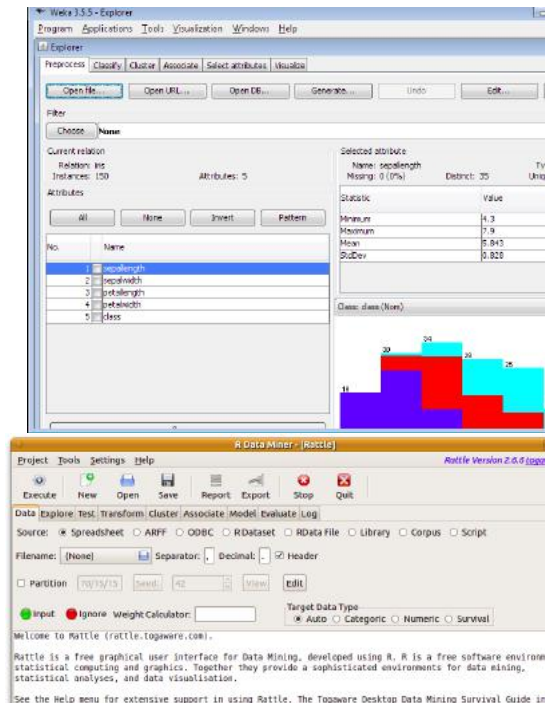
Simple graphical user interface

Process oriented

Programming oriented

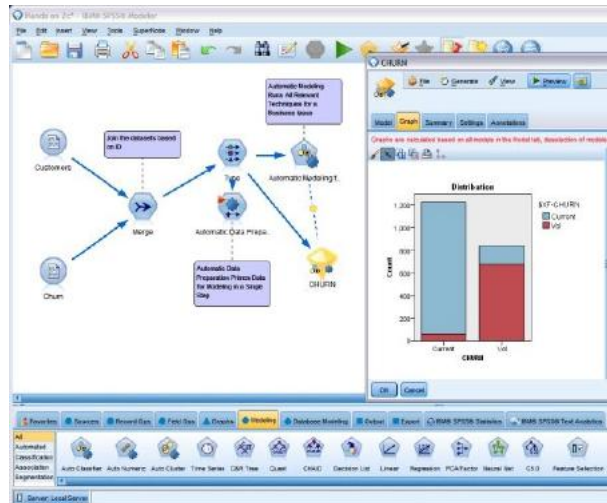
Tools: Simple GUI

- Weka: Waikato Environment for Knowledge Analysis (also has a Java API)
- Rattle: GUI for Data Mining using R



Tools: Process oriented

- SAS Enterprise Miner
- IBM SPSS Modeler
- RapidMiner
- Knime
- Orange



Tools: Programming oriented

- R
 - Rattle for beginners
 - RStudio IDE, R markdown, shiny
- Python
 - Numpy, scikit-learn, pandas
 - Jupyter notebook

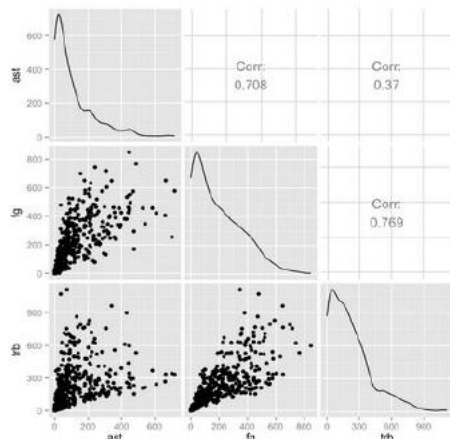


→ Both have similar capabilities but slightly different focus:

- R: statistical computing and visualization, data mining
- Python: Scripting, big data, deep learning, ML
- Interoperability via rpy2 and reticulate

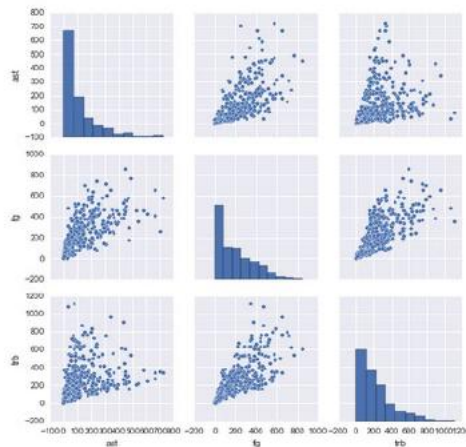
R

```
library(GGally)
ggpairs(nba[,c("ast", "fg", "trb")])
```



Python

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.pairplot(nba[["ast", "fg", "trb"]])
plt.show()
```



Agenda

- What is Data Mining?
- Data Mining Tasks
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues



Data

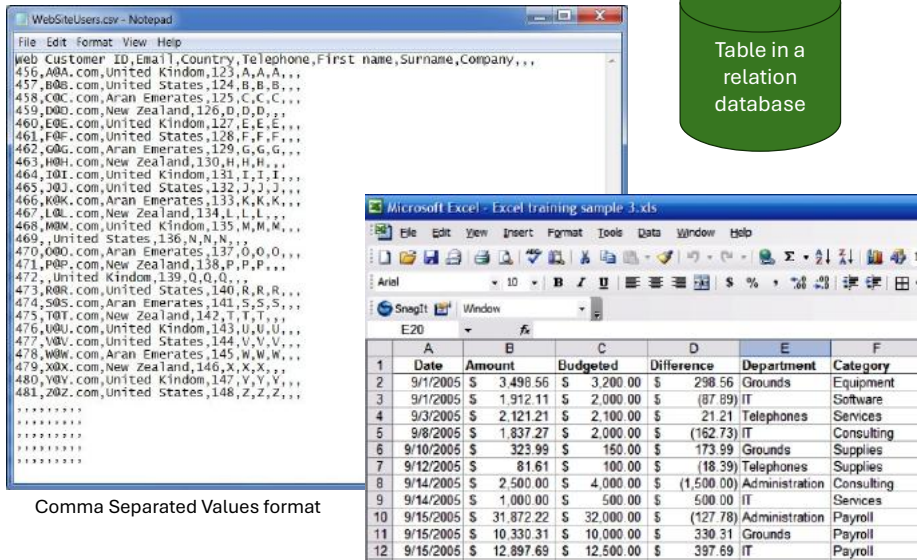
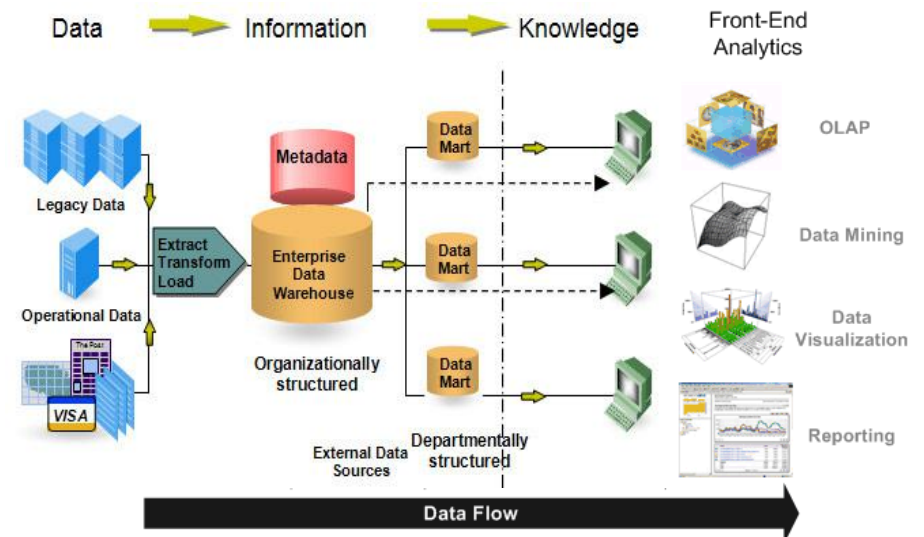


Table in a relation database

Comma Separated Values format

1	A	B	C	D	E	F
	Date	Amount	Budgeted	Difference	Department	Category
2	9/1/2005	\$ 3,498.56	\$ 3,200.00	\$ 298.56	Grounds	Equipment
3	9/1/2005	\$ 1,912.11	\$ 2,000.00	\$ (87.89)	IT	Software
4	9/3/2005	\$ 2,121.21	\$ 2,100.00	\$ 21.21	Telephones	Services
5	9/8/2005	\$ 1,837.27	\$ 2,000.00	\$ (162.73)	IT	Consulting
6	9/10/2005	\$ 323.99	\$ 150.00	\$ 173.99	Grounds	Supplies
7	9/12/2005	\$ 81.61	\$ 100.00	\$ (18.39)	Telephones	Supplies
8	9/14/2005	\$ 2,500.00	\$ 4,000.00	\$ (1,500.00)	Administration	Consulting
9	9/14/2005	\$ 1,000.00	\$ 500.00	\$ 500.00	IT	Services
10	9/15/2005	\$ 31,872.22	\$ 32,000.00	\$ (127.78)	Administration	Payroll
11	9/15/2005	\$ 10,330.31	\$ 10,000.00	\$ 330.31	Grounds	Payroll
12	9/15/2005	\$ 12,897.69	\$ 12,500.00	\$ 397.69	IT	Payroll

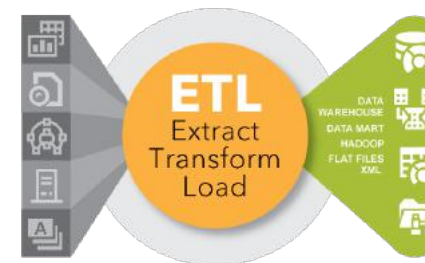
Data Warehouse



Data Warehouse

- **Subject Oriented:** Data warehouses are designed to help you analyze data (e.g., sales data is organized by product and customer).
- **Integrated:** Integrates data from disparate sources into a consistent format.
- **Nonvolatile:** Data in the data warehouse are never overwritten or deleted.
- **Time Variant:** maintains both historical and (nearly) current data.

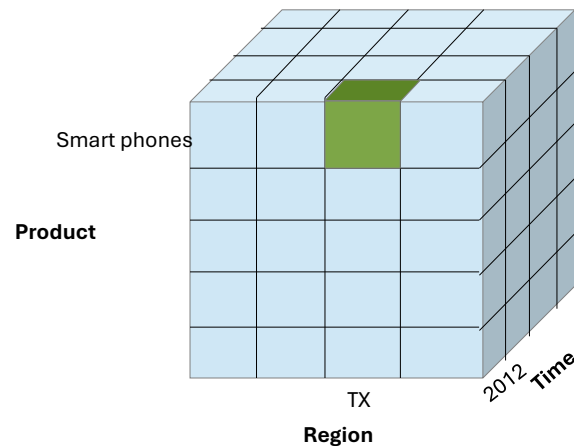
ETL: Extract, Transform and Load



Source: SAS, ETL: What it is and why it matters

- **Extracting data from outside sources**
- **Transforming data to fit analytical needs. E.g.,**
 - Clean missing data, wrong data, etc.
 - Normalize and translate (e.g., 1 → "female")
 - Join from several sources
 - Calculate and aggregate data
- **Loading data into the data warehouse**

OnLine Analytical Processing (OLAP)



Operations:

- Slice
- Dice
- Drill-down
- Roll-up
- Pivot

Store data in "data cubes" for fast OLAP operations.
Requires a special database structure (Snow-flake scheme).

Big Data



- "Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them." Wikipedia
- 3 V's: Volume, velocity, variety, (veracity) Gartner

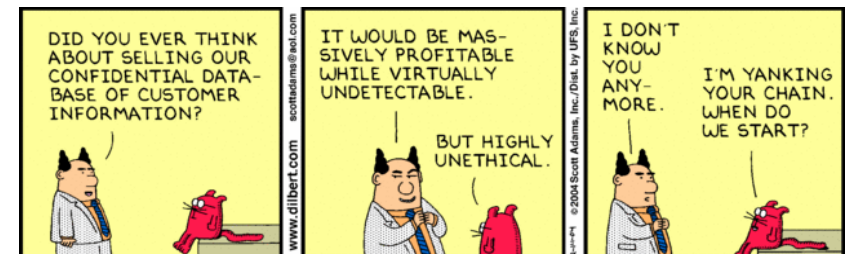


Agenda

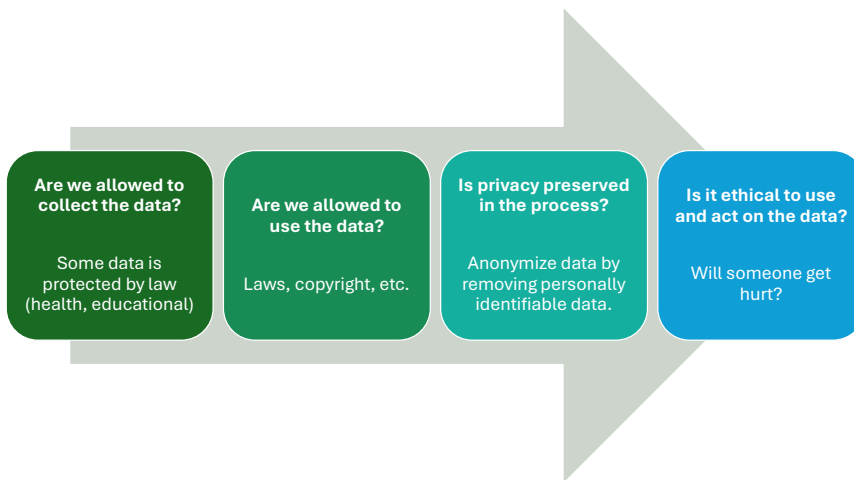
- What is Data Mining?
- Data Mining Tasks
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- **Ethics, Privacy and Security Issues**



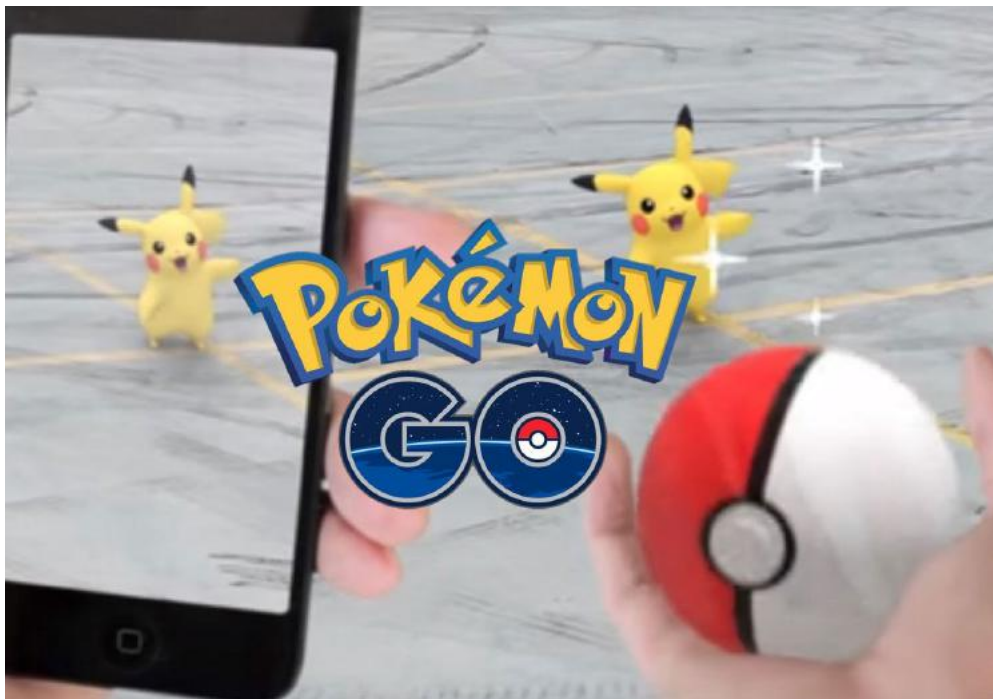
Legal, Privacy and Security Issues



Legal, Privacy and Security Issues



Problem: Internet is global, but legislation is local!



Ethics, Privacy and Security Issues

The New York Times

Data-Gathering via Apps
Presents a Gray Legal Area
By KEVIN J. O'BRIEN
Published: October 28, 2012



BERLIN — Angry Birds, the top-selling paid mobile app for the iPhone in the United States and Europe, has been downloaded more than a billion times by devoted game players around the world, who often spend hours slinging squawking fowl at groups of egg-stealing pigs.

When Jason Hong, an associate professor at the Human-Computer Interaction Institute at Carnegie Mellon University, surveyed 40 users, all but two were **unaware that the game was storing their locations so that they could later be the targets of ads....**

Here is what the small print says...

USA Today Network [Josh Hafner](#), 2:38 p.m. EDT July 13, 2016



Pokémon Go's constant location tracking and camera access required for gameplay, paired with its skyrocketing popularity, could provide data like no app before it.

"Their privacy policy is vague," Hong said. "I'd say deliberately vague, because of the lack of clarity on the business model."

...

The agreement says **Pokémon Go collects data about its users as a "business asset."** This includes data used to personally identify players such as email addresses and other information pulled from Google and Facebook accounts players use to sign up for the game. If Niantic is ever sold, the agreement states, all that data can go to another company.

Conclusion

Data Mining is interdisciplinary and overlaps significantly with many fields

- Statistics
- CS (machine learning, AI, data bases)
- Optimization (Operations Research)
- (Business) Analytics
- Data Science

Data Mining requires a team effort with members who have expertise in several areas

- Data management
- Statistics
- Programming
- Communication
- + Application domain