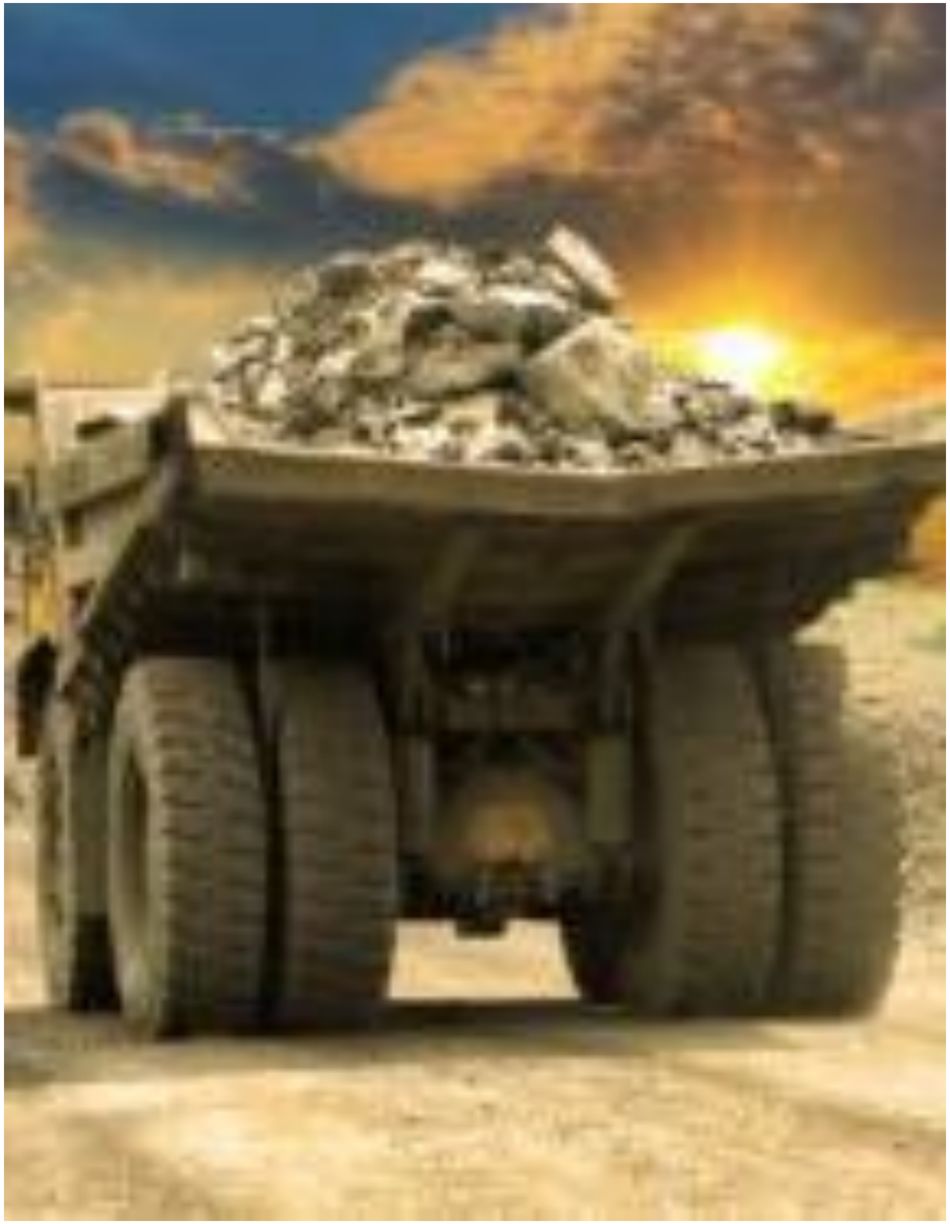
"**Data** is more valuable than **oil** and is the **most expensive asset** in the world."

Introduction to Data Mining Methods and Tools

by Michael Hahsler

Agenda

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



Evolution of Database Technology

1960s:

- Data collection, database creation, IMS and network DBMS

1970s:

- Relational data model, relational DBMS implementation

1980s:

- RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
- Application-oriented DBMS (spatial, scientific, engineering, etc.)

1990s:

- Data mining, data warehousing, multimedia databases, and Web databases

2000s

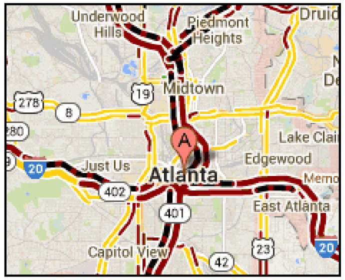
- Stream data management and mining
- Data mining and its applications
- Web technology (XML, data integration) and global information systems

Large-scale Data is Everywhere!

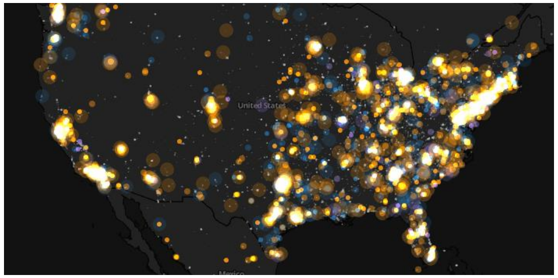
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Cyber Security

E-Commerce

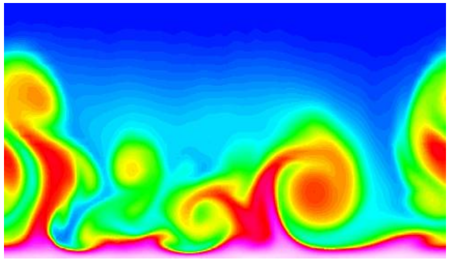


Traffic Patterns



Social Networking: Twitter

Sensor Networks



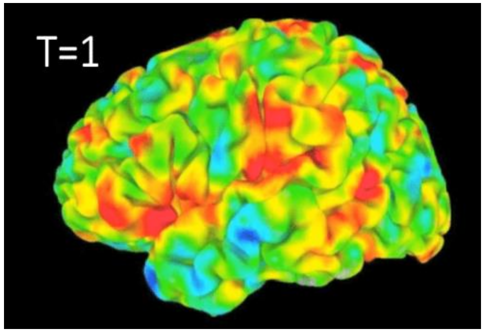
Computational Simulations

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Google has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

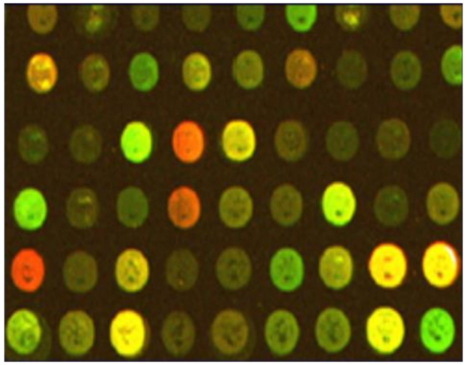
Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

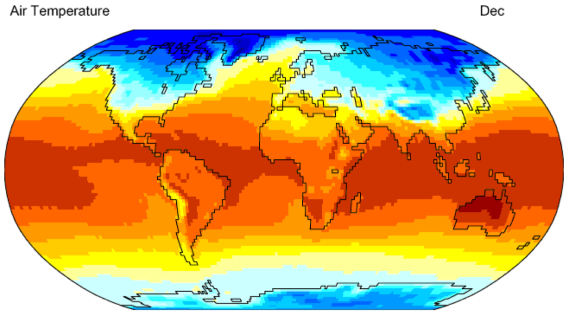


fMRI Data from Brain

Sky Survey Data

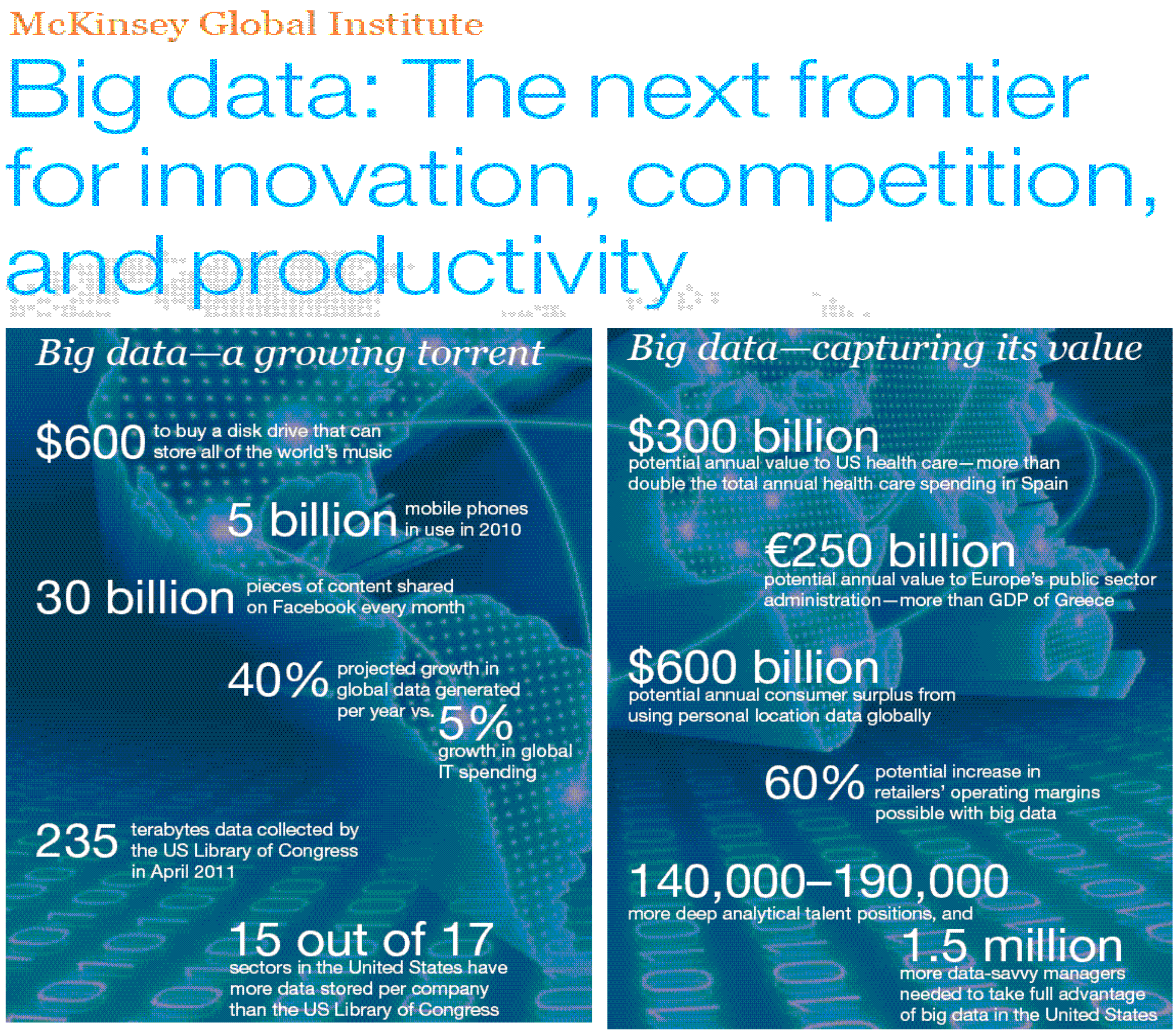


Gene Expression Data



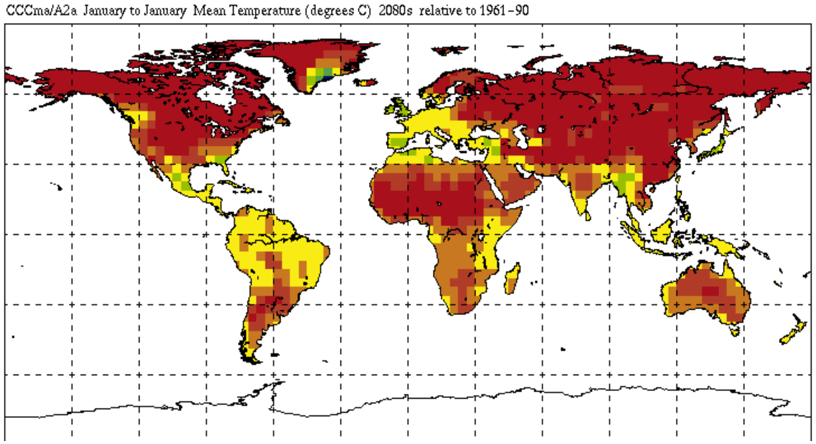
Surface Temperature of Earth

Great opportunities to improve productivity in all walks of life



Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



Predicting the impact of climate change

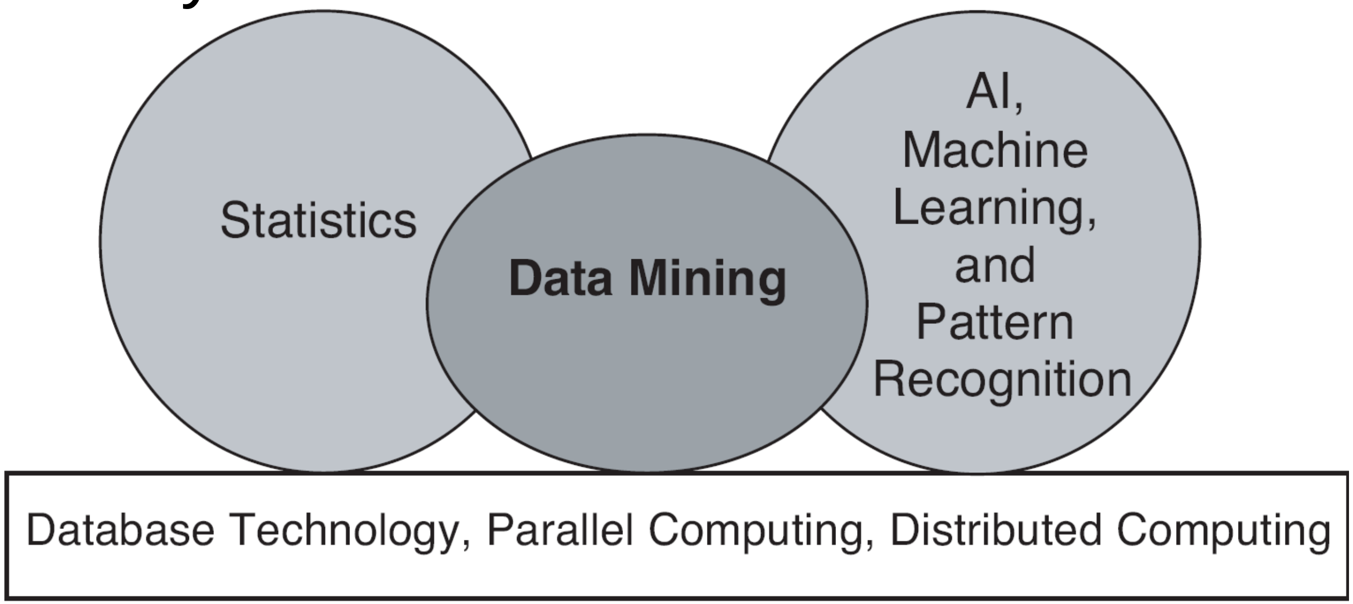
Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

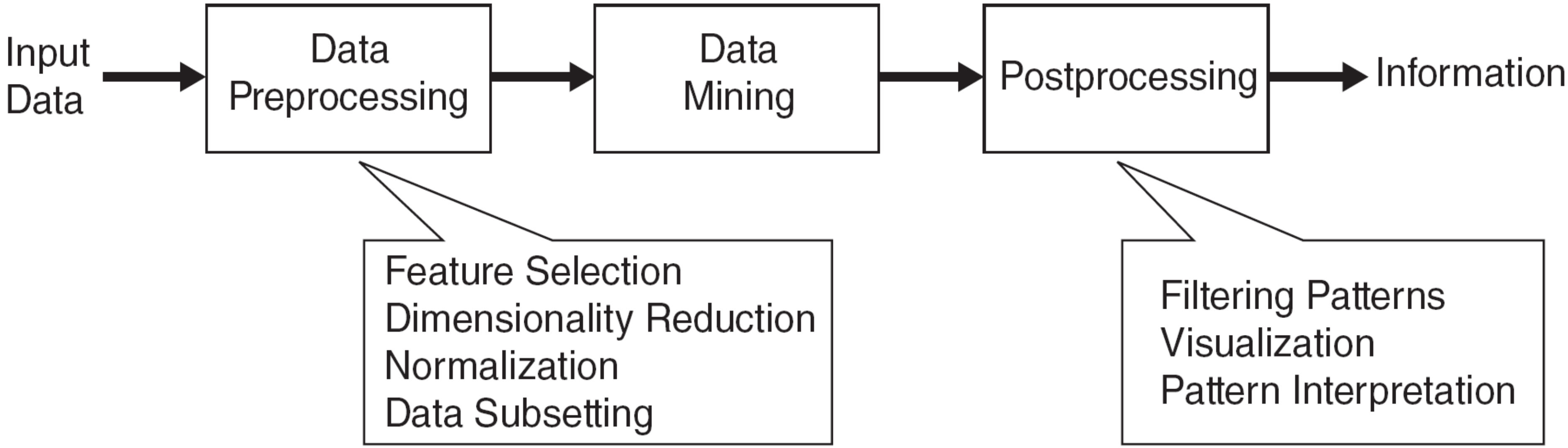
- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed



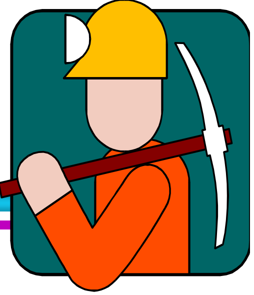
- A key component of the emerging field of data science and data-driven discovery

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



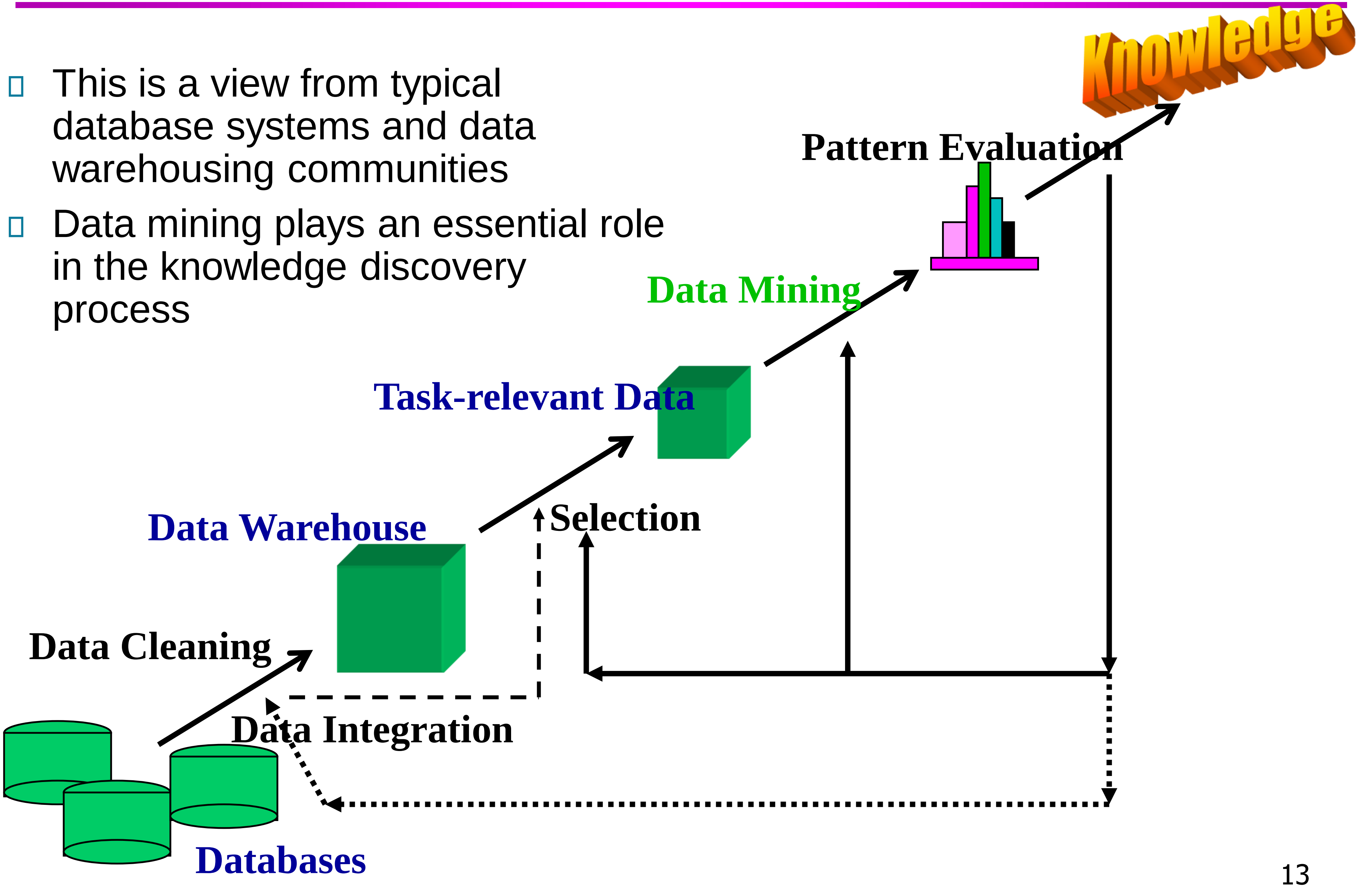
What Is Data Mining?



- 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
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems

Knowledge Discovery (KDD) Process

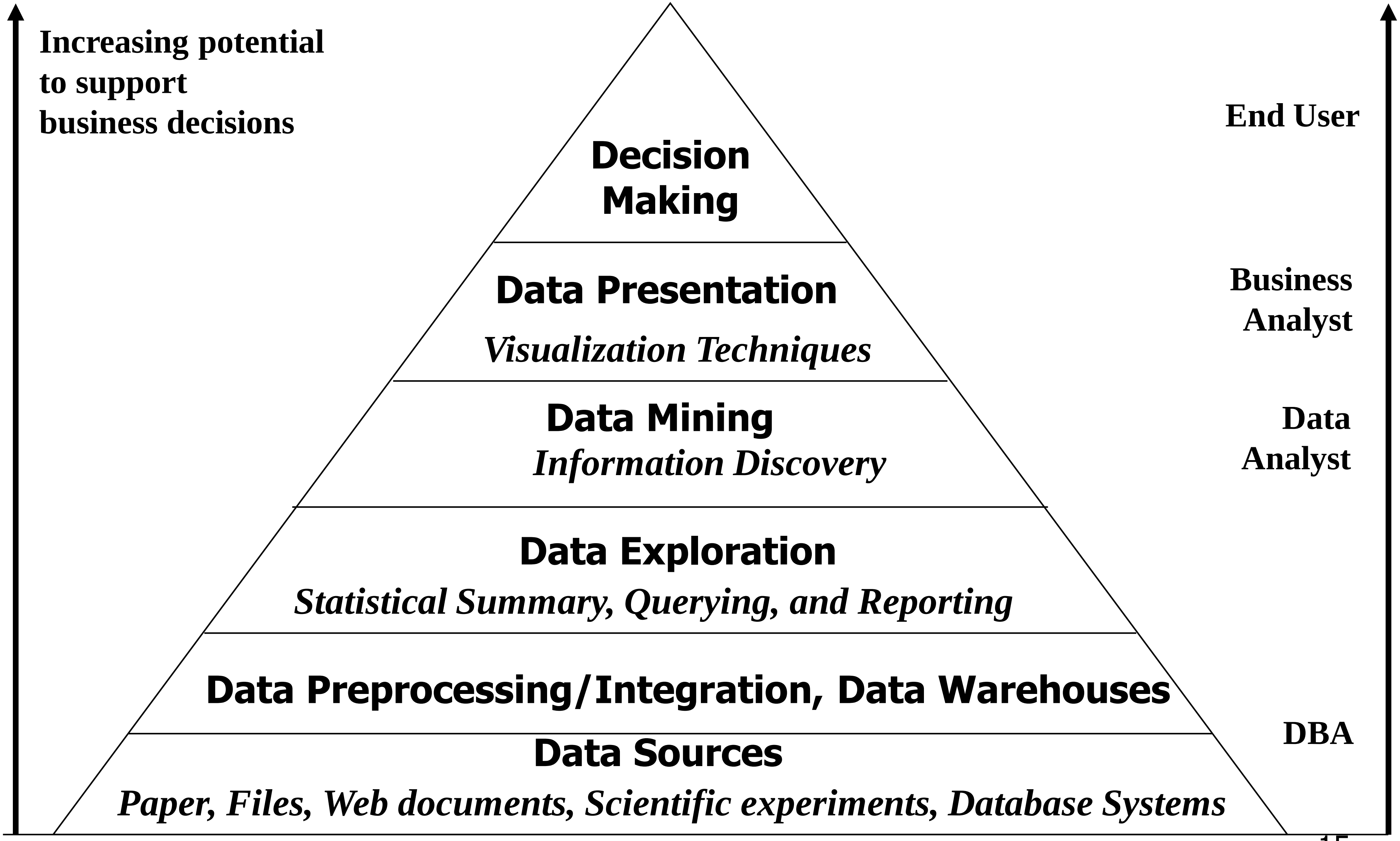
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Example: A Web Mining Framework

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

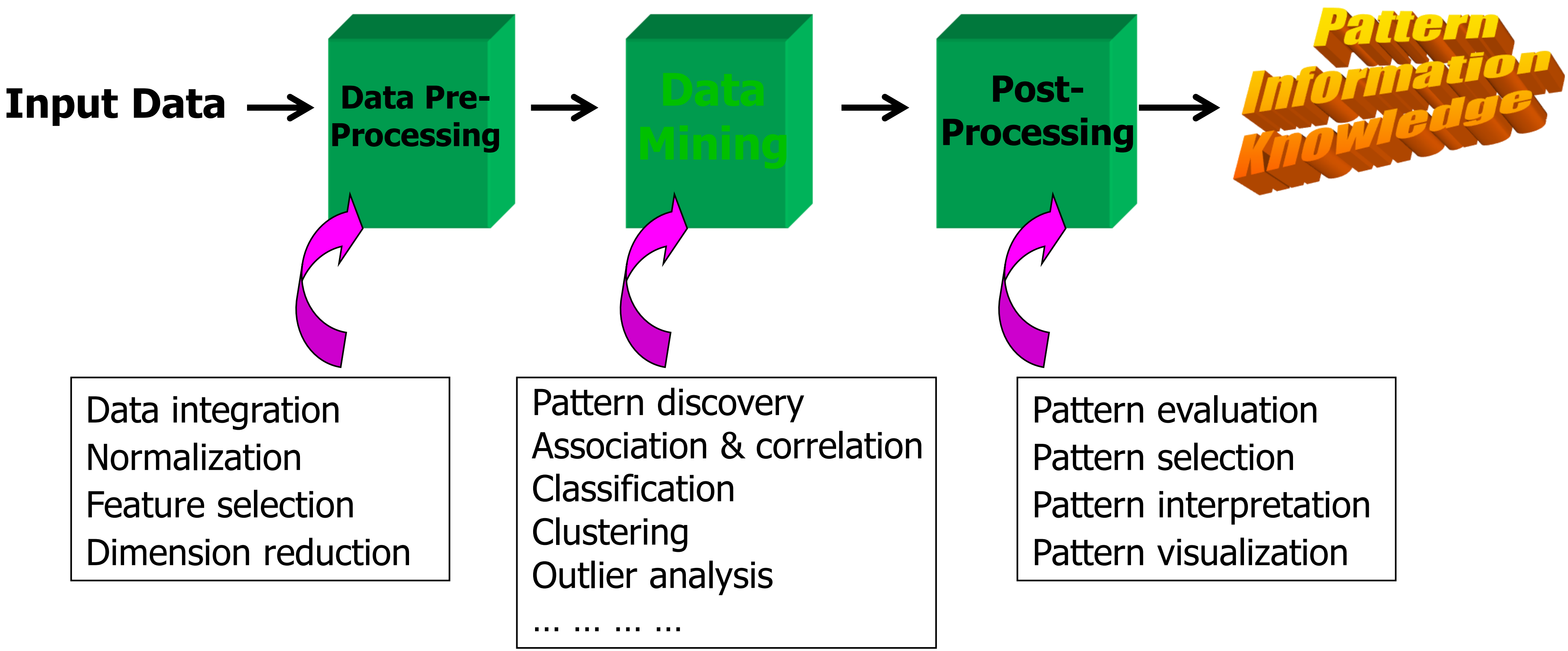
Data Mining in Business Intelligence



Example: Mining vs. Data Exploration

- Business intelligence view
  – Warehouse, data cube, reporting but not much mining
- Business objects vs. data mining tools
- Supply chain example: tools
- Data presentation
- Exploration

KDD Process: A Typical View from ML and Statistics



- This is a view from typical machine learning and statistics communities

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

What is Data Mining?

One of many definitions:

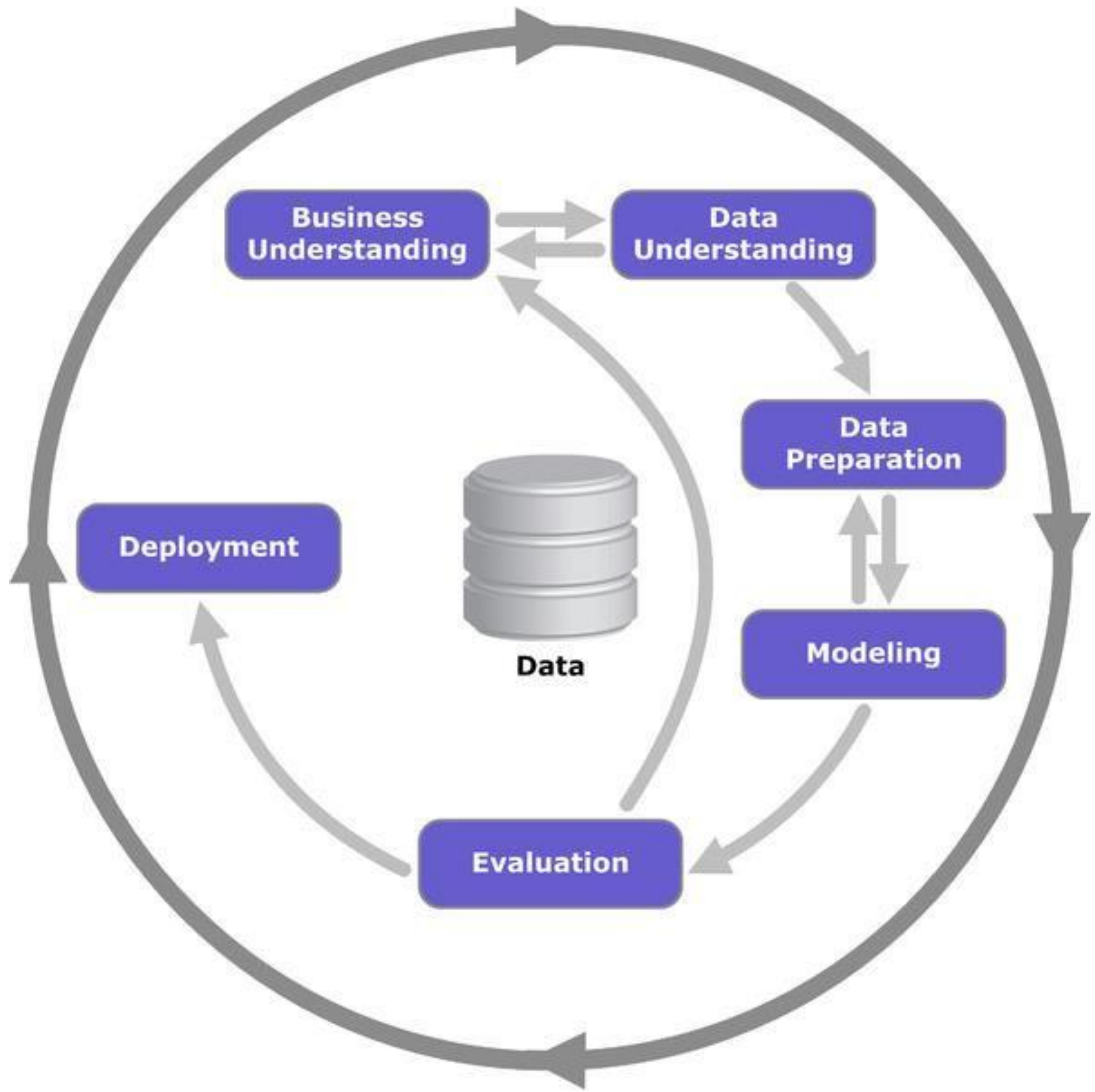
> *"Data mining is the science* ***of extracting useful knowledge*** *from huge data repositories."*

ACM SIGKDD, Data Mining Curriculum: A Proposal

http://www.kdd.org/curriculum

CRISP-DM Reference Model

- Cross Industry Standard Process for Data Mining
- Open standard process model
- Industry, tool and application neutral
- Defines tasks and outputs.
- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).
- SAS has SEMMA and most consulting companies use their own similar process.



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

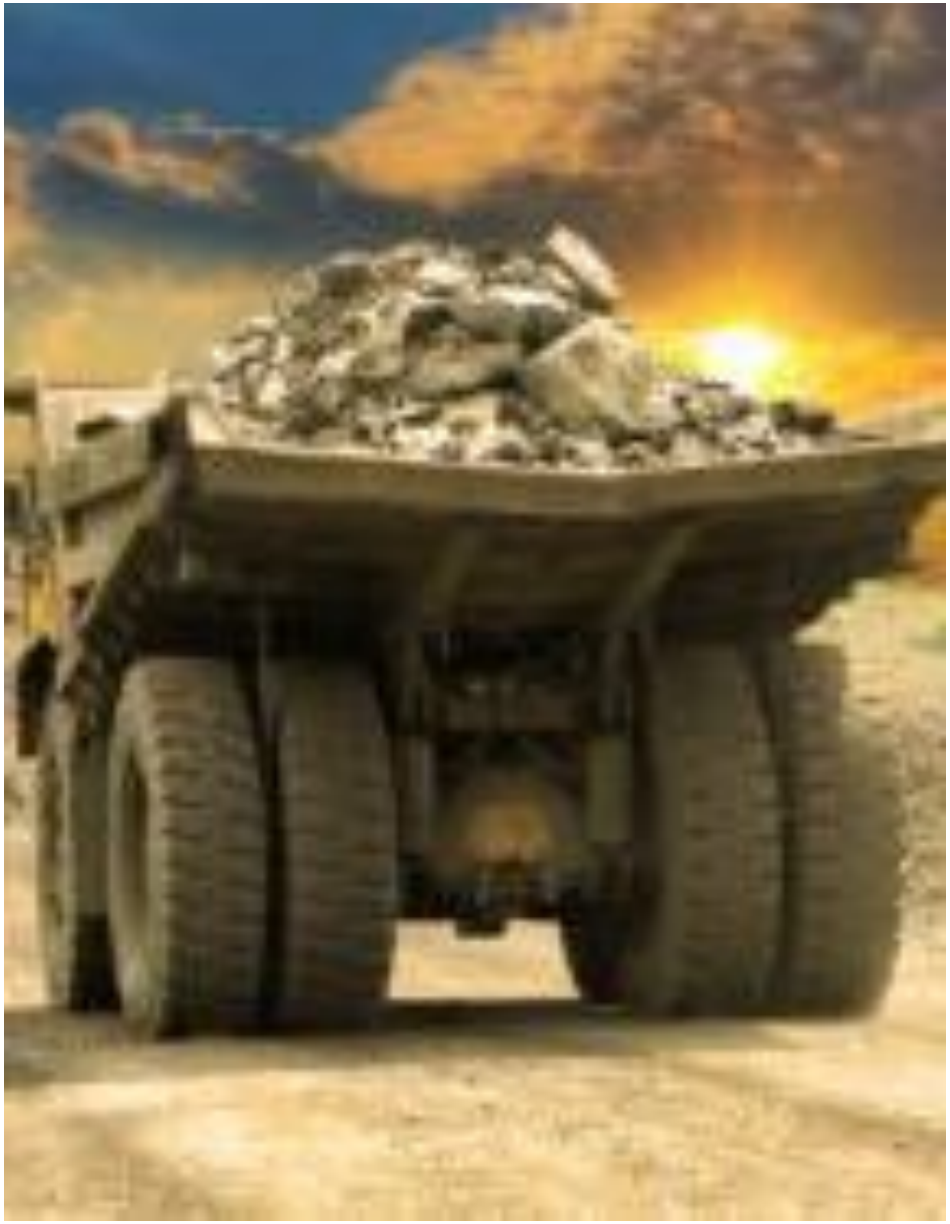
Tasks in the CRISP-DM Model

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

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

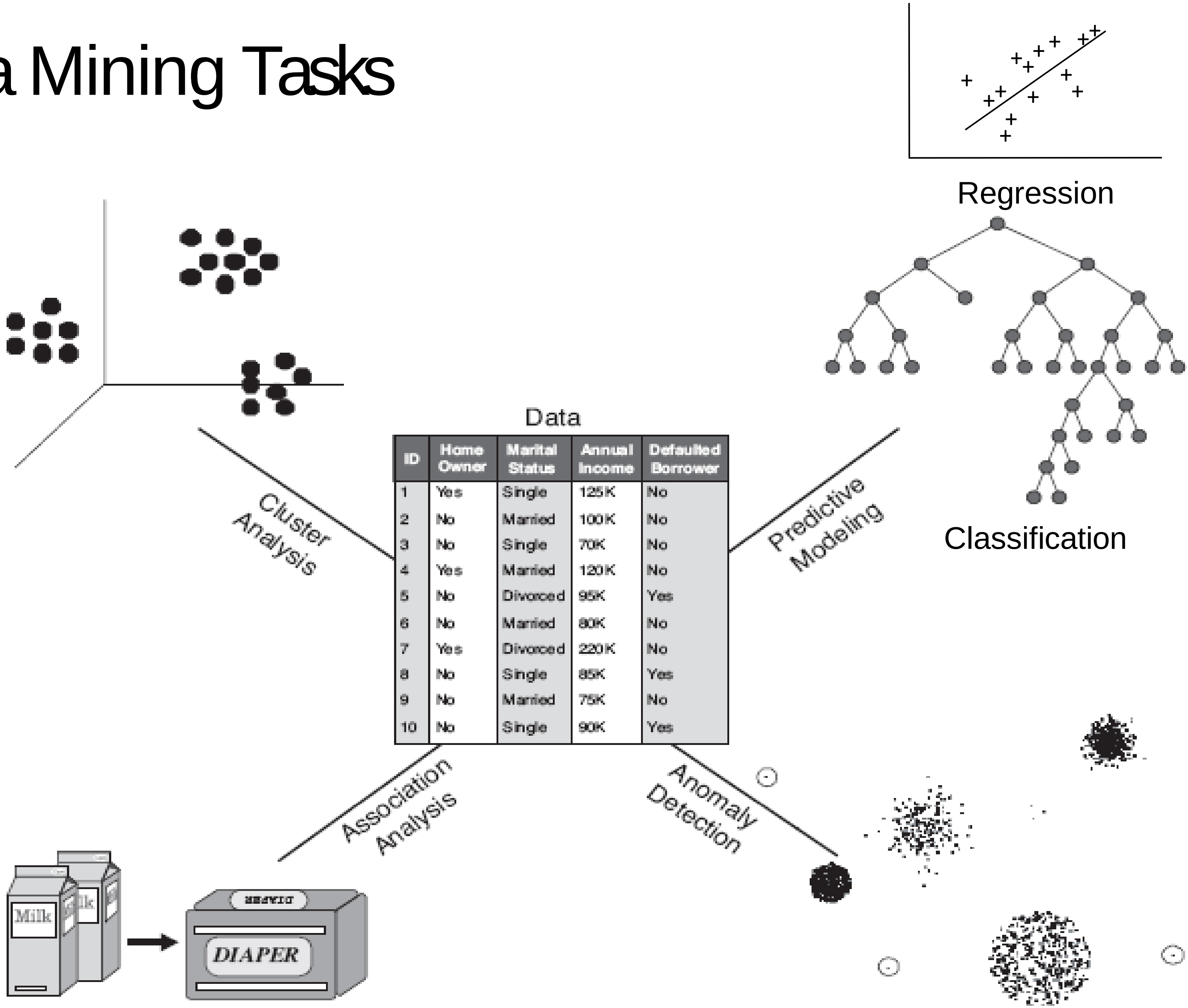
Descriptive Methods

Find human-interpretable patterns that describe the data.

Predictive Methods

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

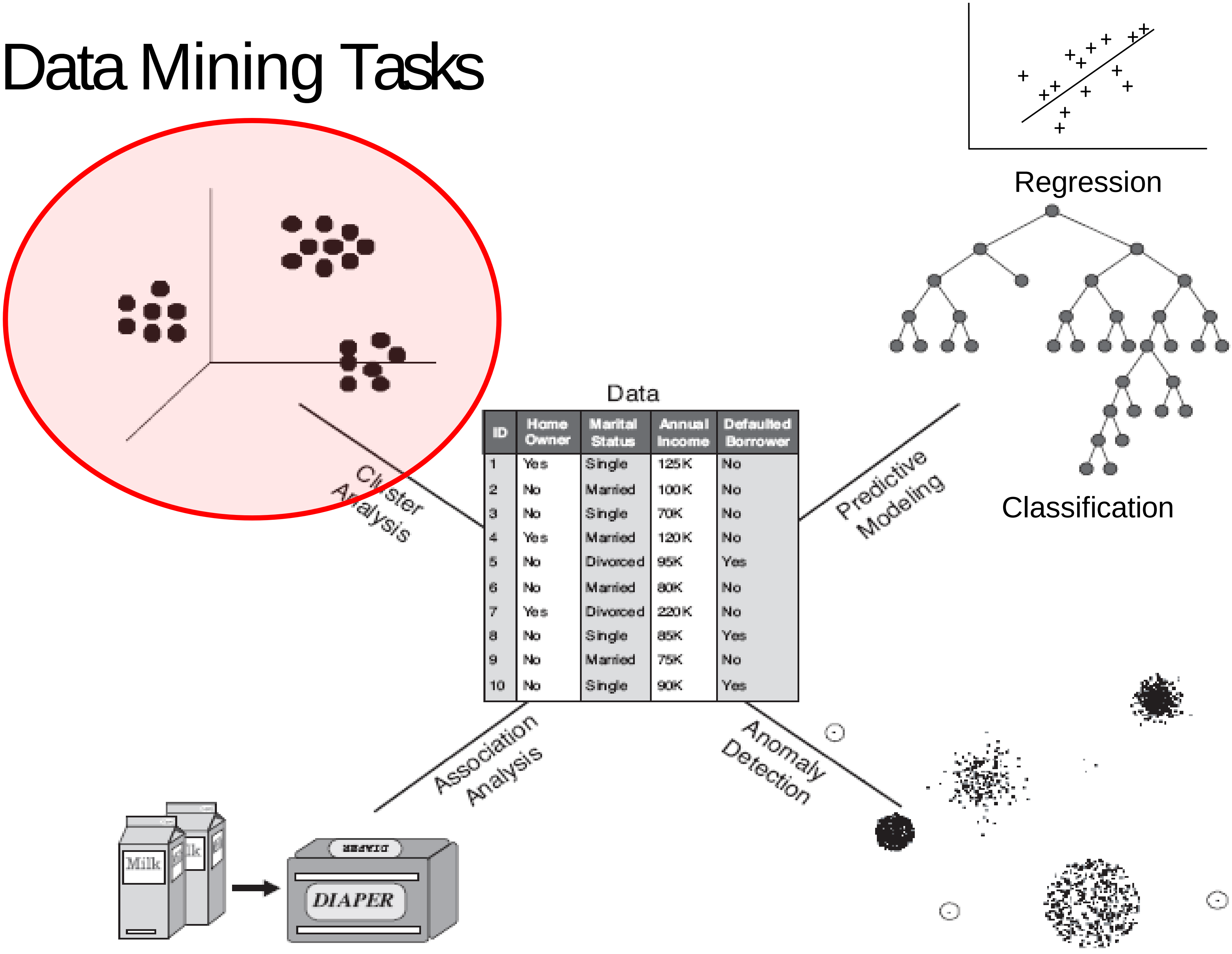
Data Mining Tasks



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley, 2006

Data Mining Tasks



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

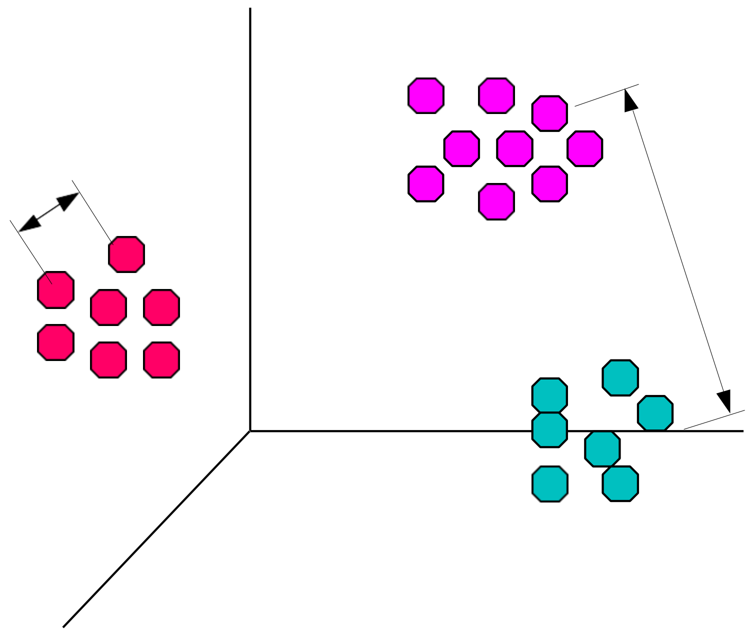
Clustering

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

Intracluster distances are minimized



Intercluster distances are maximized

Euclidean distance based clustering in 3-D space.

Clustering: Market Segmentation

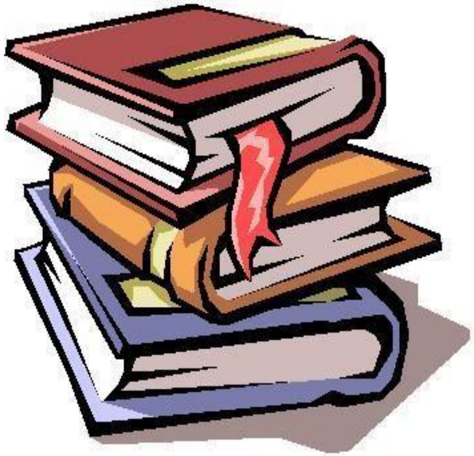
Goal: subdivide a market into distinct subsets of customers. Use a different marketing mix for each segment.

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 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.

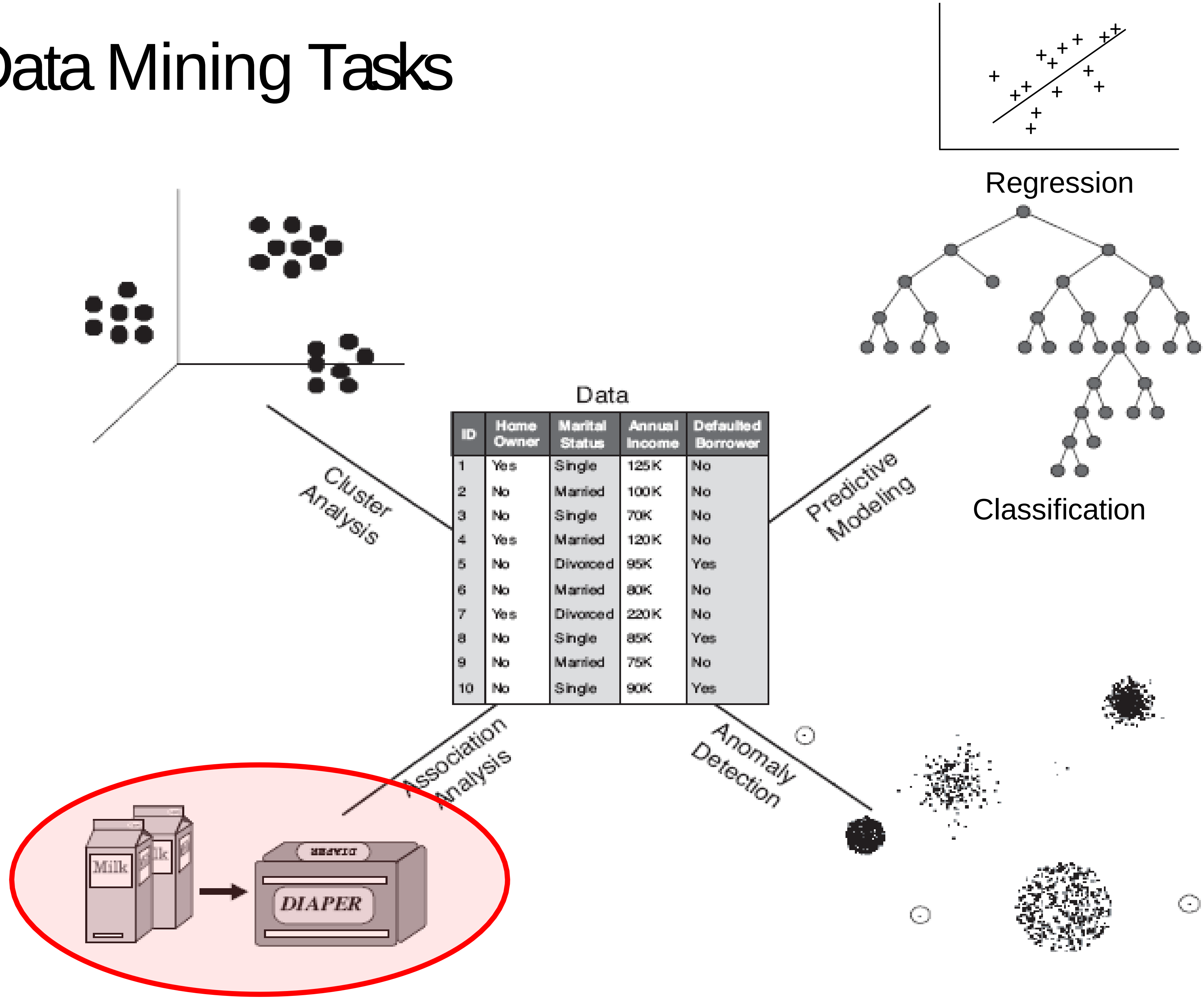
Gain: Can be used to organize documents or to create recommendations.

Clustering: 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 Tasks



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Association Rule Discovery

- Given is a set of transactions. Each contains a number of items.
- Produce dependency rules of the form

  LHS → 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

{Milk} → {Coke}

{Diaper, Milk} → {Beer}

Discovered Rules

Association Rule Discovery Marketing and Sales Promotion

- Let the rule discovered be

  {Potato Chips, … } → {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!

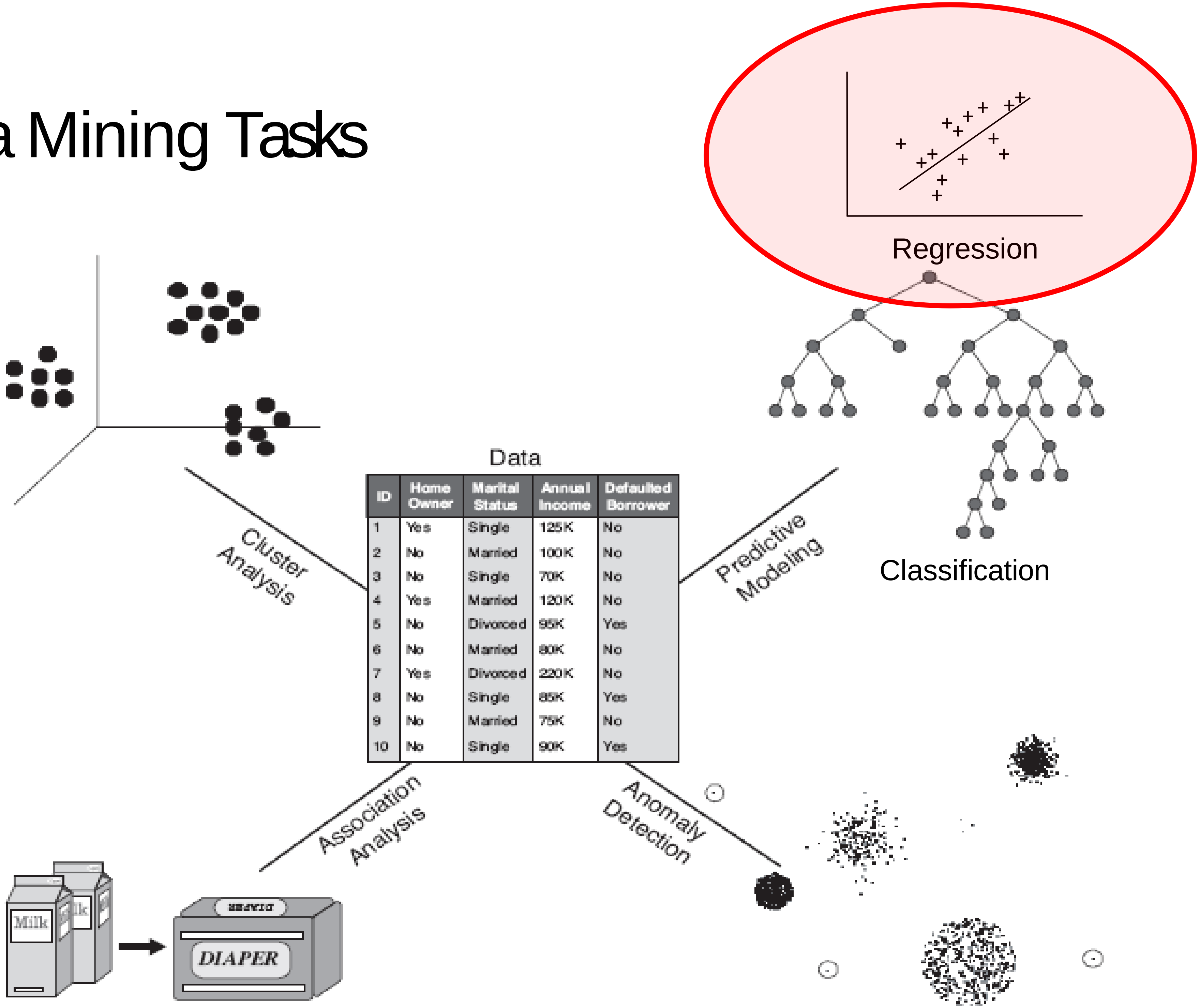
Association Rule Discovery Supermarket shelf management

- **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.

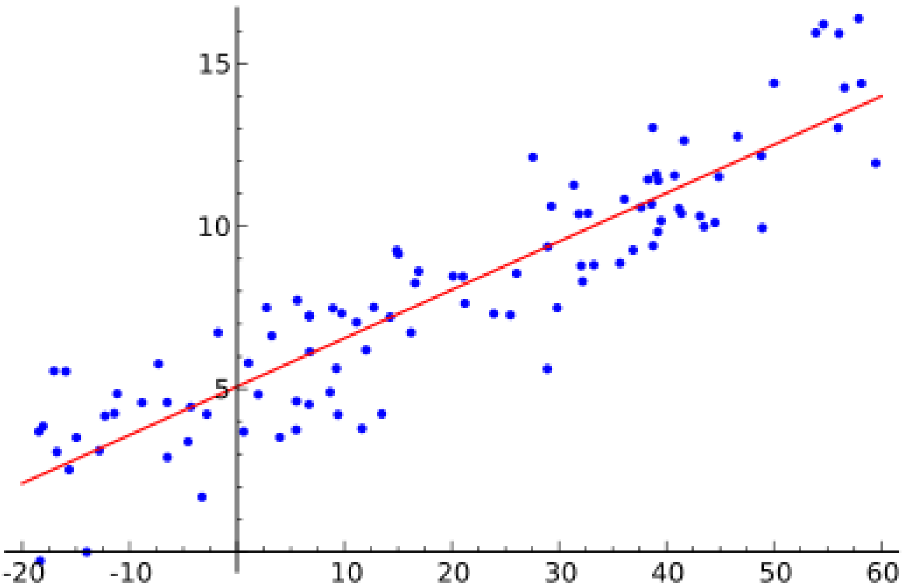
Data Mining Tasks



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

Regression

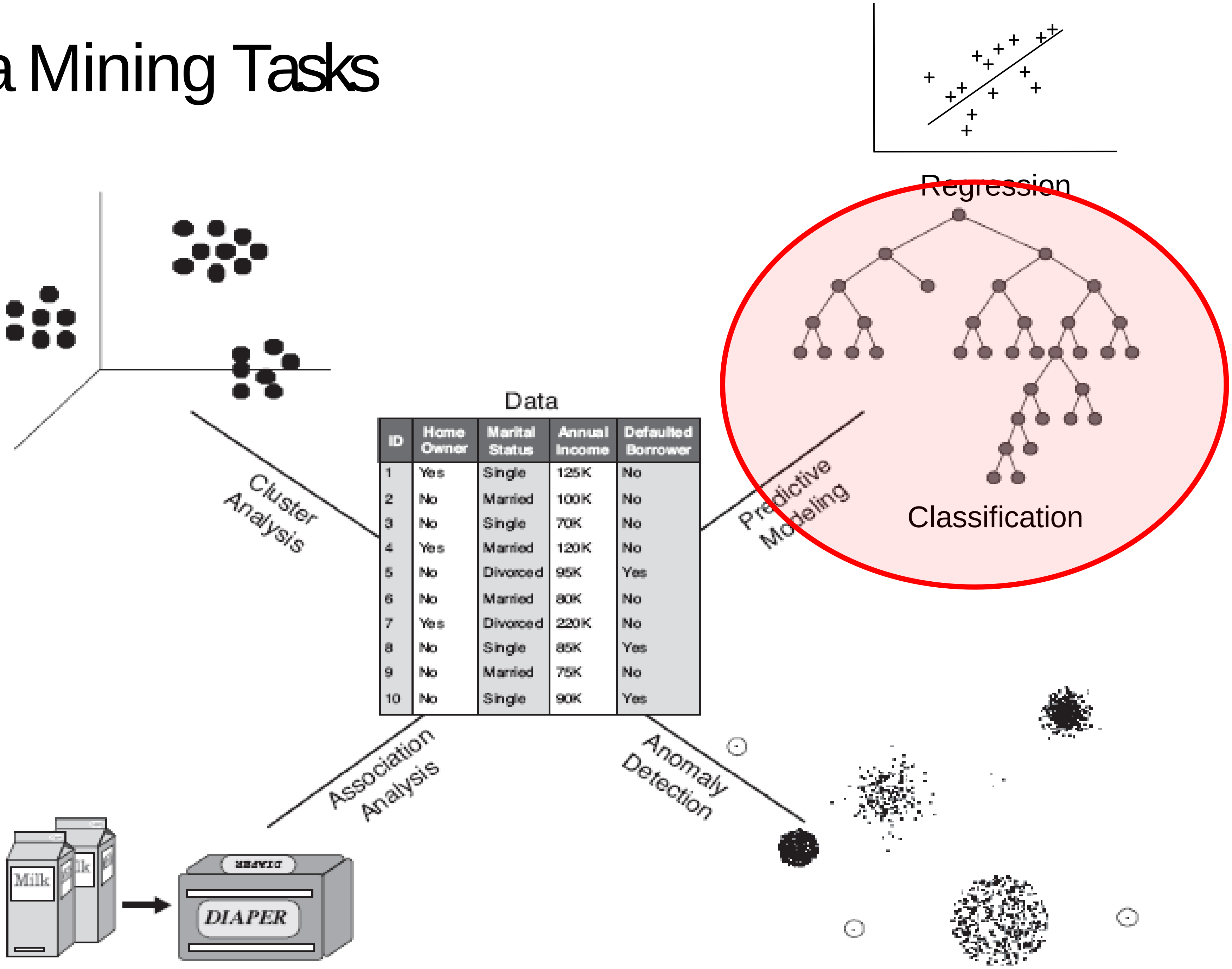
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Studied in statistics and econometrics.



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 Tasks



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

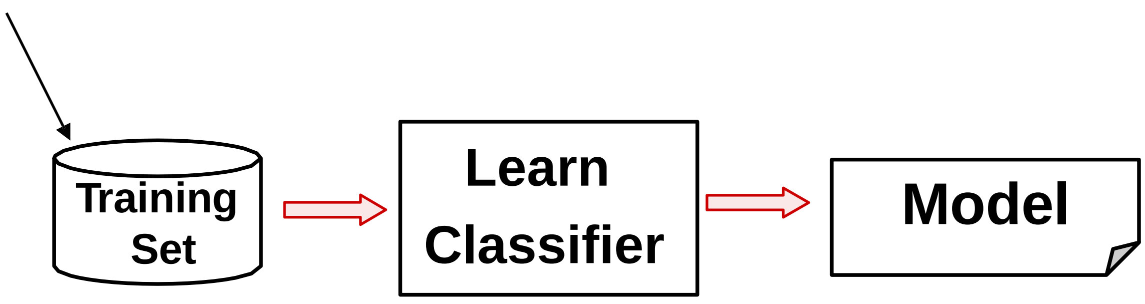
Classification

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

Class information is available → **Supervised Learning**

class

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



Classification

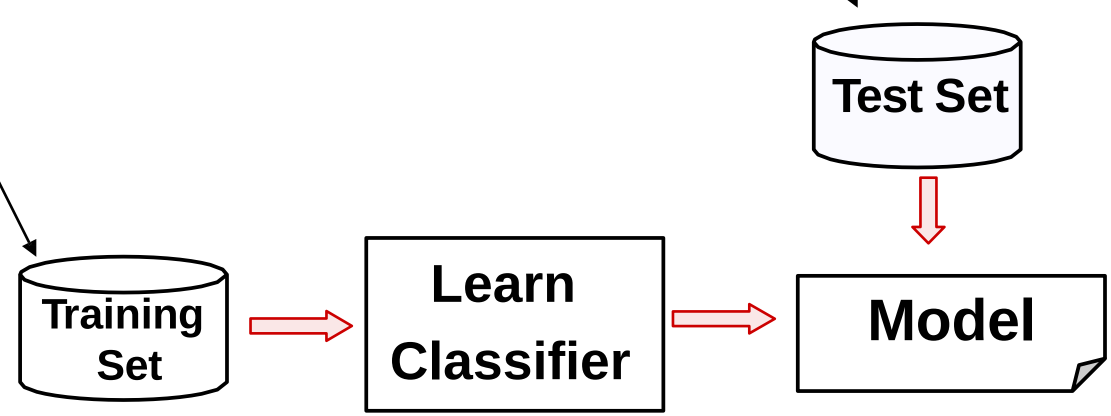
Find a **model** for the class attribute as a function of the values of other attributes/features.

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

class

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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Direct Marketing

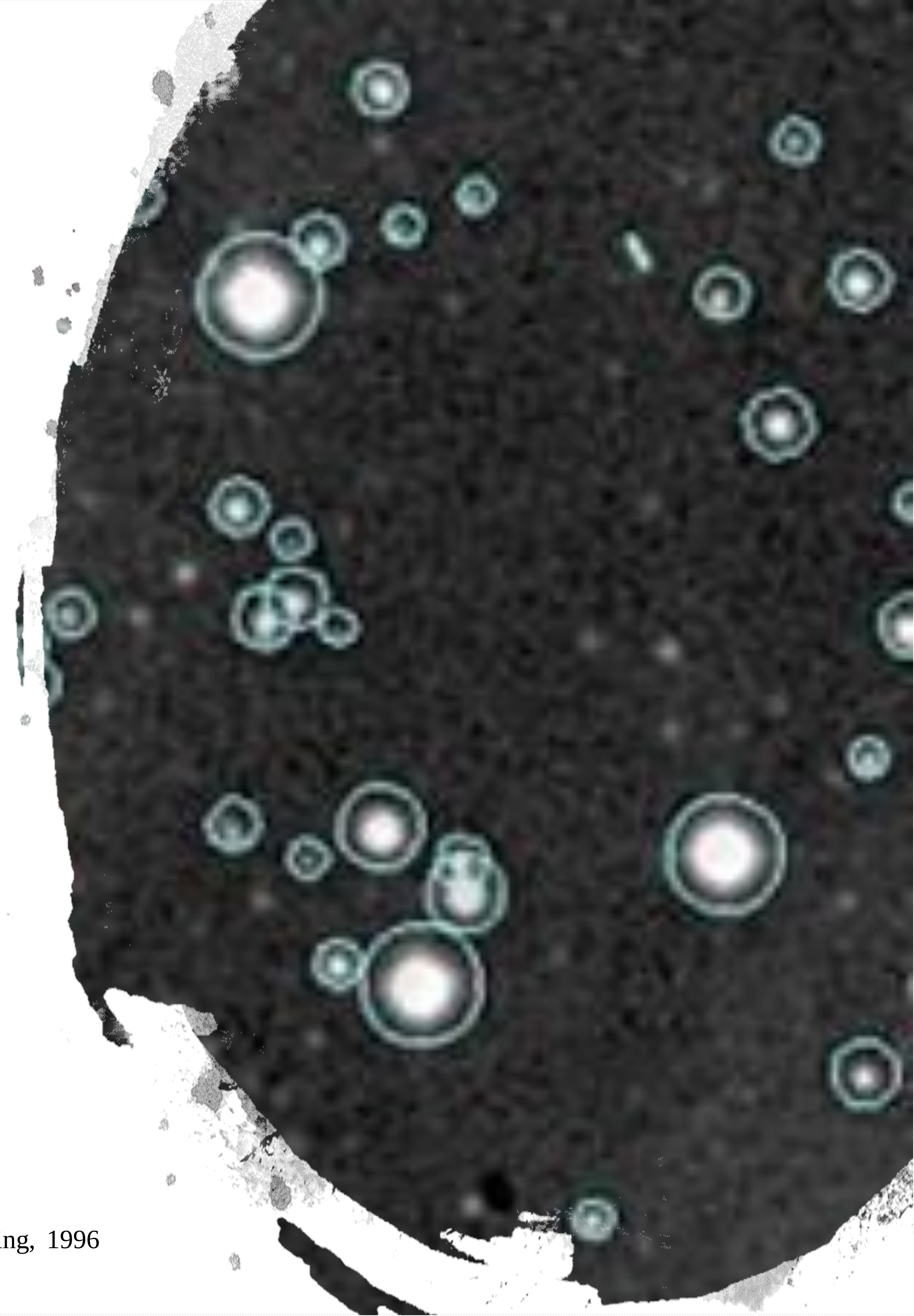
- Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new product.
- Approach:
  - Use the data for a similar product introduced before or from a focus group. We have customer information (e.g., demographics, lifestyle, previous purchases) and know which customers decided to 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: 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).

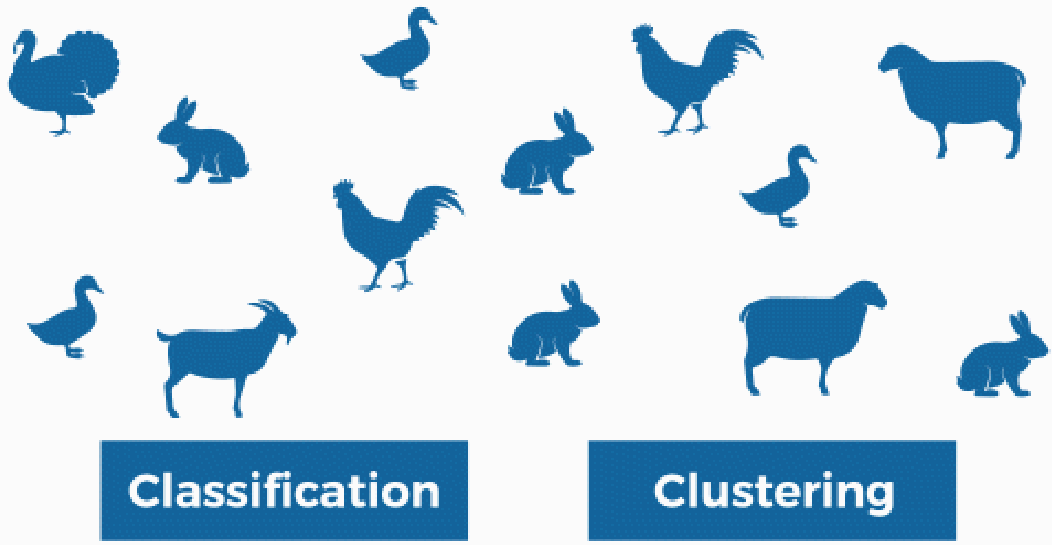
Classification: Sky Survey Cataloging

- 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.

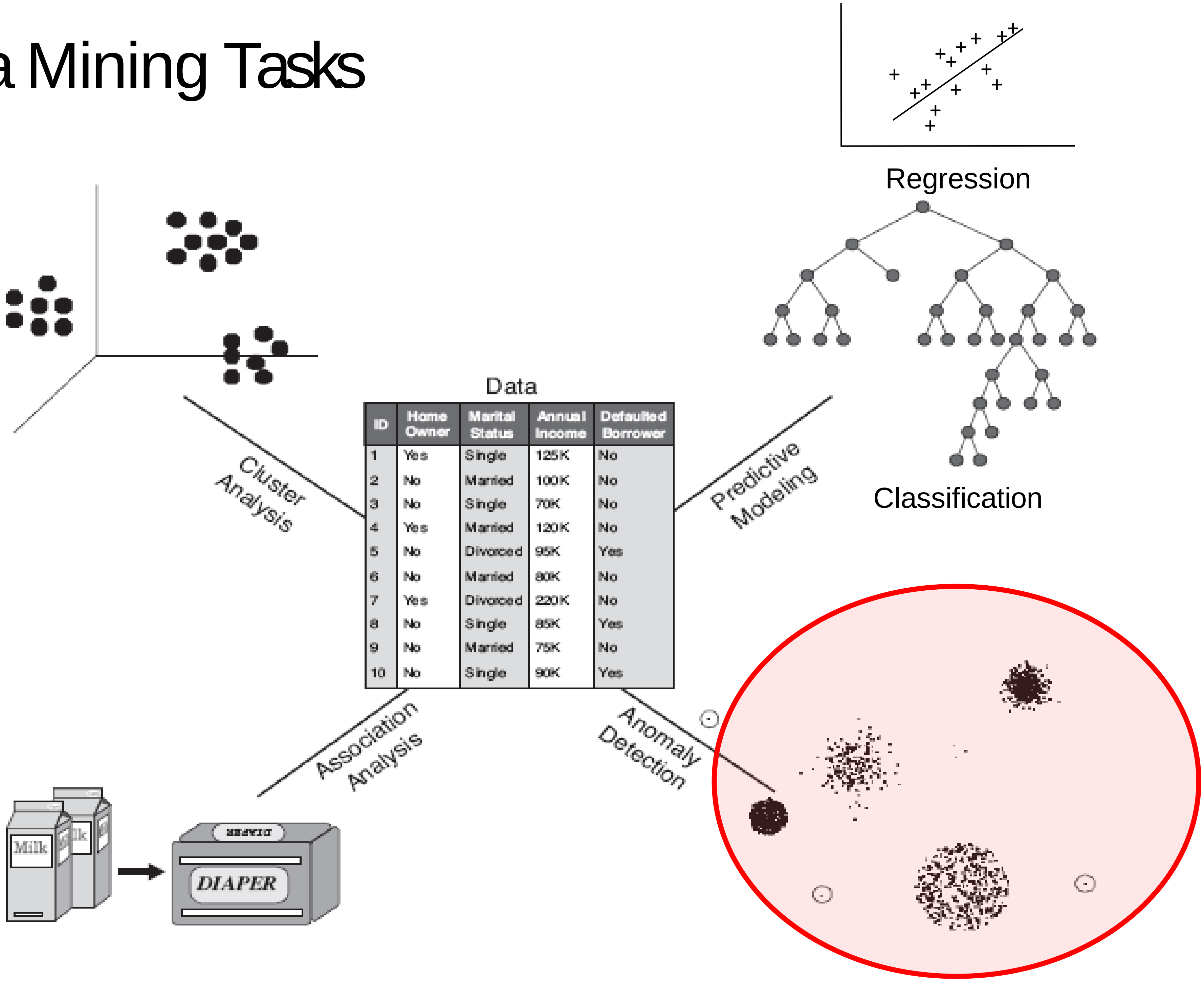


From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classification vs



Data Mining Tasks



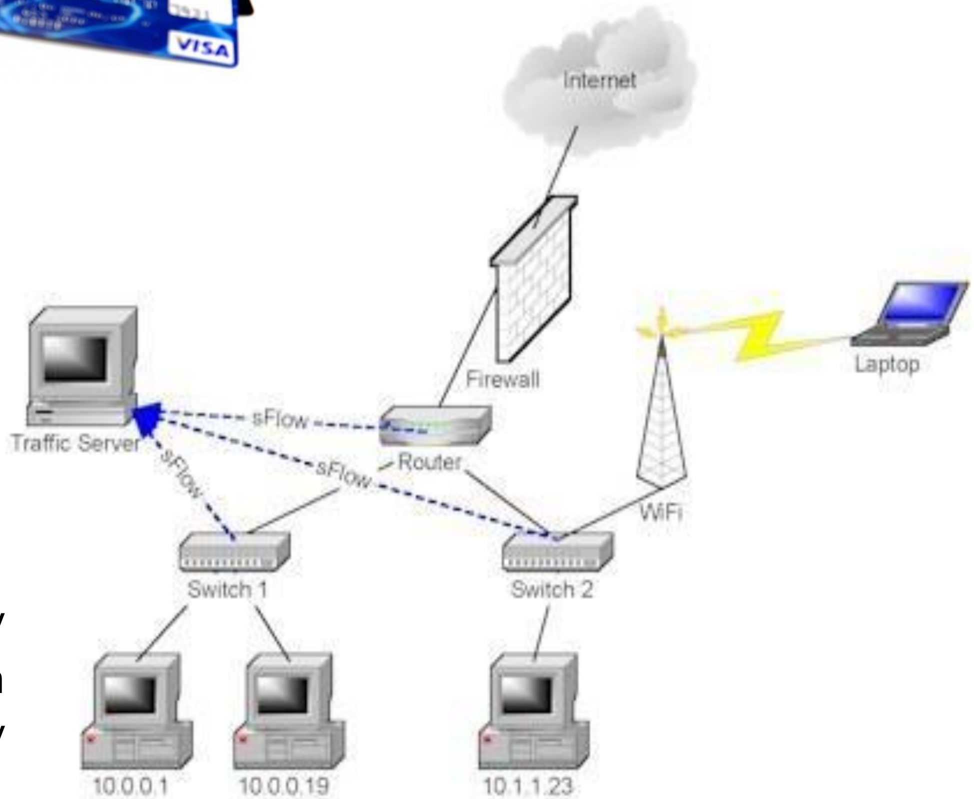
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior.

- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection

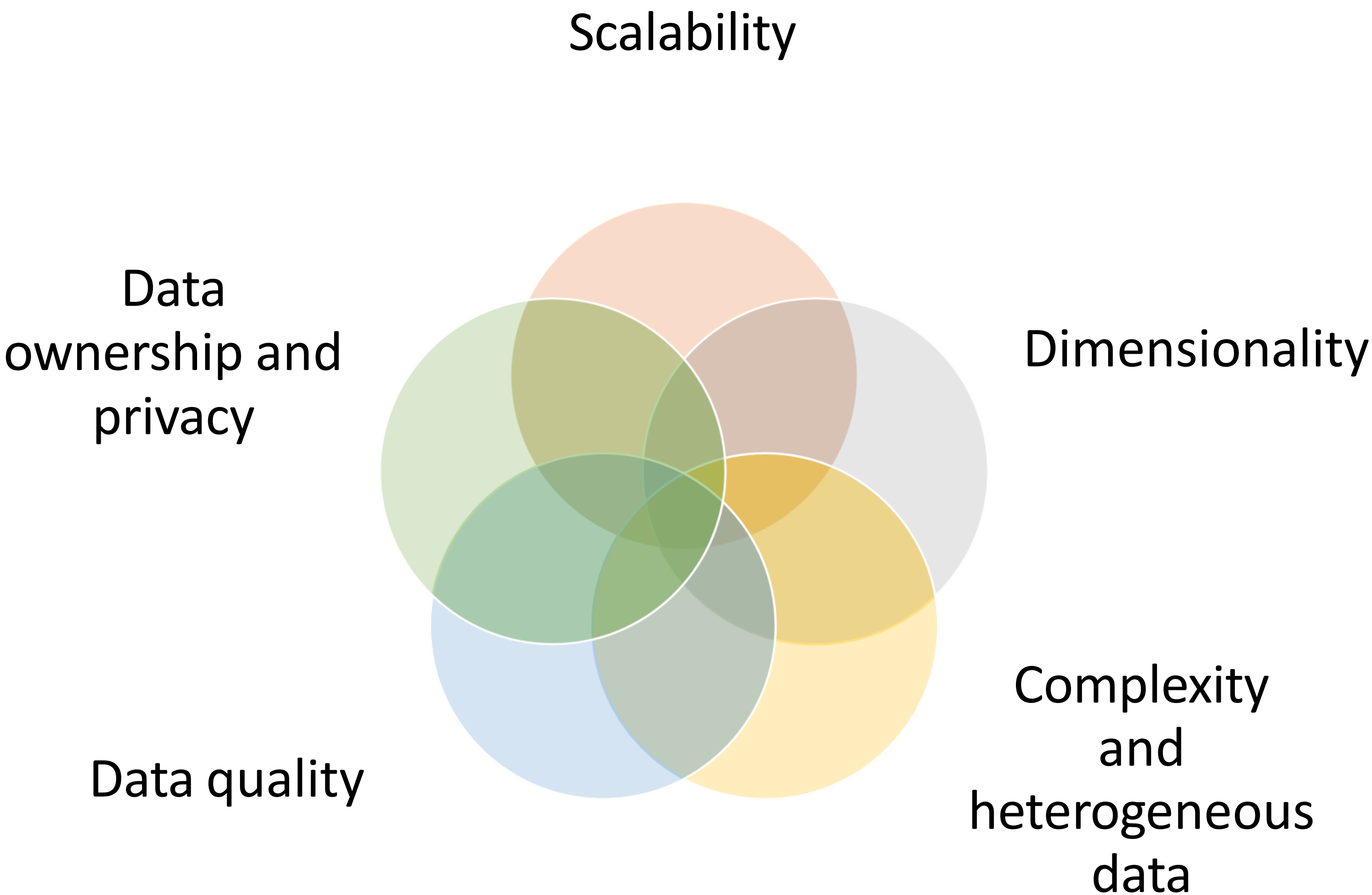


Typical network traffic at University level may reach over 100 million connections per day

Other Data Mining Tasks

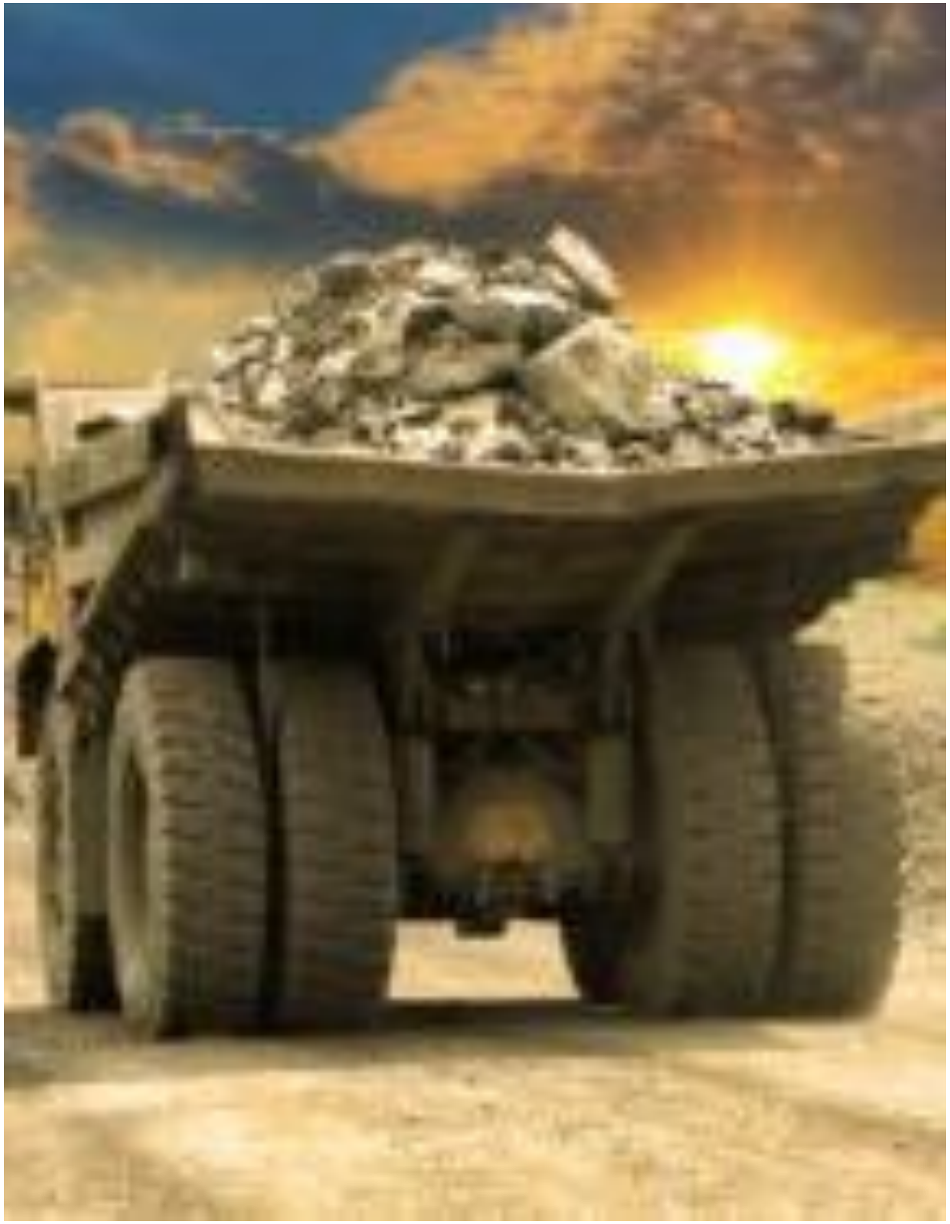
- Text mining – document clustering, topic models
- Graph mining – social networks
- Data stream mining/real time data mining
- Mining spatiotemporal data (e.g., moving objects)
- Visual data mining
- Distributed data mining

Challenges of Data Mining



Agenda

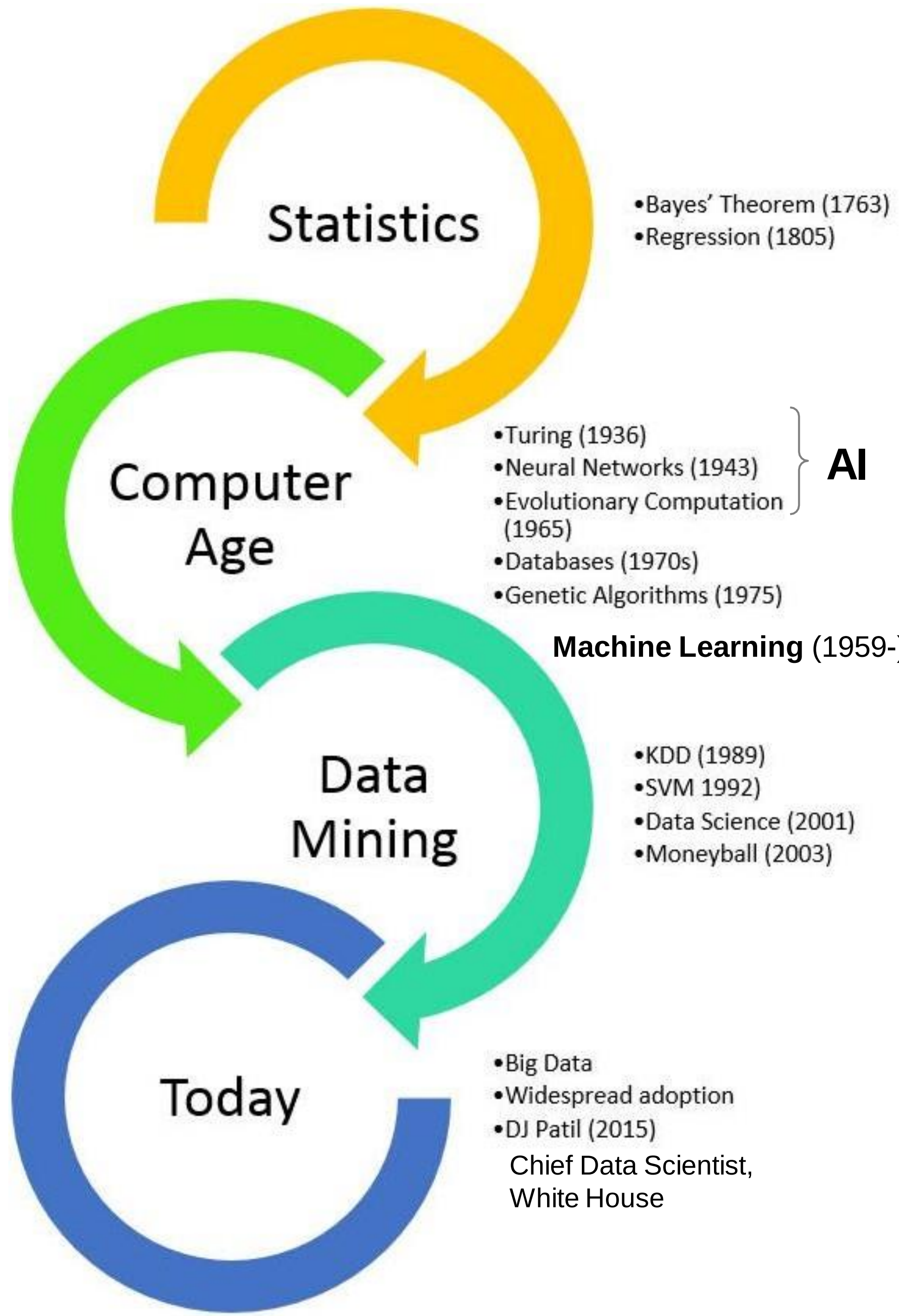
- What is Data Mining?
- Data Mining Tasks
- **Relationship to Statistics, Optimization, Machine Learning and AI**
- Tools
- Data
- Legal, Privacy and Security Issues



Origins of Data Mining

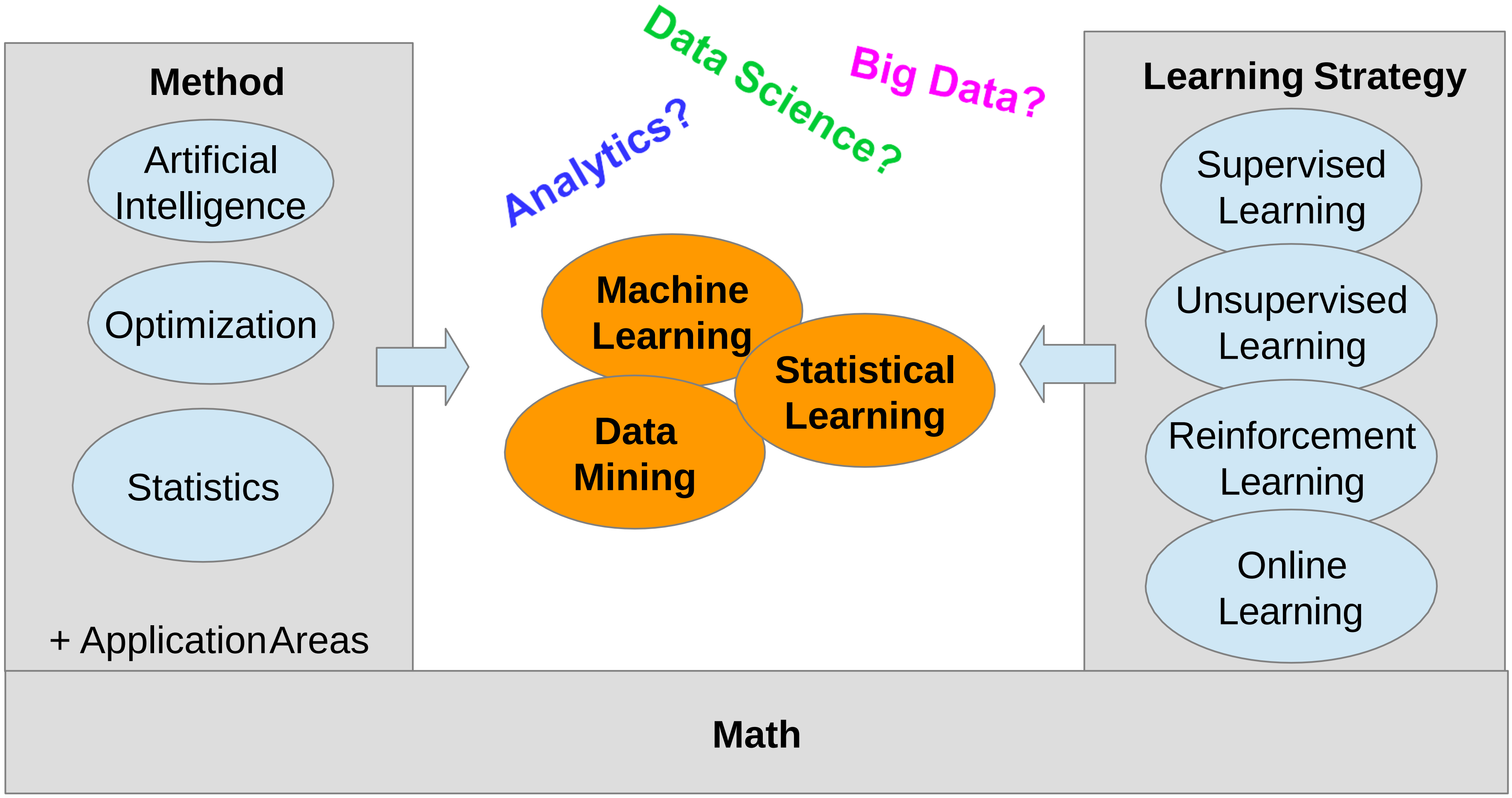
- Draws ideas from AI, machine learning, pattern recognition, statistics, and database systems.

- There are differences in terms of
  - —used data and
  - —the goals.

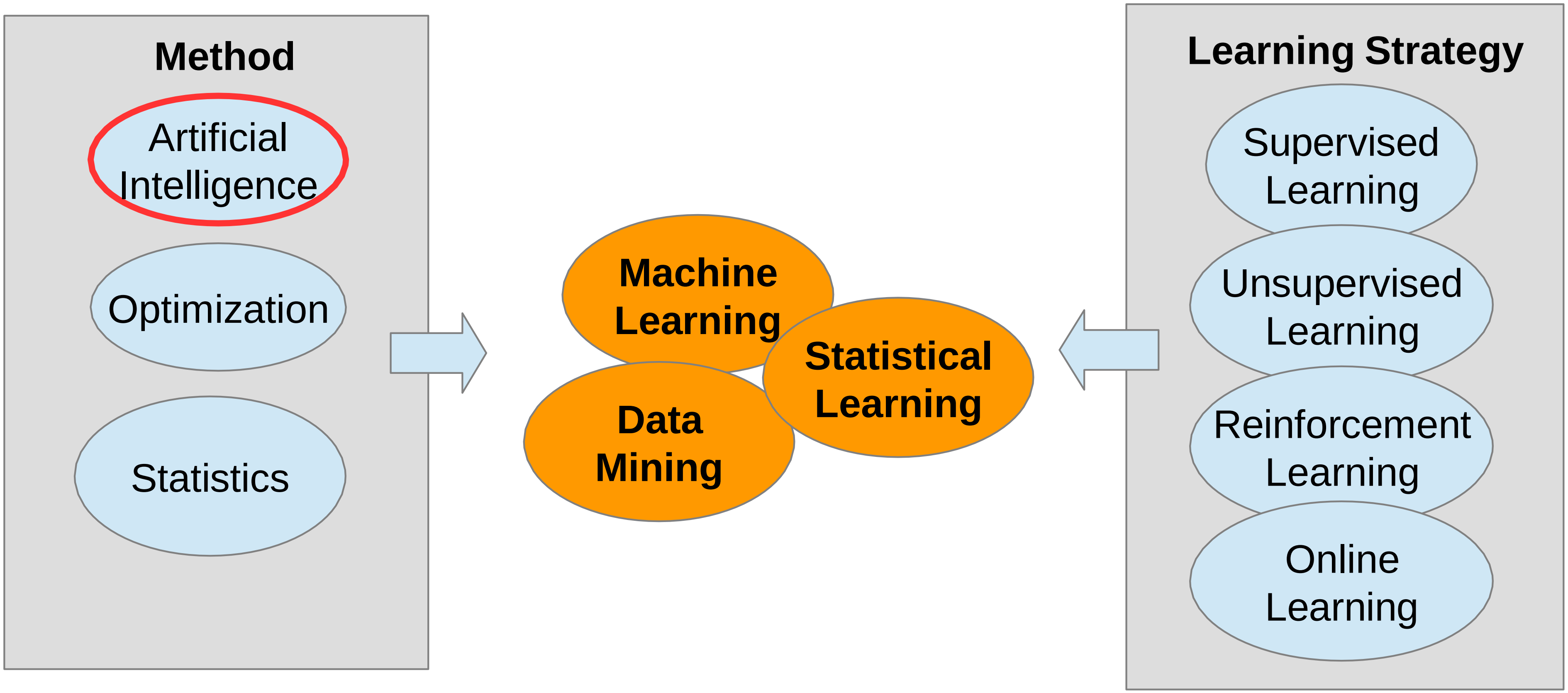


https://rayli.net/blog/data/history-of-data-mining/

Relationship to other Fields

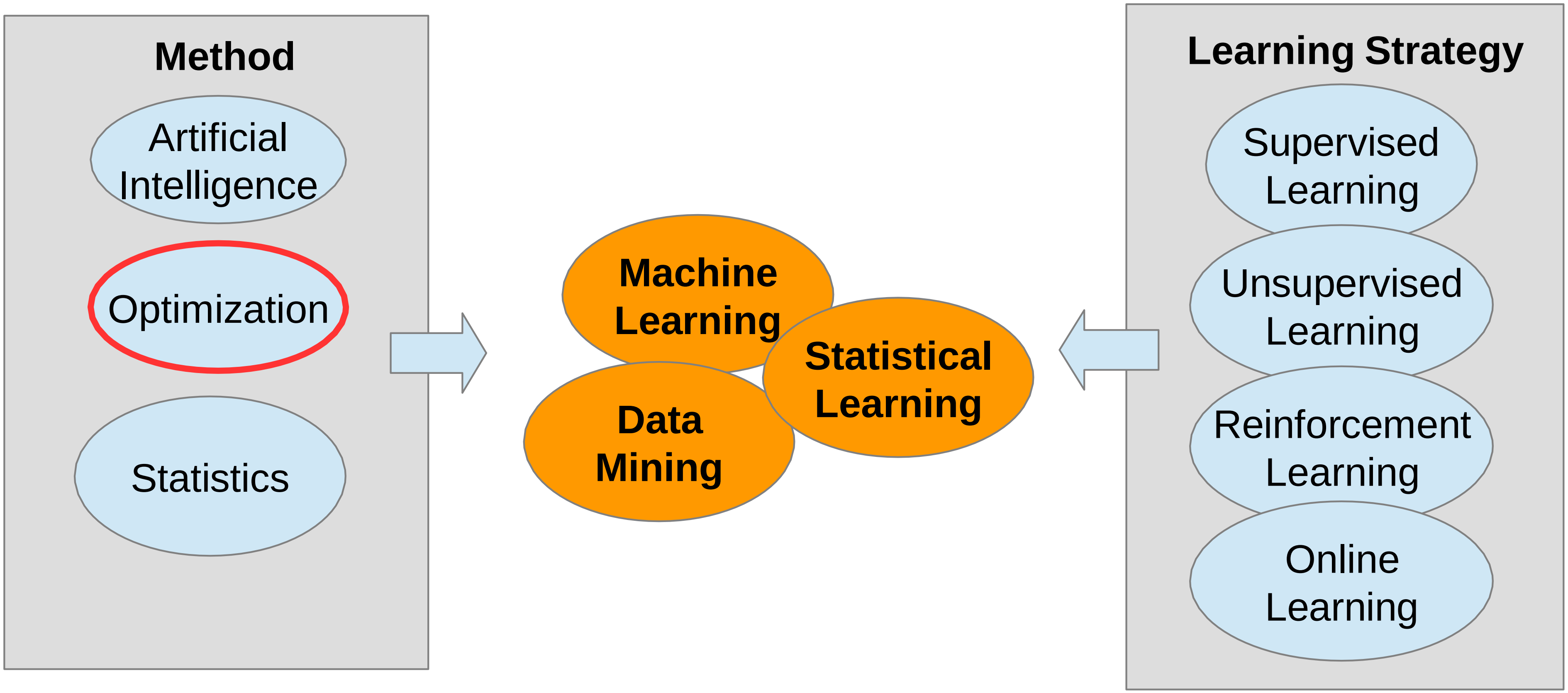


Relationship to other Fields



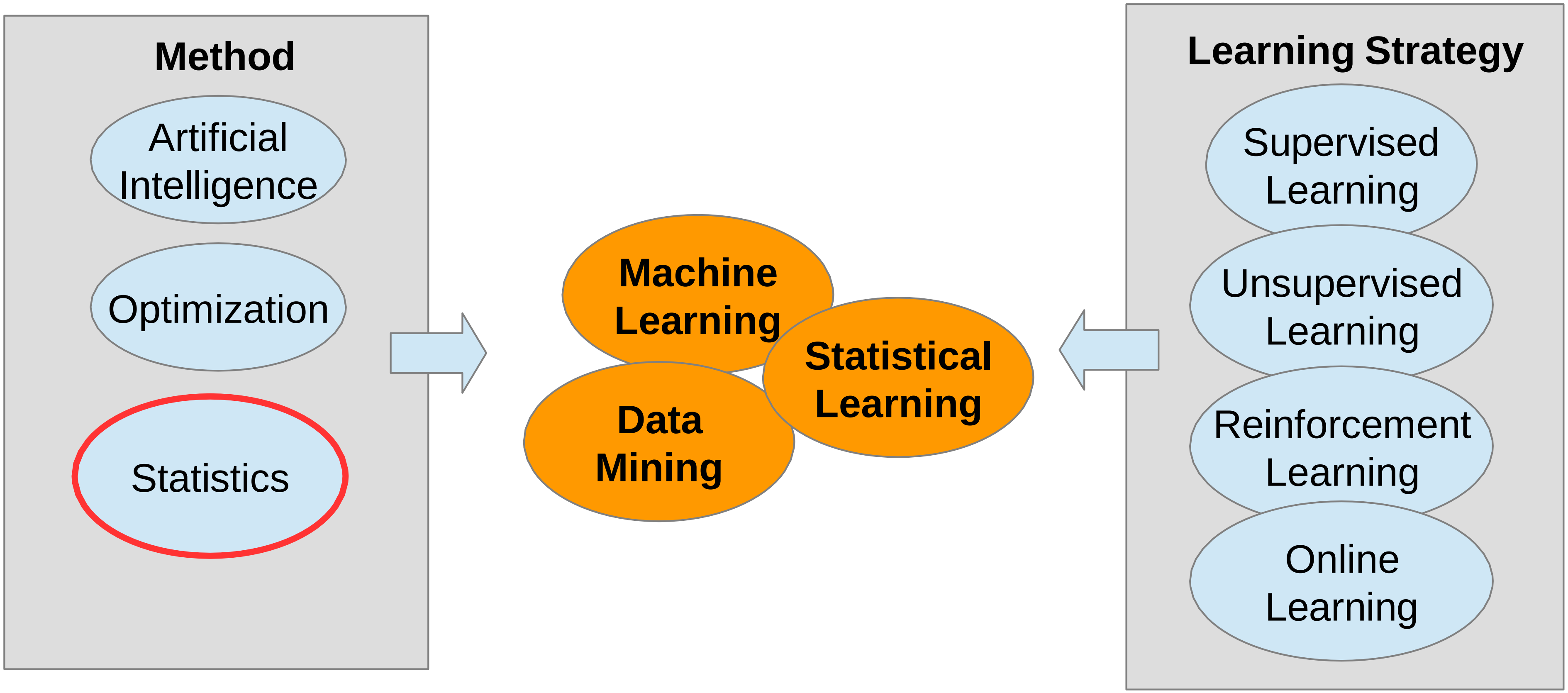
Artificial Intelligence: Create an **autonomous agent** that perceives its environment and takes actions that maximize its chance of reaching some goal.
Areas: reasoning, knowledge representation, planning, learning, natural language processing, and vision.

Relationship to other Fields



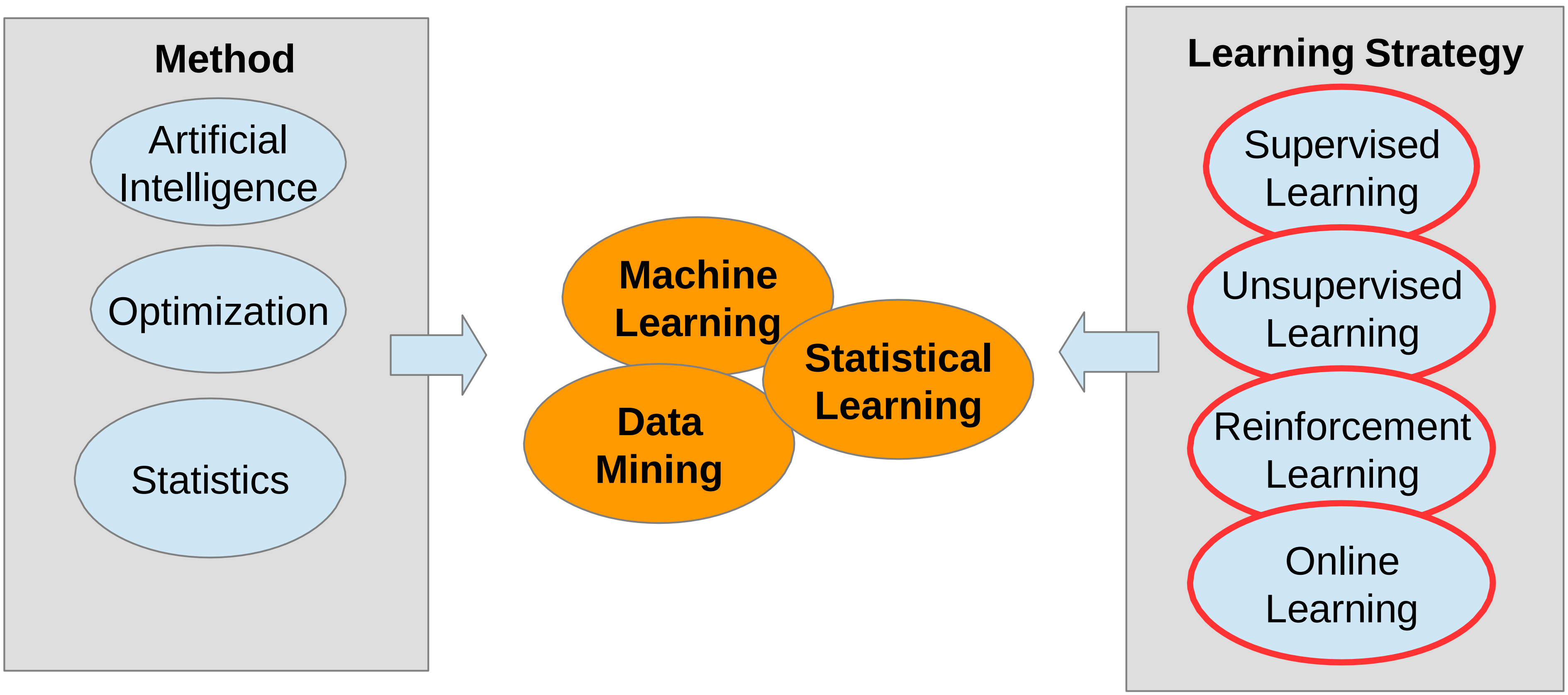
Optimization: Selection of a best alternative from some set of available alternatives with regard to some criterion.
Techniques: Linear programming, integer programming, nonlinear programming, stochastic and robust optimization, heuristics, etc.

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.

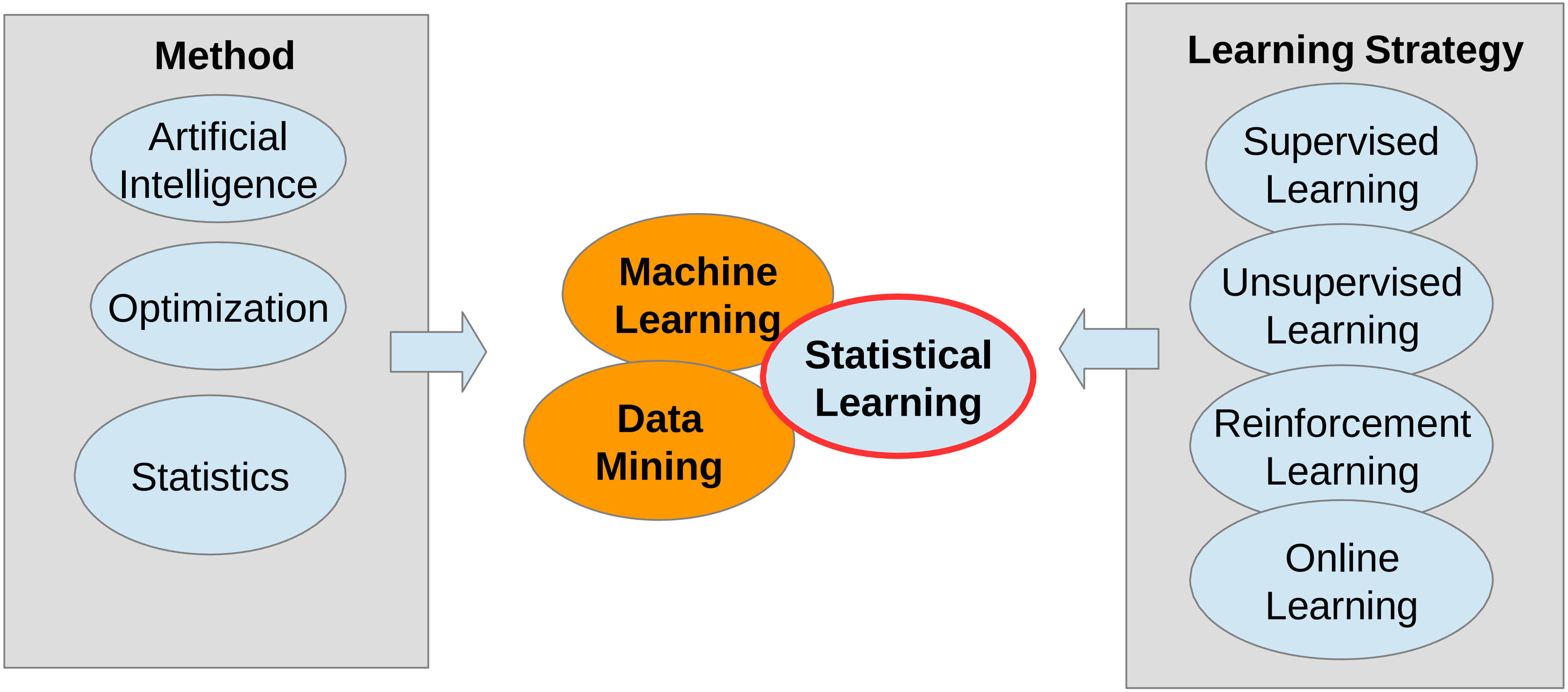
Relationship to other Fields



Learning Strategy: From what data do we learn?

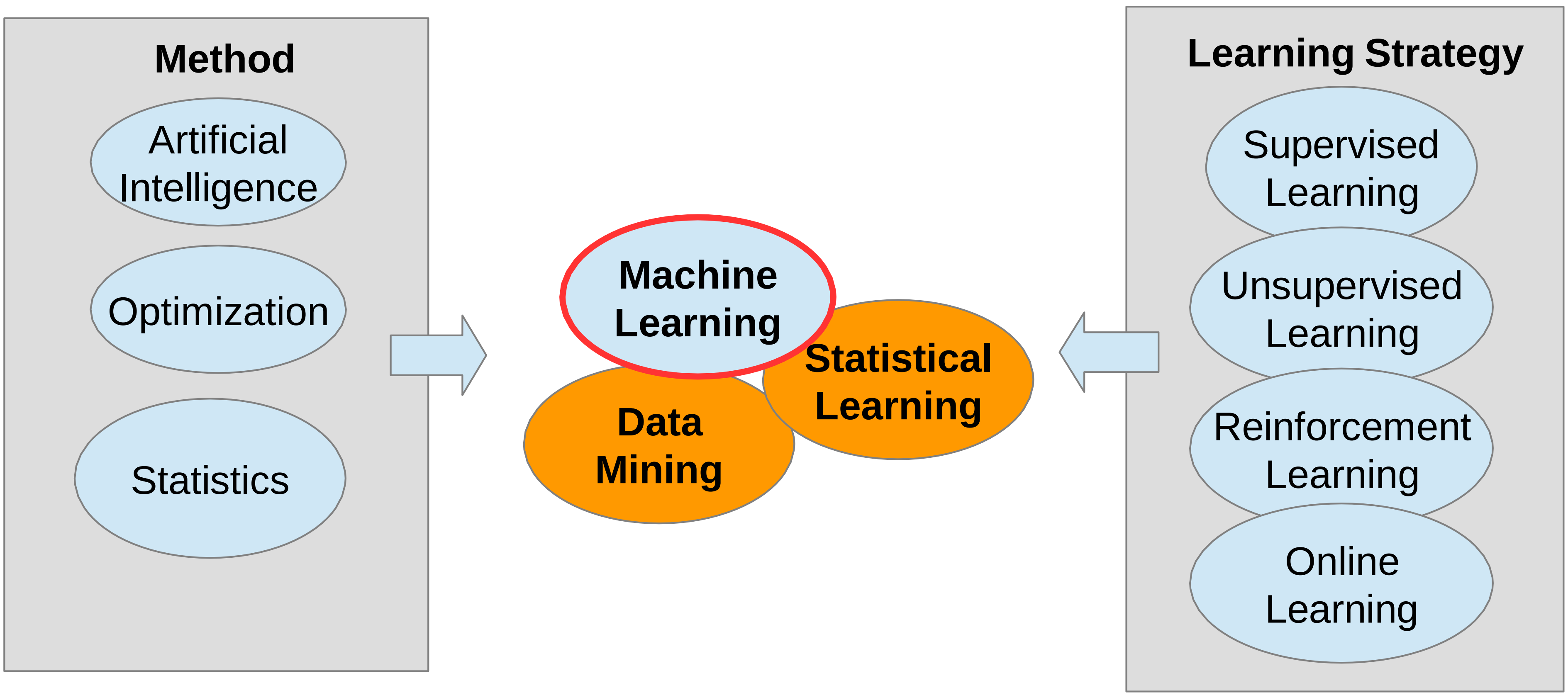
- Is a training set with correct answers available? → Supervised learning
- Long-term structure of rewards? → Reinforcement learning
- No answer and no reward structure? → Unsupervised learning
- Do we have to update the model regularly? → Online learning

Relationship to other Fields



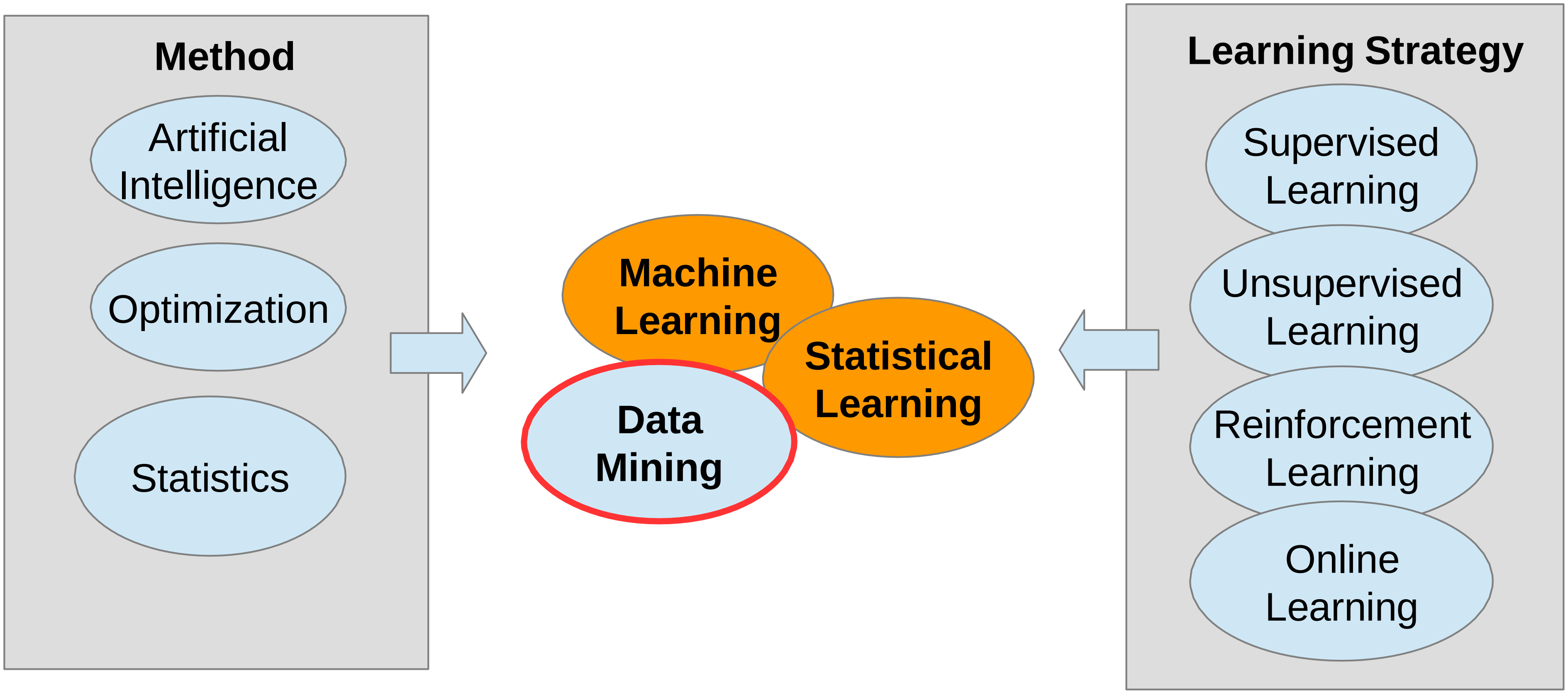
Statistical learning: deals with the problem of finding a **predictive function** based on data.
Tools: (Linear) classifiers, regression and regularization.

Relationship to other Fields



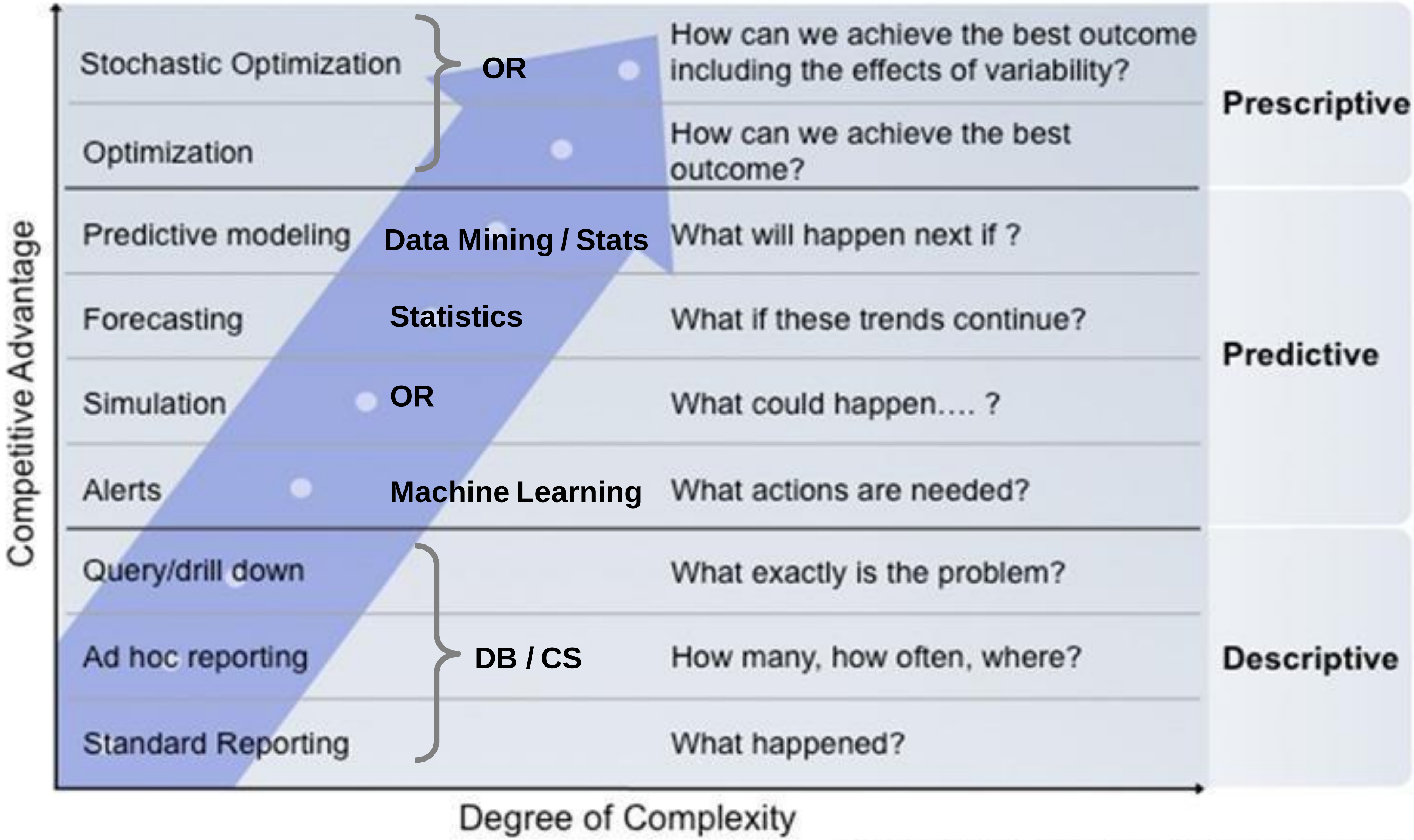
Machine Learning involves the study of algorithms that can extract information **automatically**, i.e., without on-line human guidance.
Techniques: Focus on supervised learning.

Relationship to other Fields



Data Mining: Manually analyze a given dataset to gain insights and predict potential outcomes.
Techniques: Any applicable technique from databases, statistics, machine/statistical learning. New methods were developed by the Data Mining community.

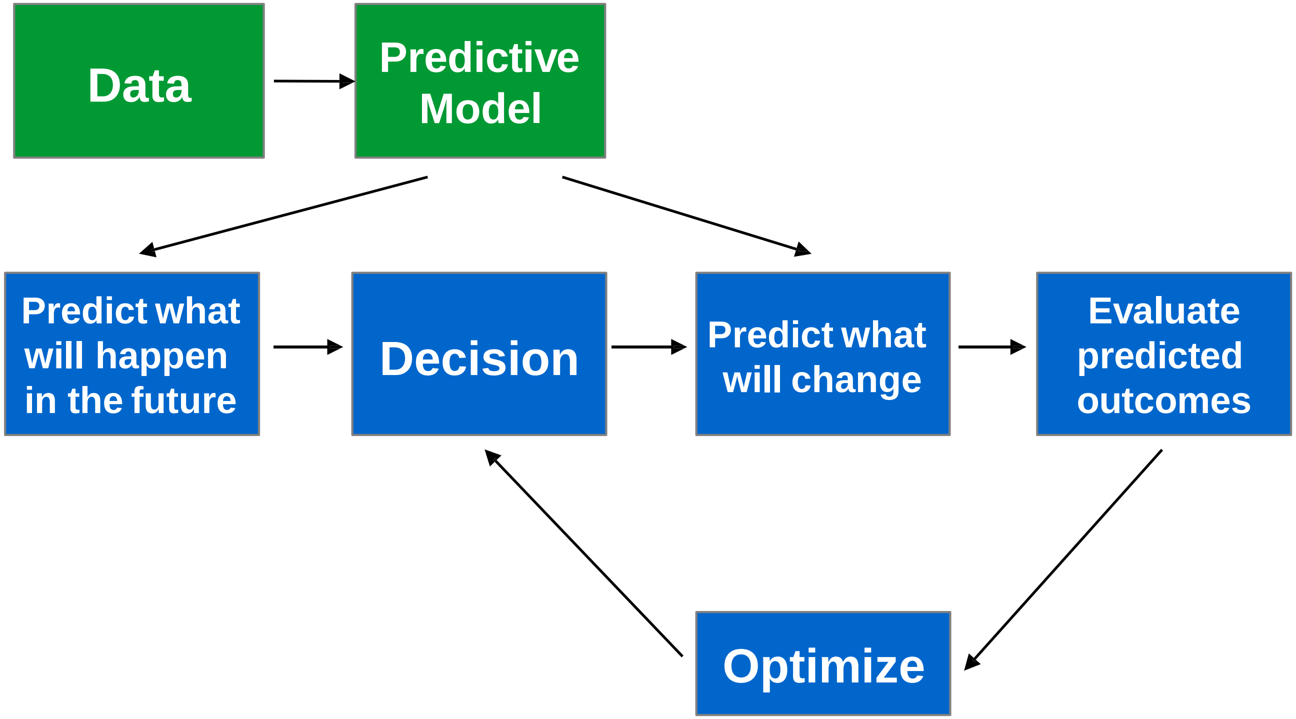
Data Mining & Analytics



Based on: Competing on Analytics, Davenport and Harris, 2007

Prescriptive Analytics

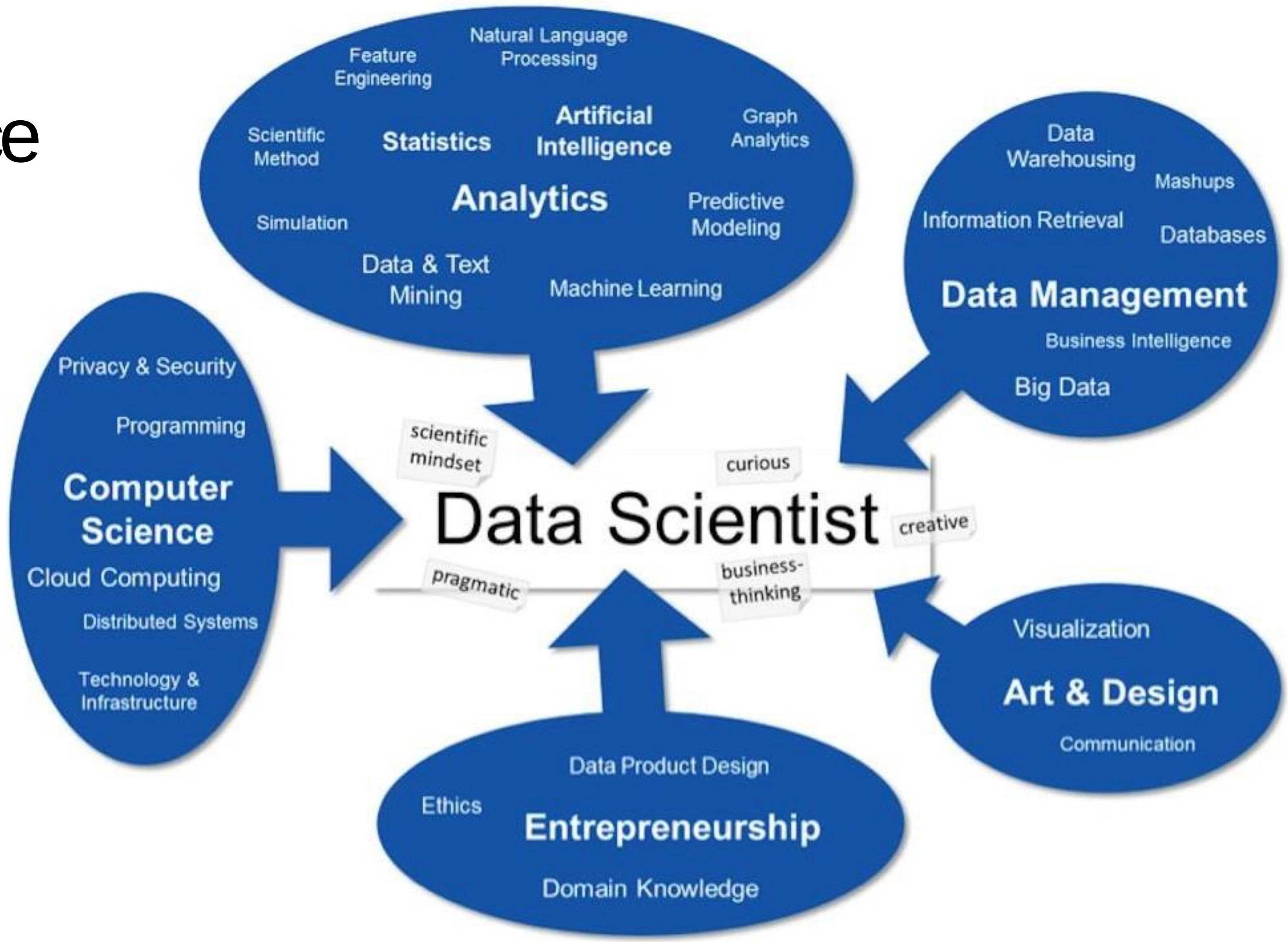
What decisions should we make now
to achieve the best future outcome?



Issues:

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

Data Science

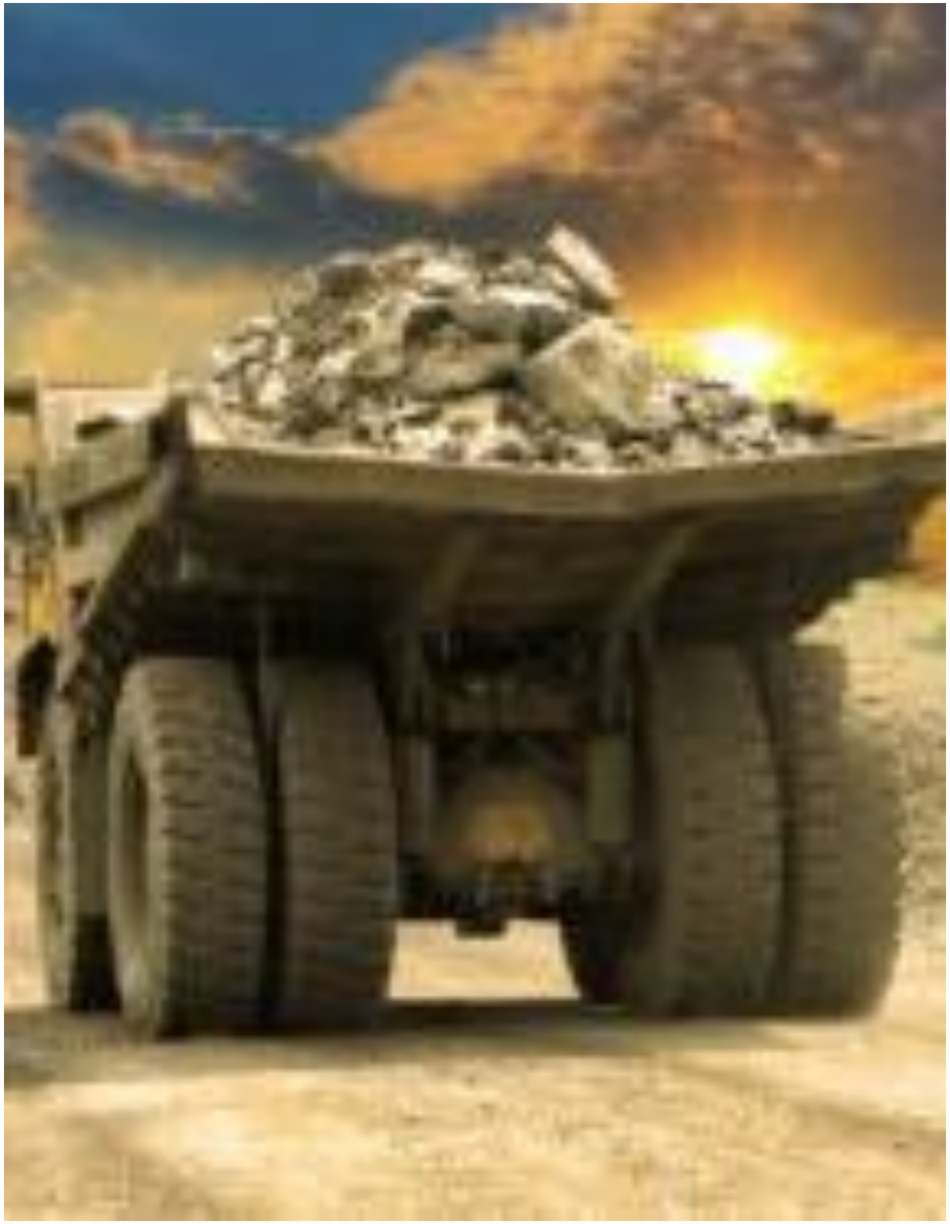


Source: T. Stadelmann, et al., Applied Data Science in Europe

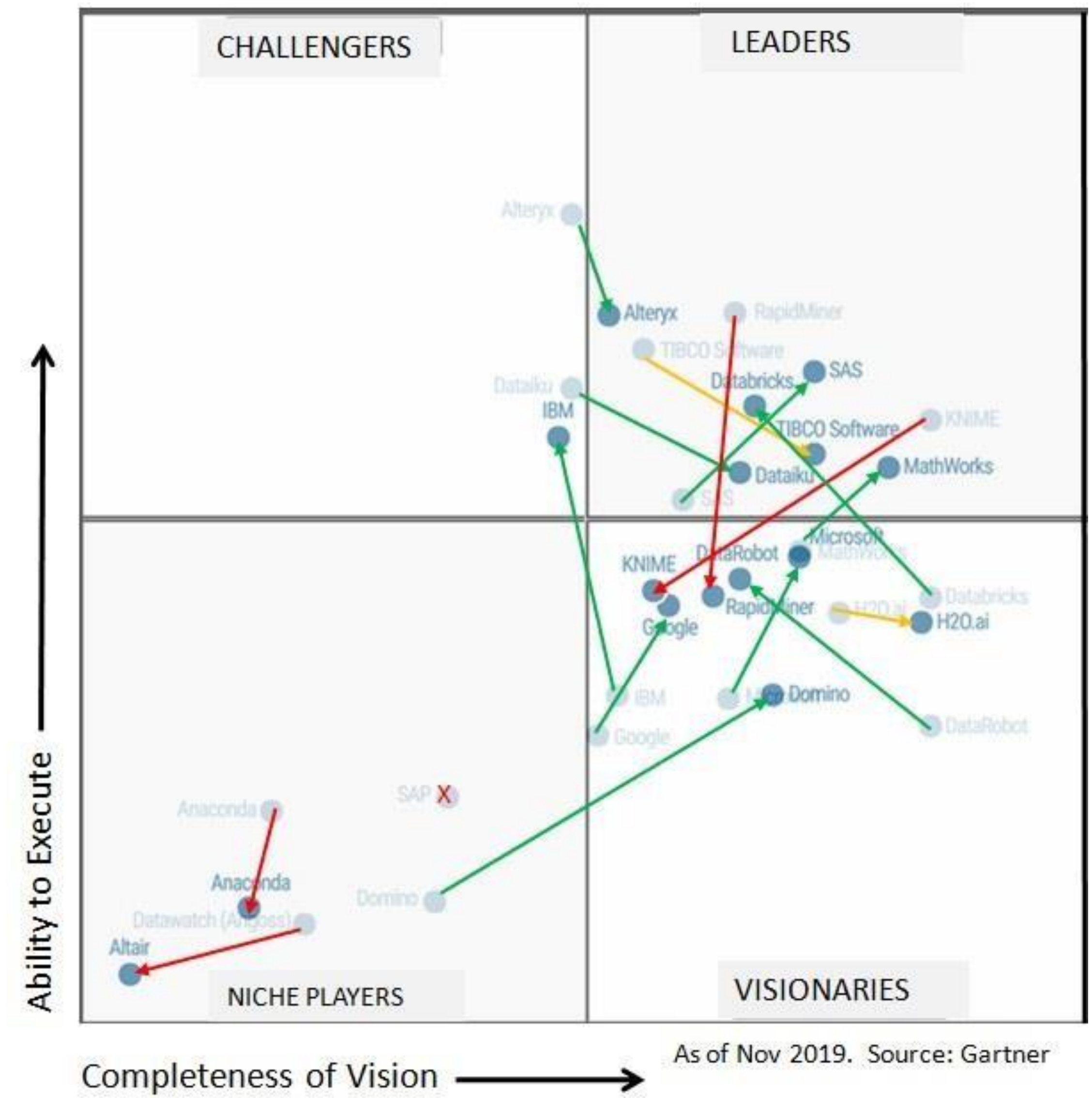
Good luck finding this person!
Probably a team effort!

Agenda

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

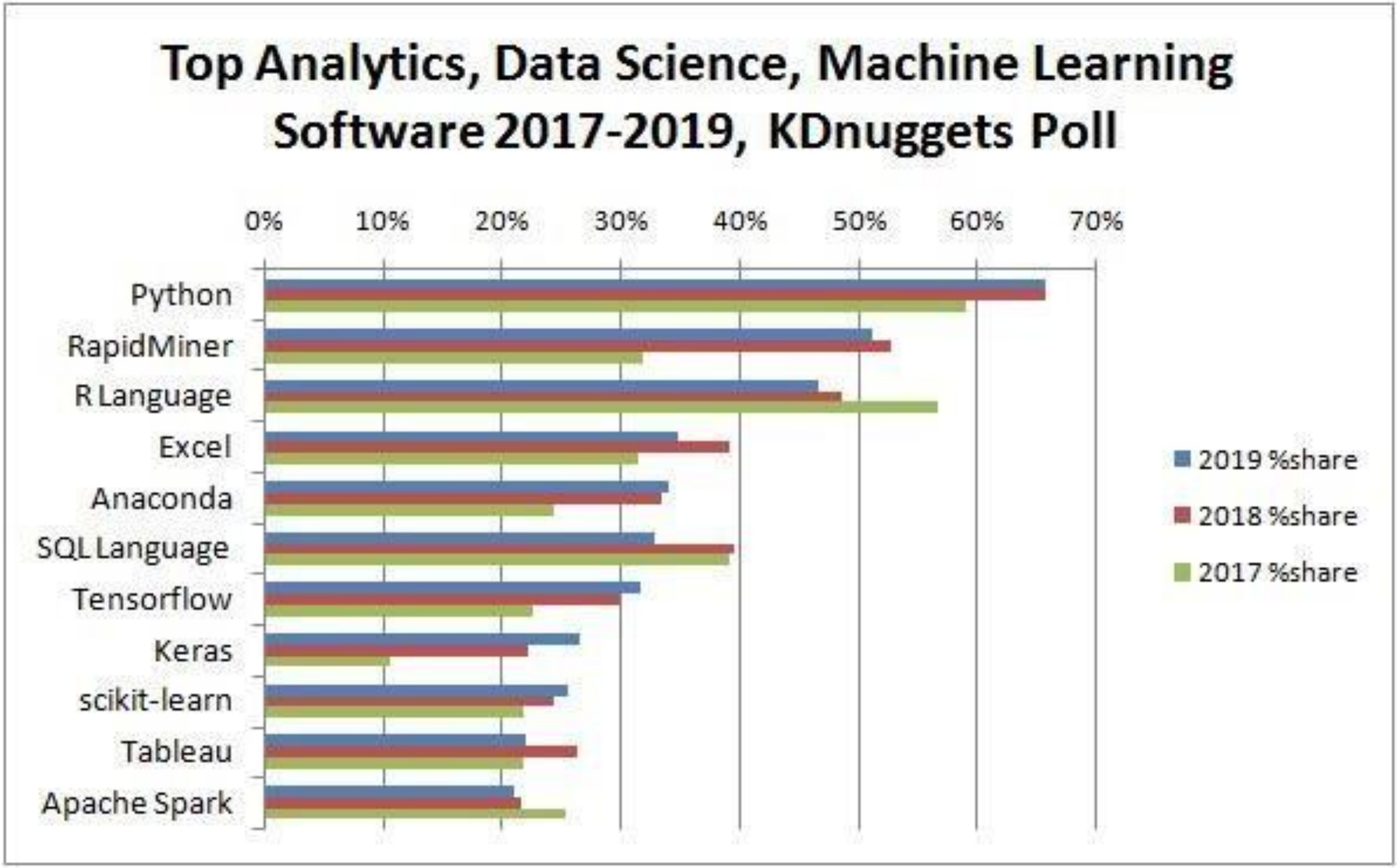


Tools: Commercial Players



Gartner MQ for Data Science and Machine Learning Platforms, 2020 vs 2019 changes.

Tools: Popularity



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

Tools: Types

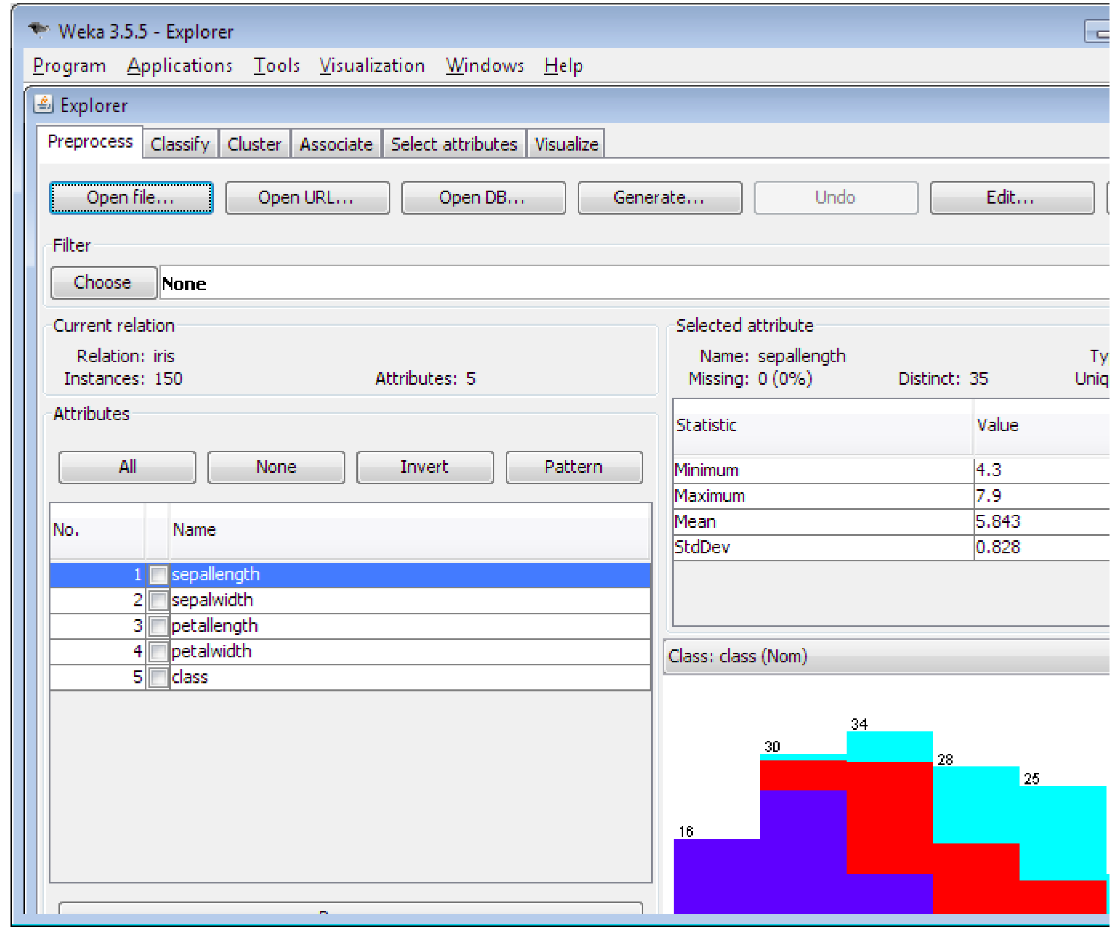
Simple graphical user interface

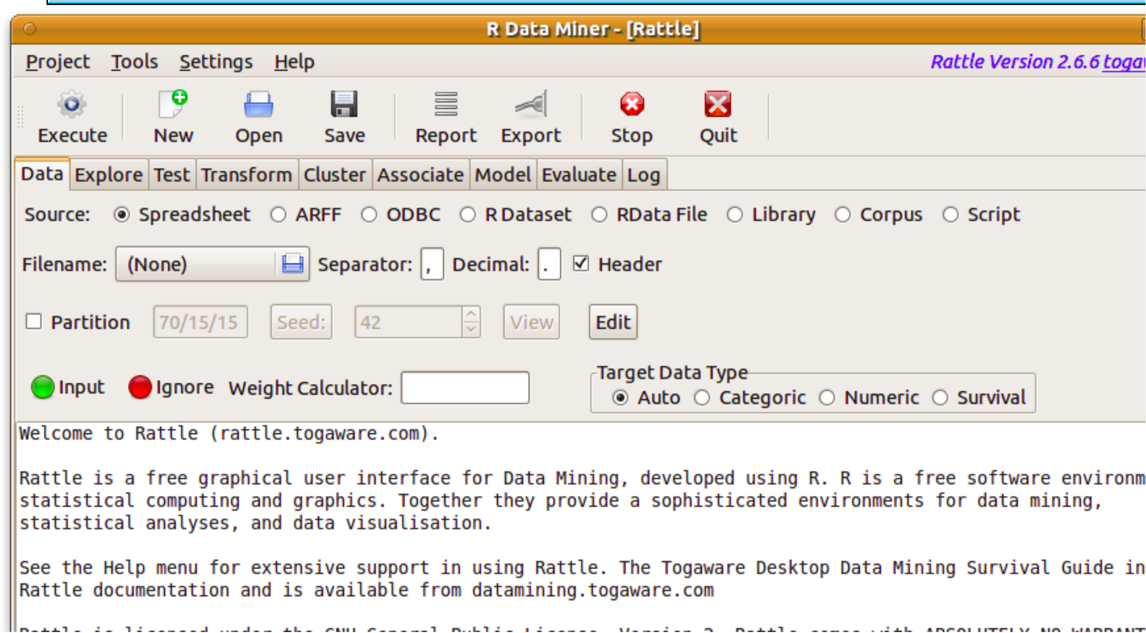
Process oriented

Programming oriented

Tools: Simple GUI

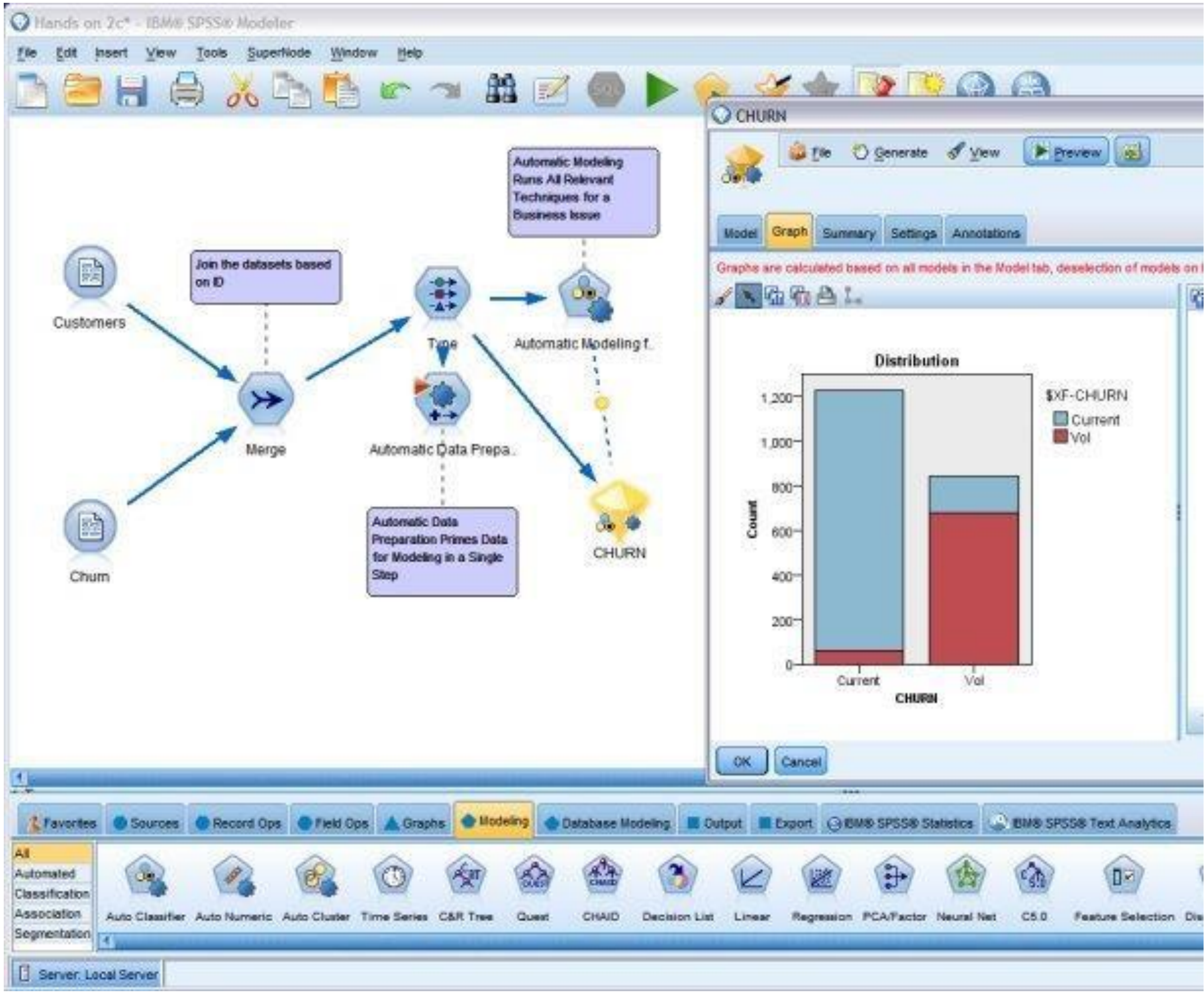
- Weka: Waikato Environment for Knowledge Analysis (Java API)
- Rattle: GUI for Data Mining using R





Tools: Process oriented

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



Tools: Programming oriented

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

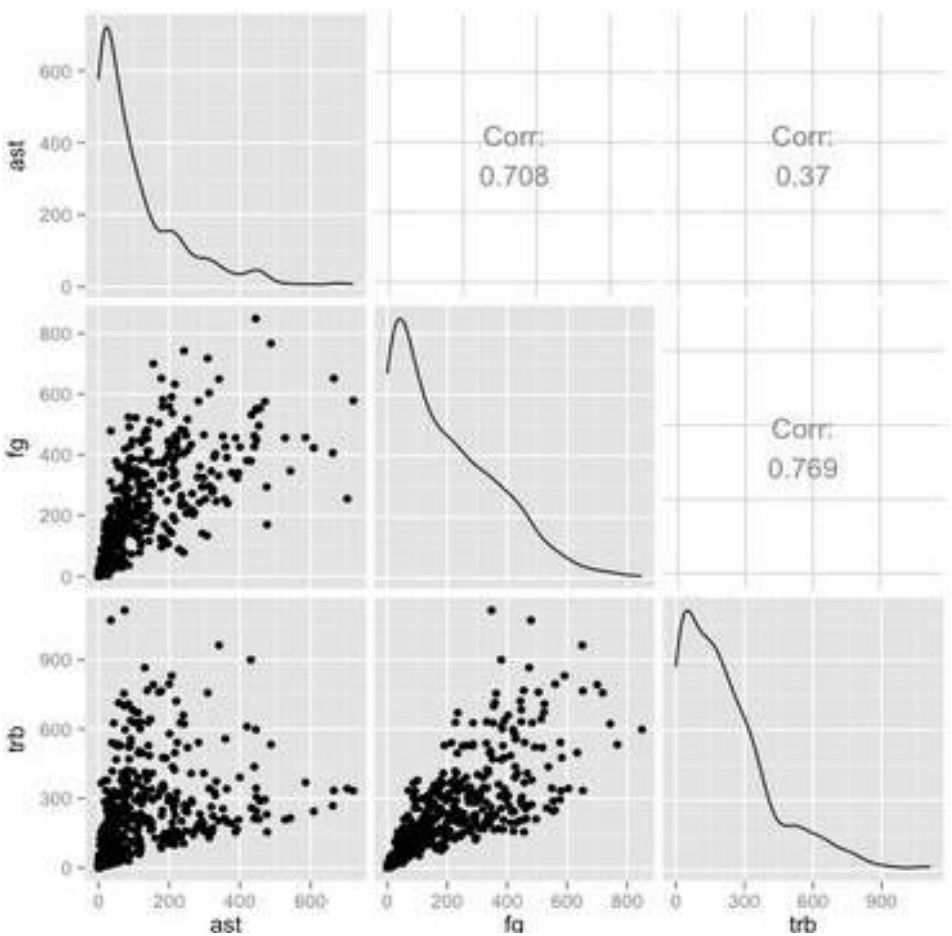
→ Both have similar capabilities. Slightly different focus:
  - R: statistical computing and visualization
  - Python: Scripting, big data
  - Interoperability via rpy2 and rediculate

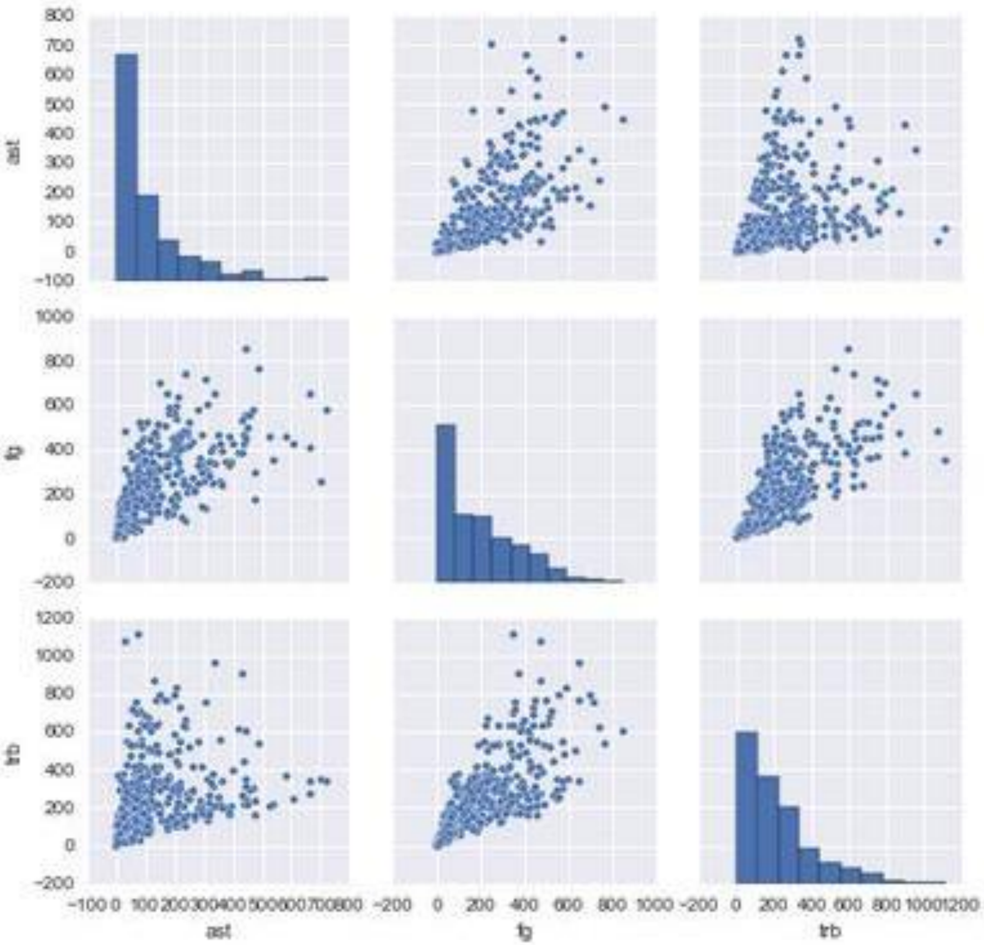
R

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

Python

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

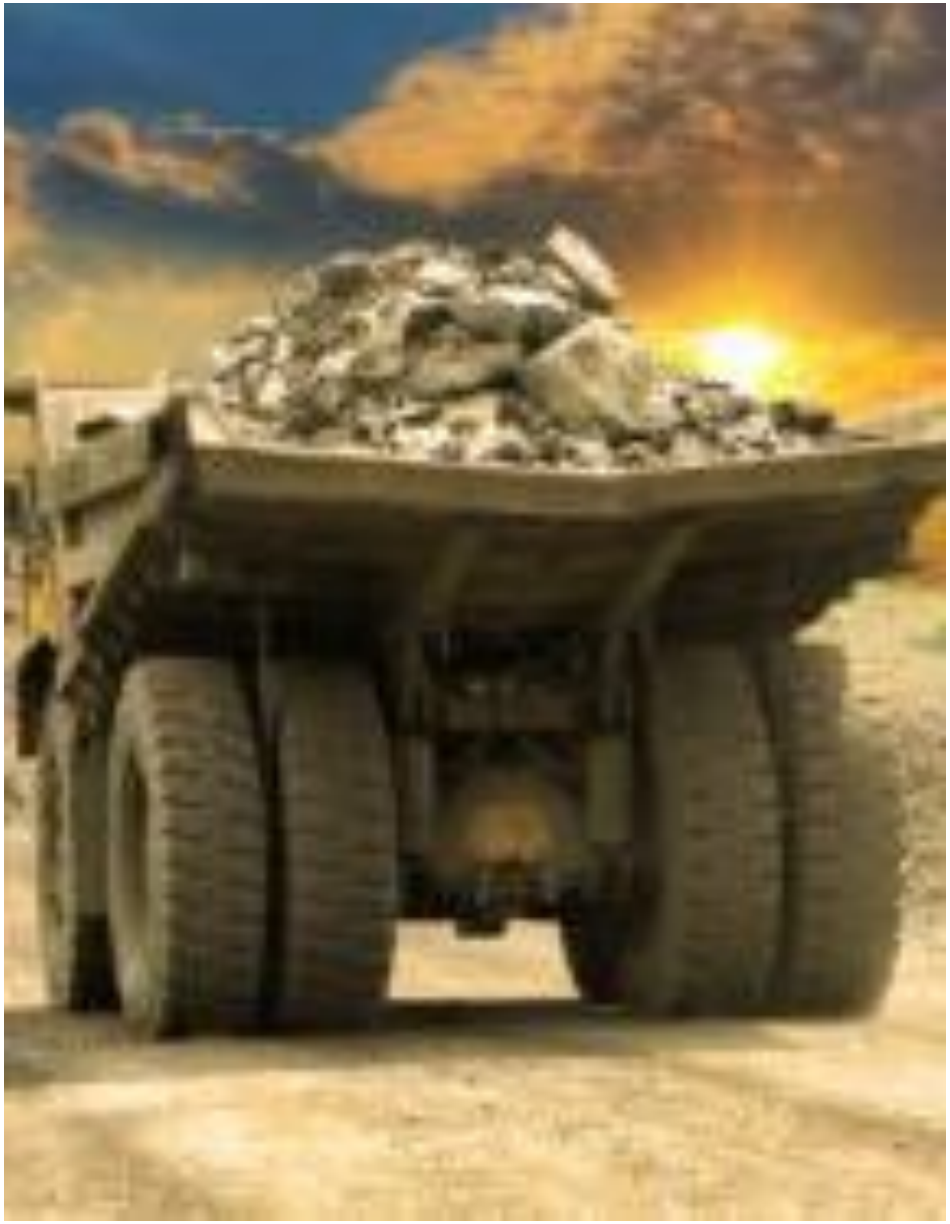




https://www.dataquest.io/blog/python-vs-r/

Agenda

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



References

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2020). Introduction to Data Mining. New York, NY: Pearson.

Hahsler, Michael. (2021). An R Companion for Introduction to Data Mining.
Online Book.
https://mhahsler.github.io/Introduction_to_Data_Mining_R_Examples/book/