DATA MINING
Course Overview

Dino Pedreschi, Riccardo Guidotti

UNIVERSITÀ DI PISA

Teachers

- **Dino Pedreschi**
  - Computer Science Department
  - Email: dino.pedreschi@unipi.it

- **Riccardo Guidotti**
  - Computer Science Department
  - Email: riccardo.guidotti@unipi.it

- **Alessio Cascione (Assistant)**
  - Computer Science Department
  - Email: alessio.cascione@di.unipi.it

Classes

- Classes
  - **Monday, 9:00 - 11:00, Room Fibonacci E**
  - **Thursday, 9:00 - 11:00, Room Fibonacci E**
  - Actual times: Monday 9:15 - 10:45 - Thursday 9:15-10:45 (OK?)

- Office Hours
  - RG: Thursday, 15-17, in presence and MS Teams
  - Appointment [DM1 Meeting] at riccardo.guidotti@unipi.it
  - DP: Monday, 15-17, in presence and MS Teams
  - Appointment [DM1 Meeting] at dino.pedreschi@unipi.it

- Teaching Assistant
  - Alessio Cascione [DM1 Meeting] at alessio.cascione@di.unipi.it

Topics

DM1

- Introduction to Data Mining
- Data Understanding
- Data Preparation
- Clustering
- Foundations of Classification
- Foundations of Regression
- Frequent & Sequential Pattern Mining

DM2

- Imbalanced Learning
- Dimensionality Reduction
- Anomaly Detection
- Advanced Classification/Regression
- Time Series Analysis
- Transactional Clustering
- Explainability

DM1 Topics

- **Module 1: Data Understanding**
  - KDD & CRISP
  - Data Understanding
  - Data Preparation
  - Data Similarity
  - Basic Statistics

- **Module2: Clustering**
  - Taxonomy & Problem Setting
  - Centroid-based Clustering
  - Hierarachical Clustering
  - Density-based Clustering
  - Advanced Approaches

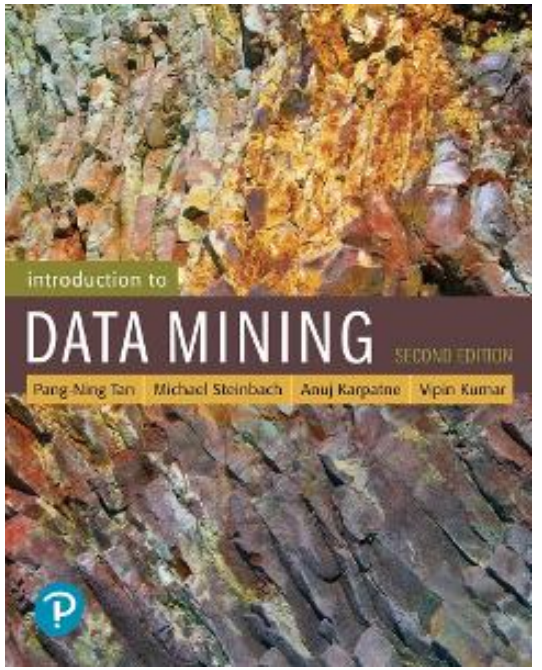
- **Module 3: Pattern Mining**
  - Frequent Pattern Mining
  - Association Rules
  - Sequential Pattern Mining
  - Generalized Sequential Pattern
  - Advanced Approaches

- **Module 4: Classification & Regression**
  - Problem Setting
  - K Nearest Neighbor
  - Naive Bayes Classifier
  - Decision Tree Classifier
  - Linear Regression

Material

- Web Site: http://didawiki.cli.di.unipi.it/doku.php/dm/start

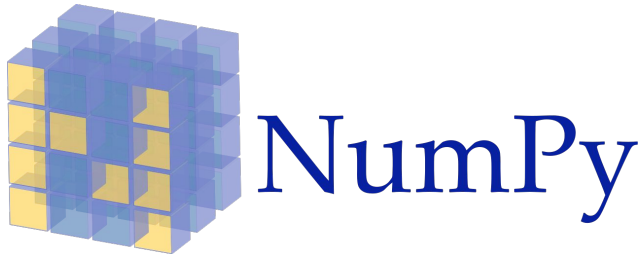
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar. **Introduction to Data Mining**. Addison Wesley, ISBN 0-321-32136-7, 2006, 2° Edition (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)

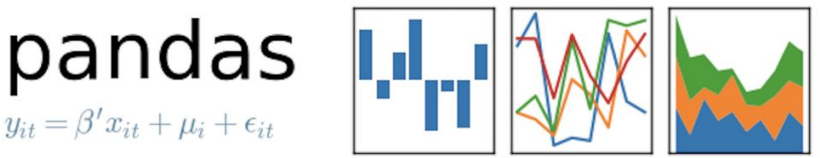


- Berthold, M.R., Borgelt, C., Höppner, F., Klawonn, F. Guide to Intelligent Data Analysis. Springer Verlag, 1st Edition., 2010. ISBN 978-1-84882-259-7
- Laura Igual et al. Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications.
- Slides, Exercises and Notebook

Laboratory

- Python
- Jupyter Notebook





Exam

- Project
  - Topics presented during the classes
  - A single report to be sent periodically and one week before the oral exam
  - Groups composed of up to 3 people
- Oral or Written Exam
  - Short discussion of the project (group presentation, where possible), plus
  - Questions on all topics presented during the classes
  - Exercises and questions about all topics

DM1 Mark = 0.6*Oral + 0.4*Project
DM2 Mark = 0.6*Oral + 0.4*Project
DM Mark = (DM1 + DM2) /2

Dataset (last year)

IMDb dataset

- The IMDb Dataset contains data about movies, TV shows, and other forms of visual entertainment, along with their ratings, which is generated by the internet community. Each record includes basic information from the “IMDb Non-Commercial Datasets” provided by IMDb itself, such as title, duration, and release year. Additional features present on each title's webpage enrich the dataset, covering aspects like statistical information (e.g., highest and lowest ratings, awards nominations, and total number of critic reviews), and technical specifications (e.g., colorations). The dataset is updated as of September 1, 2024.
- The IMDb dataset for the project can be found on the web page of the course.
- Detailed guidelines for the project will be presented next lecture and made on the web page of the course.

Homework and Suggestions (last year)

Homework

- Declare Project Groups by next Thursday adding your information at https://docs.google.com/spreadsheets/d/1RFWIwKM5Myaehh4tHceaf3olMYm_CktGvoNOFX2Oovc/edit?usp=sharing

Suggestions

- Download and start to play with the dataset
- Use a Github repository for python and ipython files.
- Use a shared Overleaf project (LaTex) for the report.

Questions?

dino.pedreschi@unipi.it

riccardo.guidotti@unipi.it

alessio.cascione@di.unipi.it

DATA MINING 1
Introduction

Dino Pedreschi

UNIVERSITÀ DI PISA

What is Data Mining?

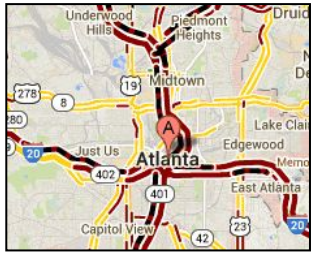
- It is the use of **efficient** techniques for the analysis of **very large collections of data** and the **extraction** of useful and possibly unexpected patterns in data (hidden knowledge).

Big Data is Everywhere!

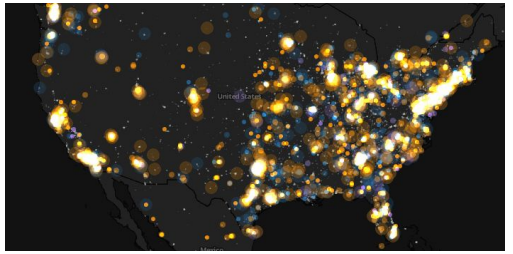
- **Enormous data growth in both commercial and scientific databases**
  - due to advances in data generation and collection technologies
- **New mantra**
  - Gather whatever data you can whenever and wherever possible
- **Expectations**
  - Gathered data will have value either for the purpose collected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

Cyber Security

E-Commerce

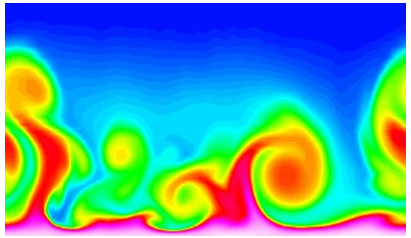


Traffic Patterns



Social Networking: Twitter

Sensor Networks



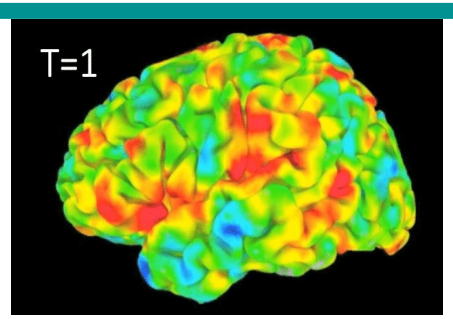
Computational Simulations

Why Data Mining? Commercial Viewpoint

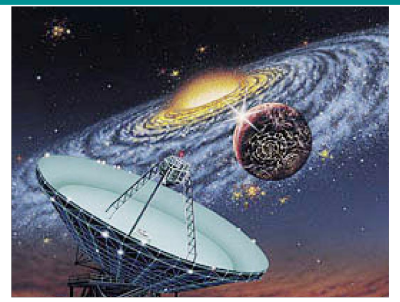
- **Lots of data is being collected and warehoused**
  - Web data
    - Google has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- **Computers have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

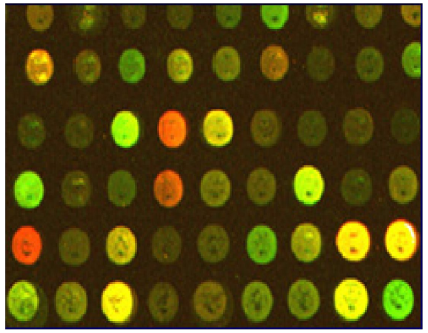
Why Data Mining? Scientific Viewpoint

- **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

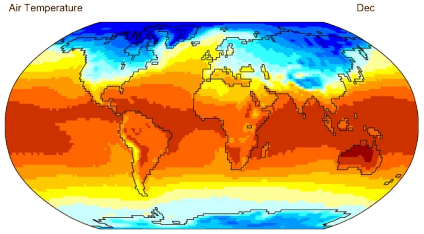


fMRI Data from Brain


Sky Survey Data


Gene Expression Data

- **Data mining helps scientists**
  - in automated analysis of massive datasets
  - In hypothesis formation

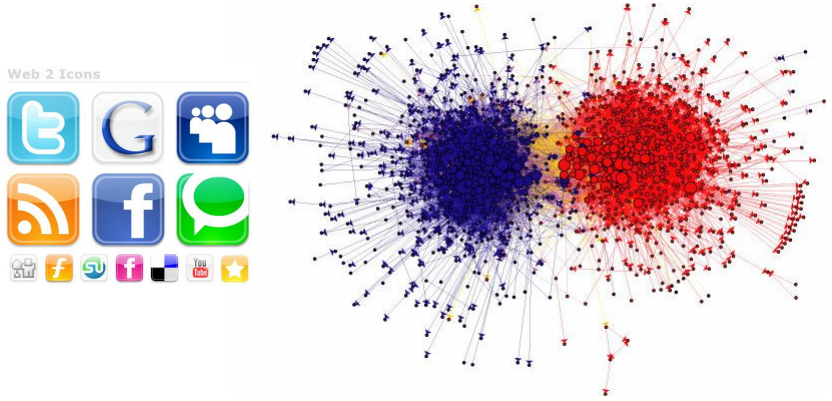


Surface Temperature of Earth

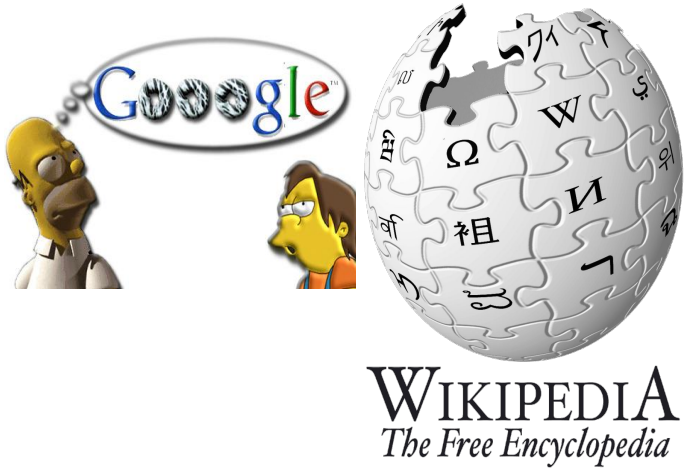
Big Data as Proxies of Social Life

SHOPPING PATTERNS & LIFESTYLE

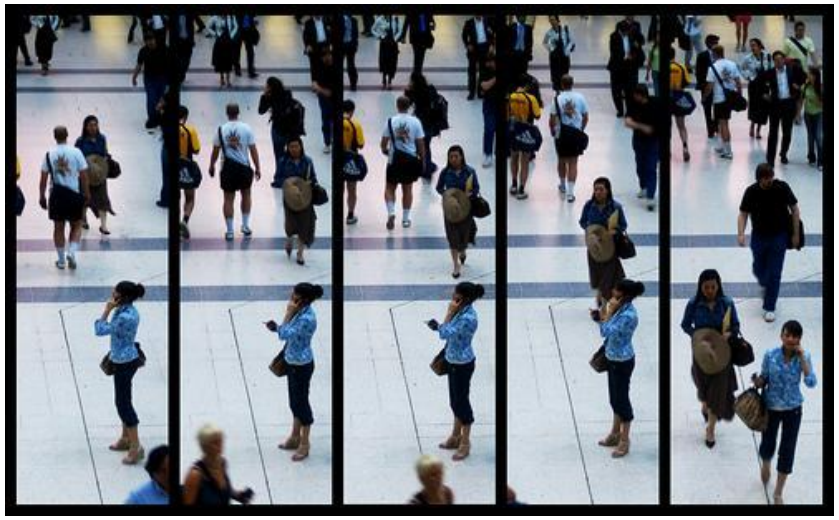
RELATIONSHIPS & SOCIAL TIES

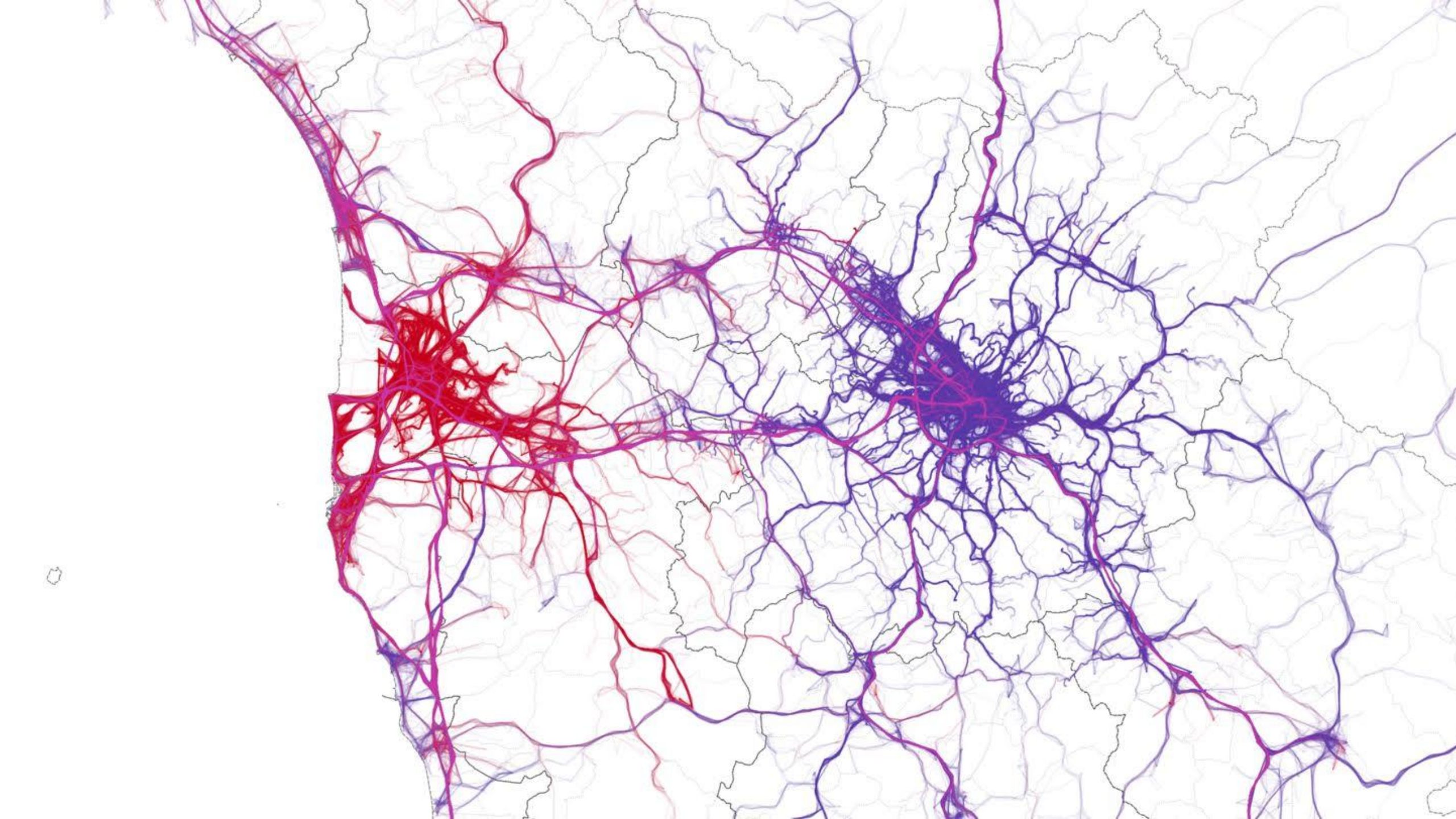


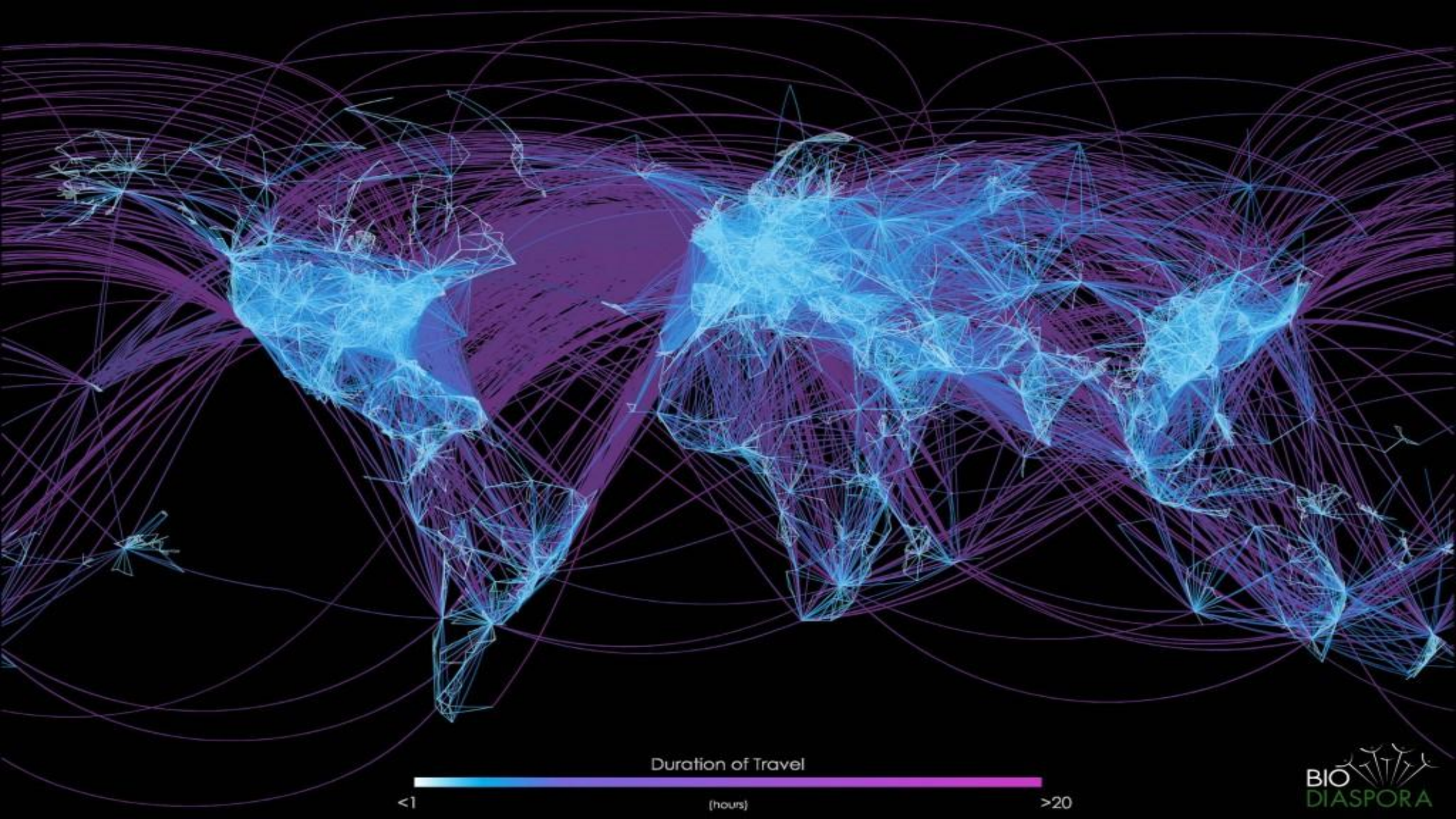
DESIRES, OPINIONS, SENTIMENTS



MOVEMENTS



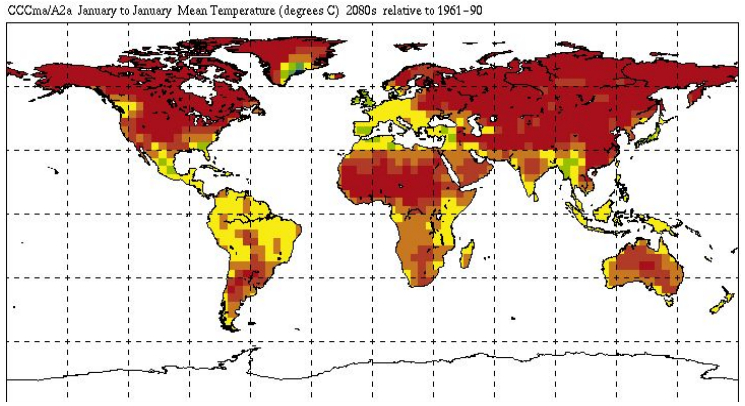



Duration of Travel
<1
(hours)
>20
BIO
DIASPORA

Great Opportunities to Solve Global Challenges

Improving health care and reducing costs

Finding alternative/ green energy sources



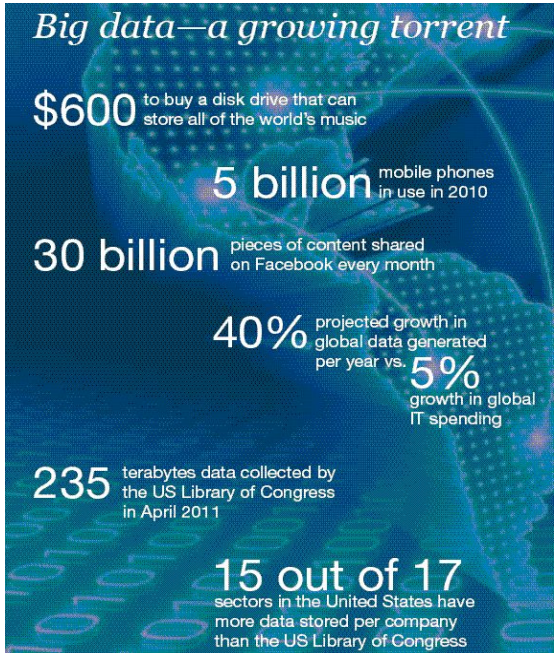
Predicting the impact of climate change

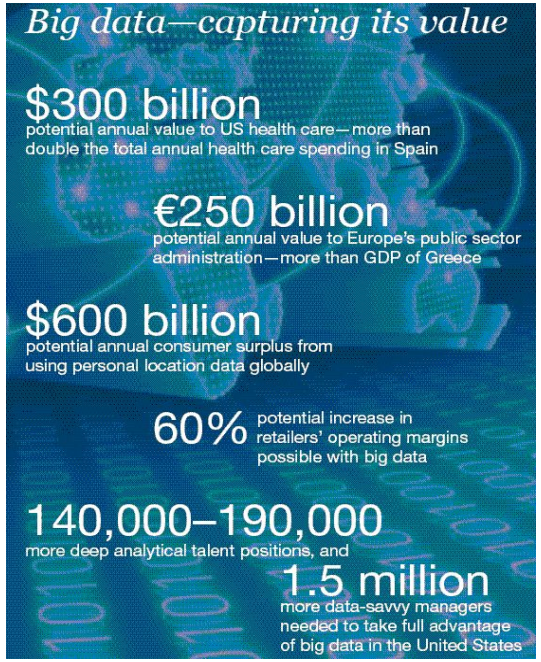
Reducing hunger and poverty by increasing agriculture production

Great Opportunities to Improve Productivity

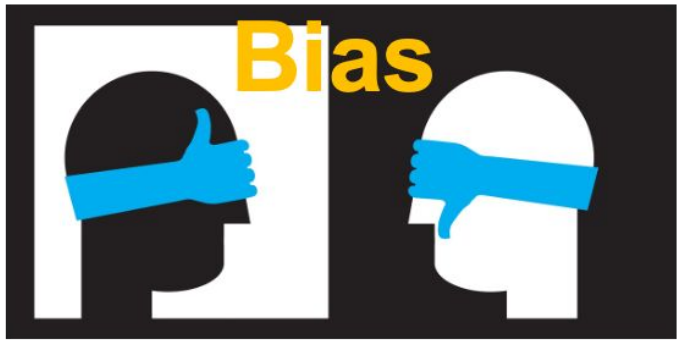
McKinsey Global Institute

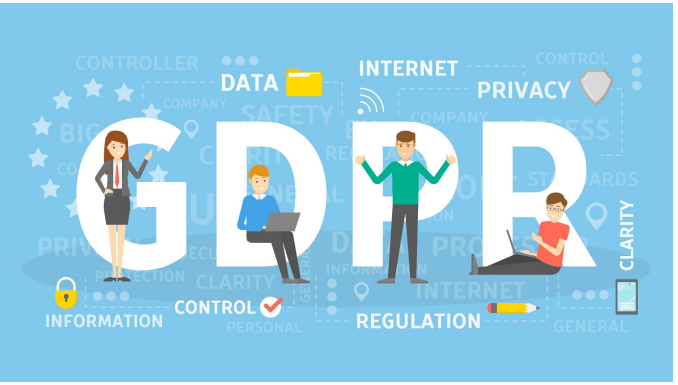
Big data: The next frontier for innovation, competition, and productivity

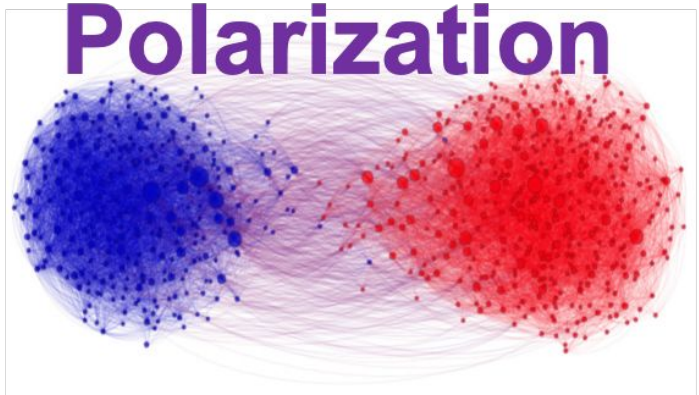




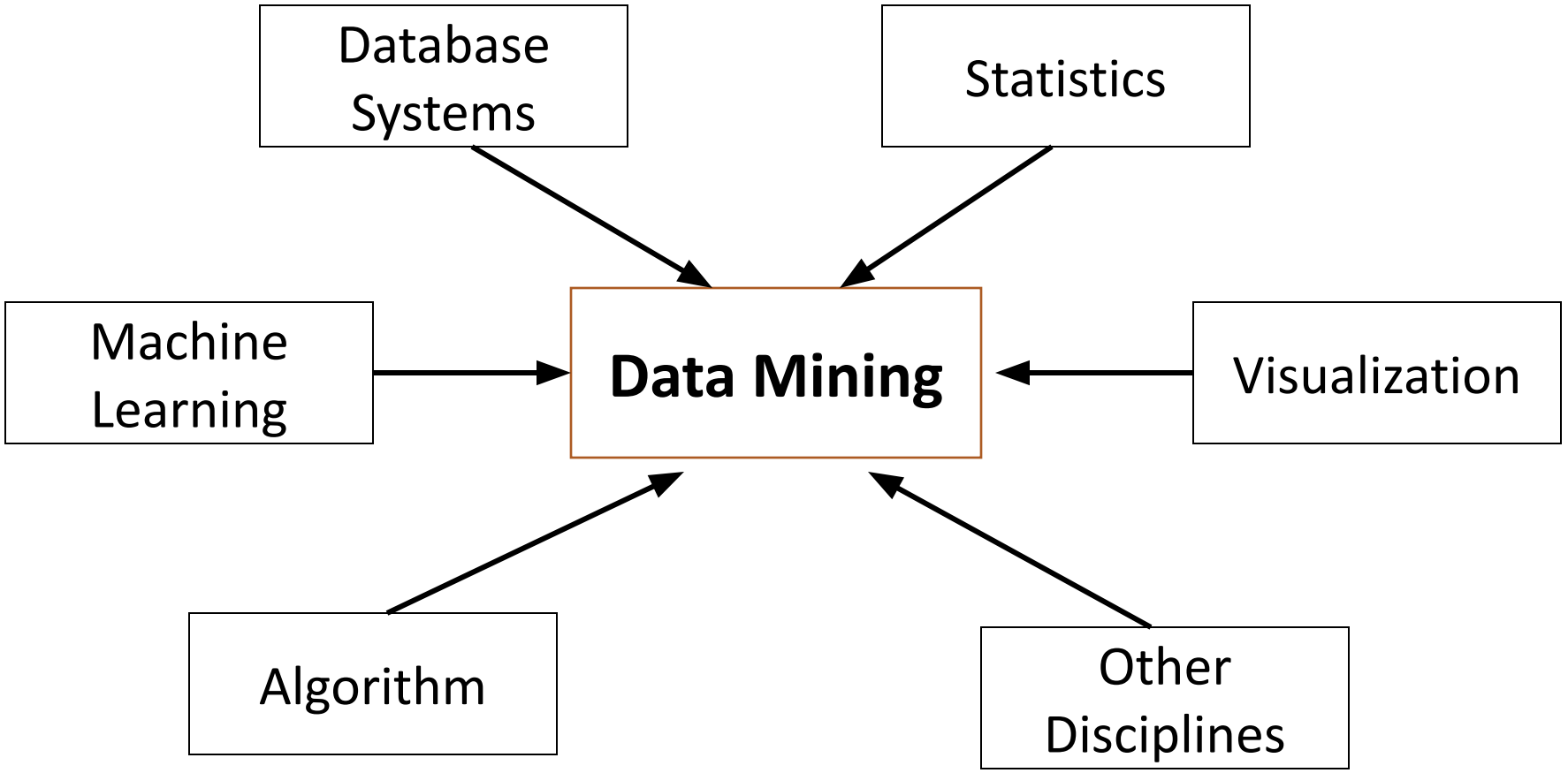
The ethical and legal dimensions of data science



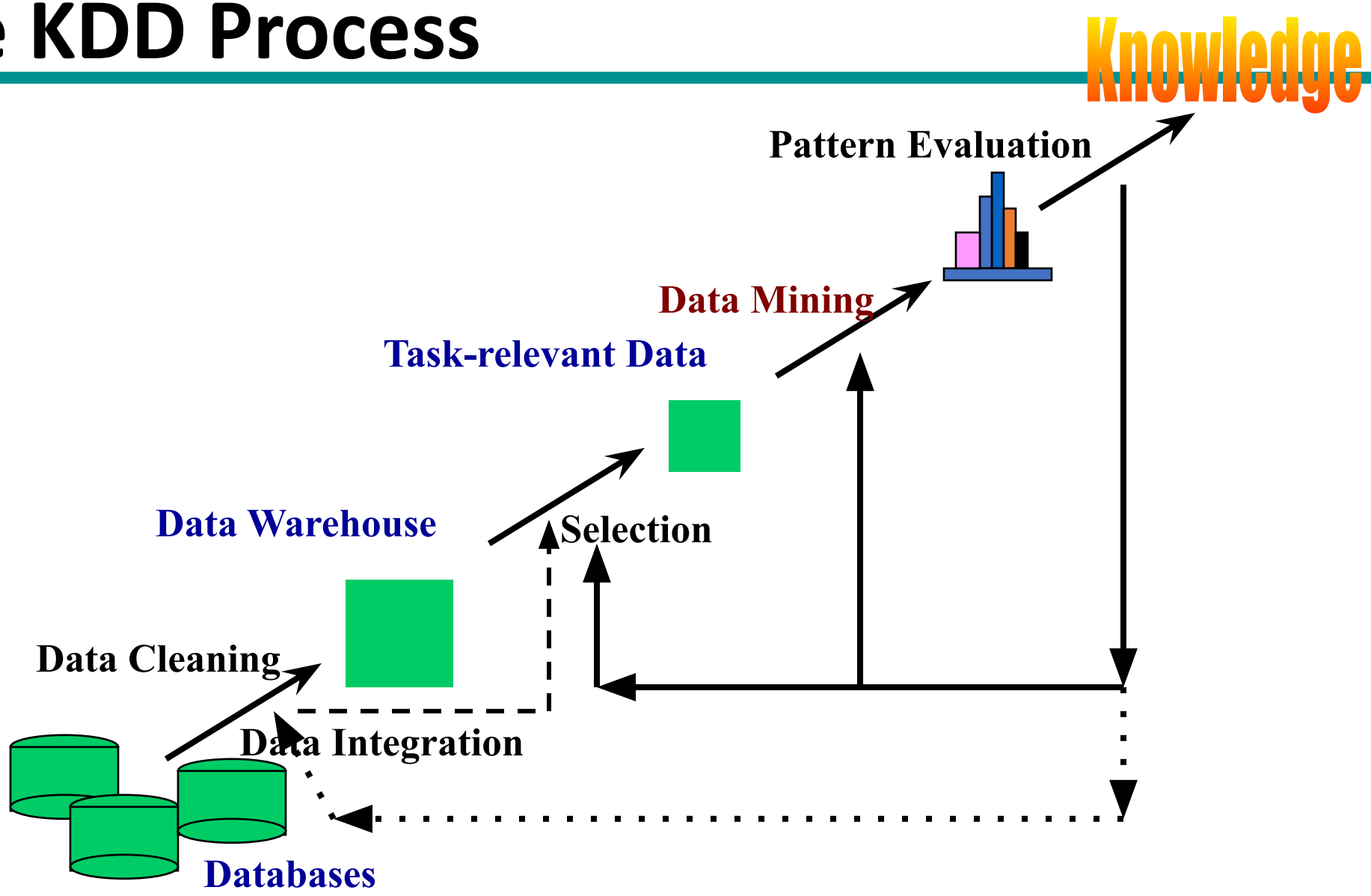




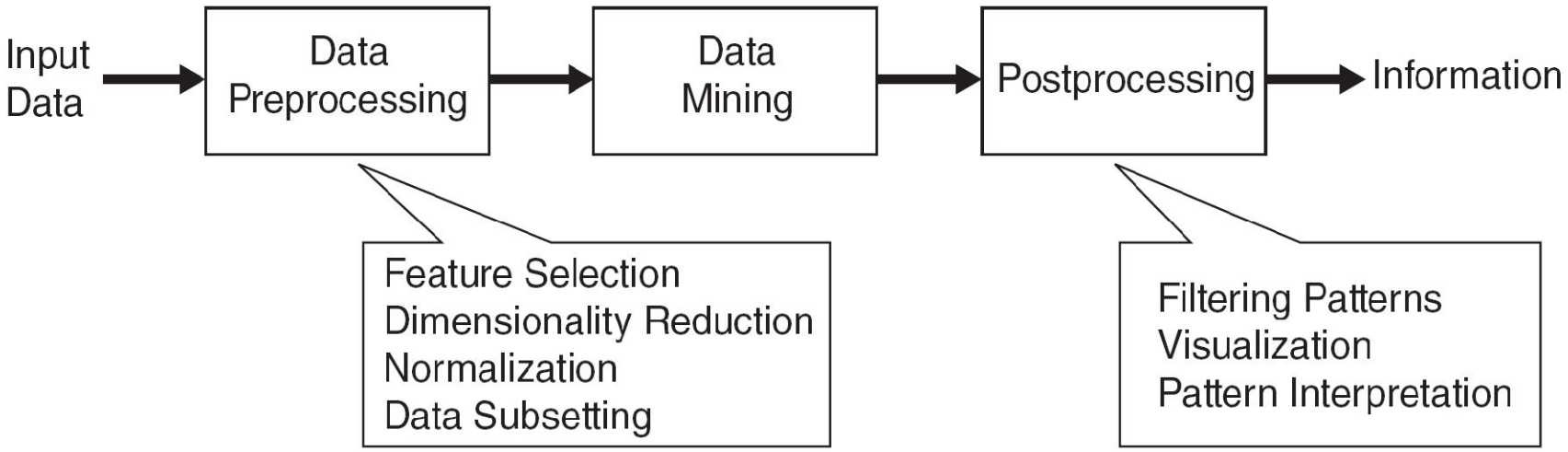
Data Mining: Confluence of Multiple Disciplines



The KDD Process



What is Data Mining?



Primary & Secondary Data

Primary Data

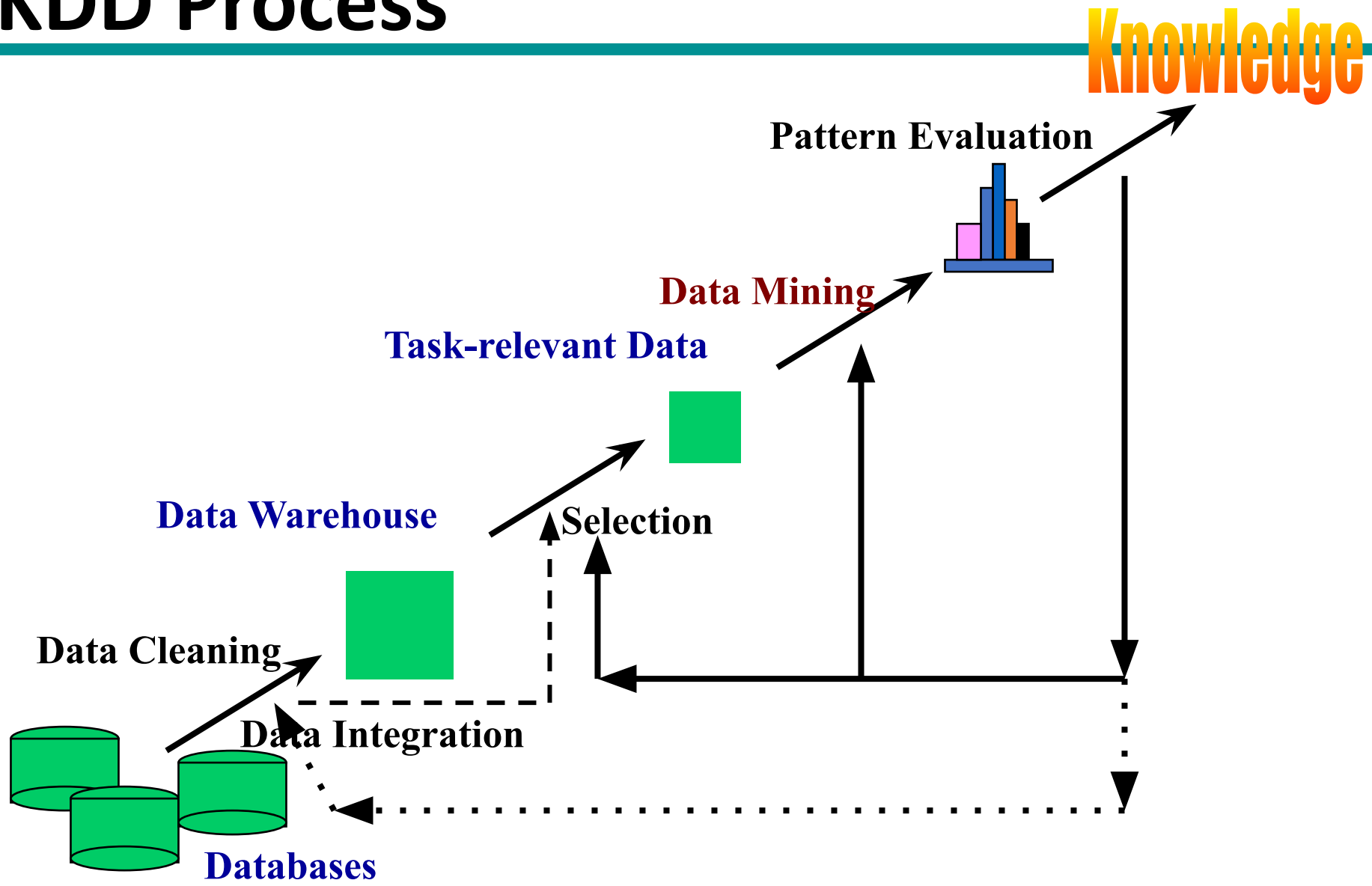
- **Original data** that has been collected for a specific purpose
- Primary data **is not altered by humans**

Secondary Data

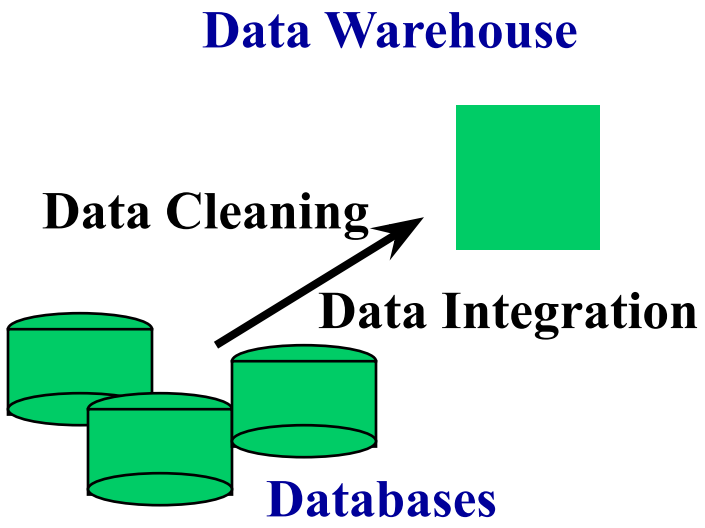
- **Data that** has been already collected and **made available for other purposes**
- Secondary data may be obtained **from many sources**

Varierty of Data Sources

The KDD Process



Data Integration and Preparation

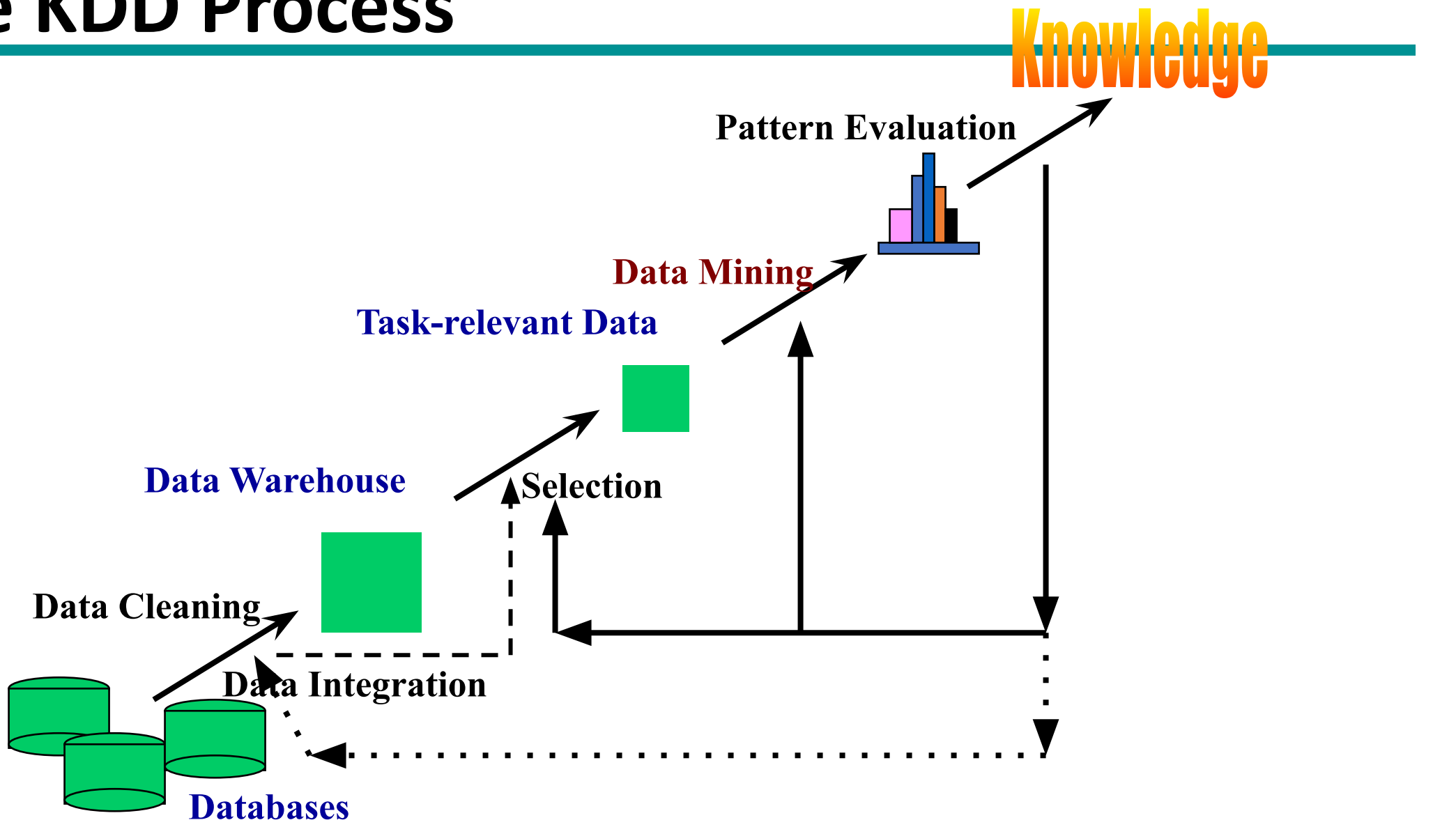


Data Integration involves the process of data understanding, data cleaning, merging data coming from multiple sources and transfoming them to load them into a **Data Warehouse**

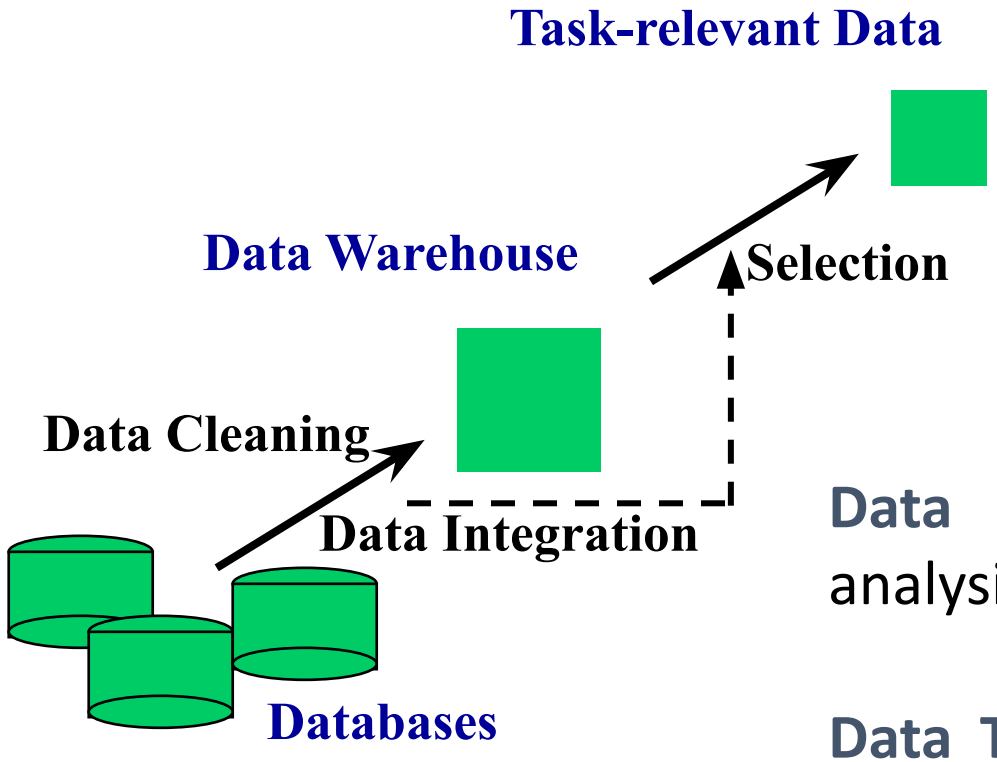
Data Warehouse is a database targeted to answer **specific business questions**

Developing a real data analytics project also requires the
BUSINESS UNDERSTANDING

The KDD Process



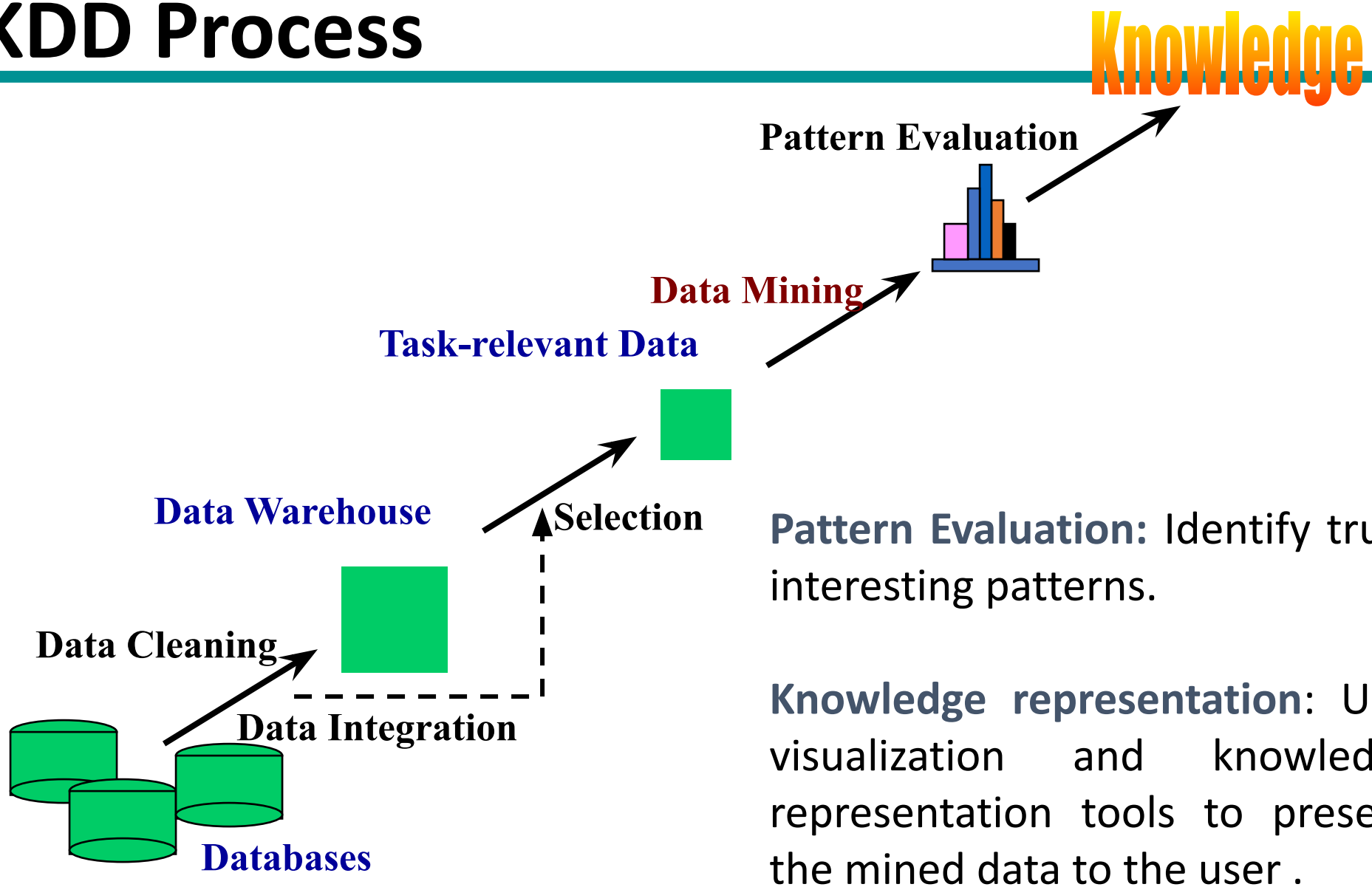
Data Selection and Transformation



Data Selection: Data relevant to analysis tasks are retrieved from data

Data Transformation: Transform data into appropriate form for mining (summary, aggregation, etc.)

The KDD Process



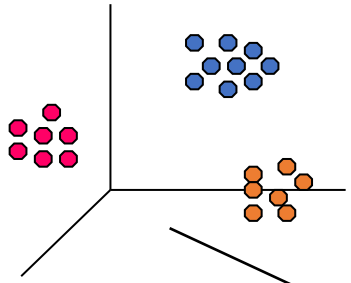
Pattern Evaluation: Identify truly interesting patterns.

Knowledge representation: Use visualization and knowledge representation tools to present the mined data to the user .

Data Mining Tasks

- Predictive Methods
  - Use some variables to predict unknown or future values of other variables.
- Descriptive Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks



Clustering

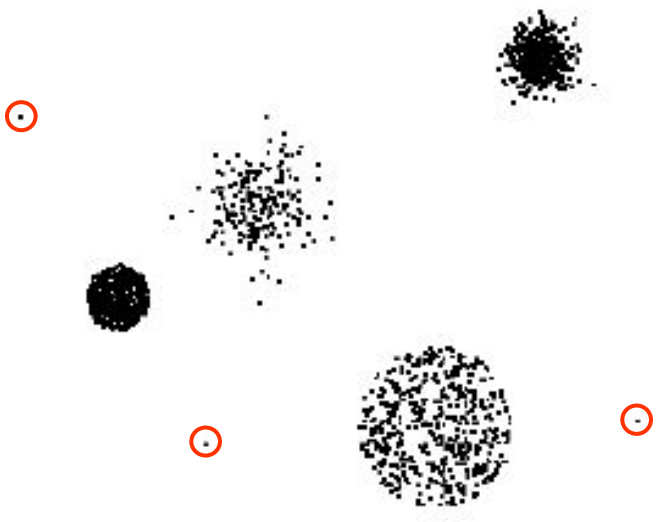
Data

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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling

Association Rules

Anomaly Detection



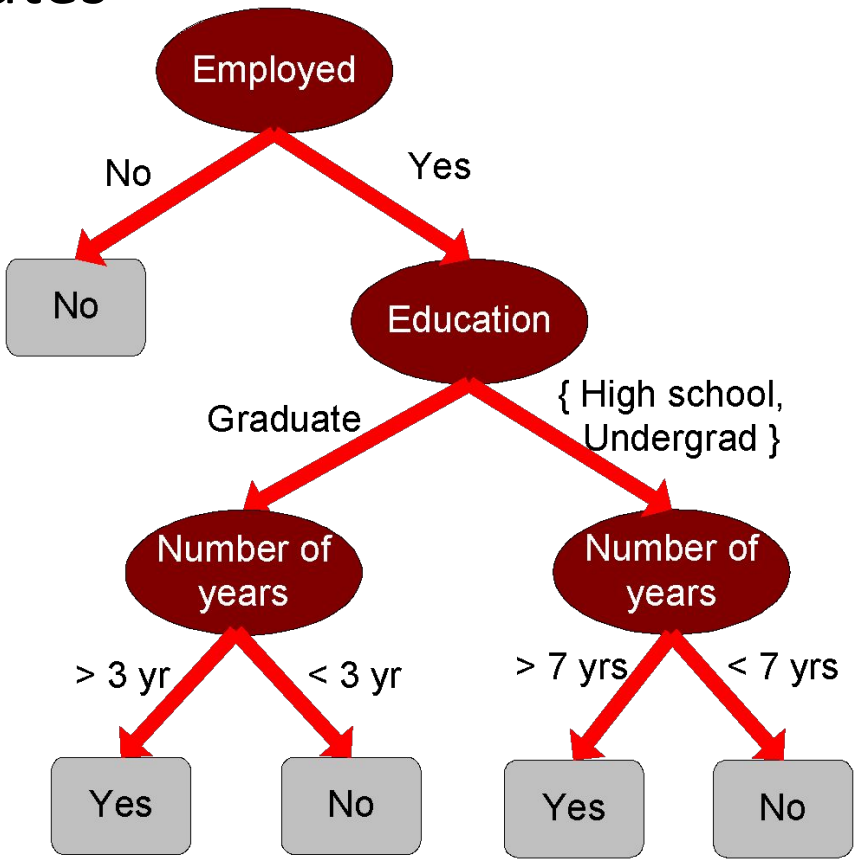
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

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
…	…	…	…	…

Model for predicting credit worthiness

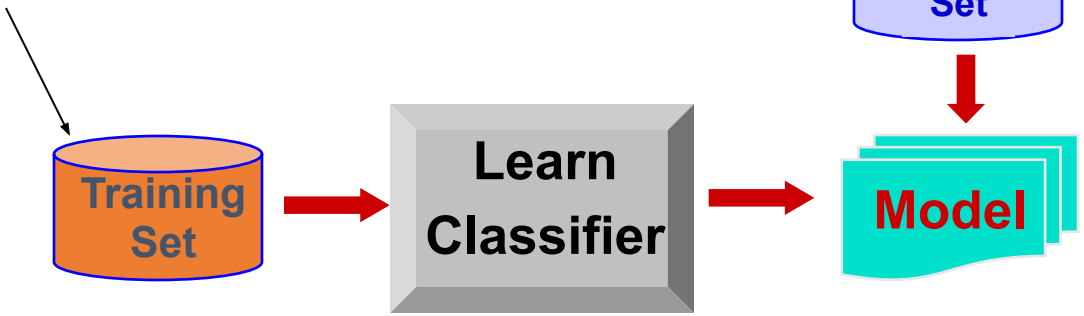


Classification Example

categorical categorical quantitative class

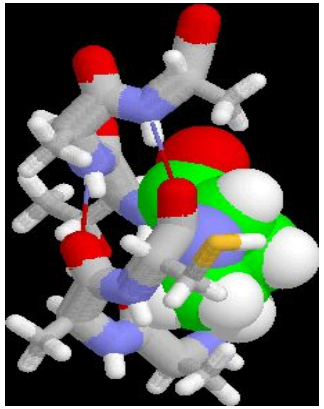
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

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

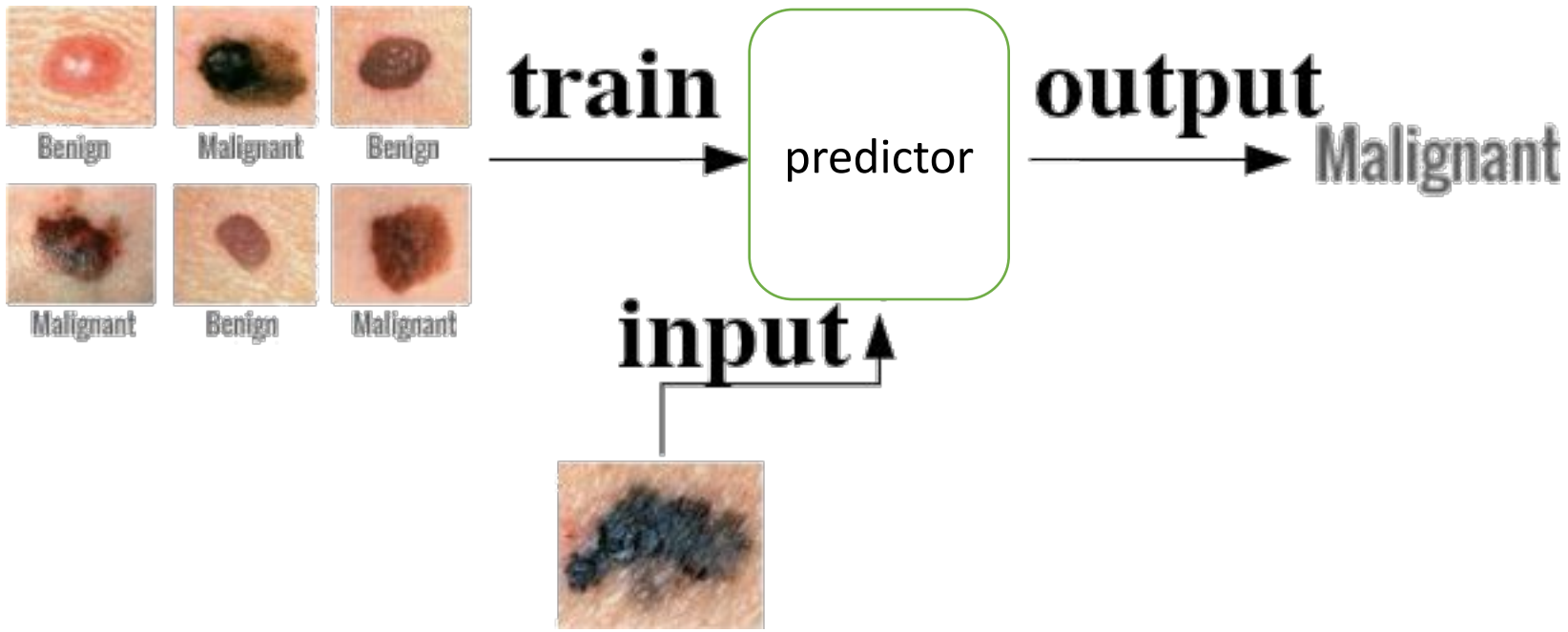


Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



AI = Machine Learning + Big Data



Classification: Application 1

Fraud Detection

- **Goal:** Predict fraudulent cases in credit card transactions.
- **Approach:**
  - Use credit card transactions and the information on its account-holder as attributes.
  - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions.
  - Use this model to detect fraud by observing credit card transactions on an account.

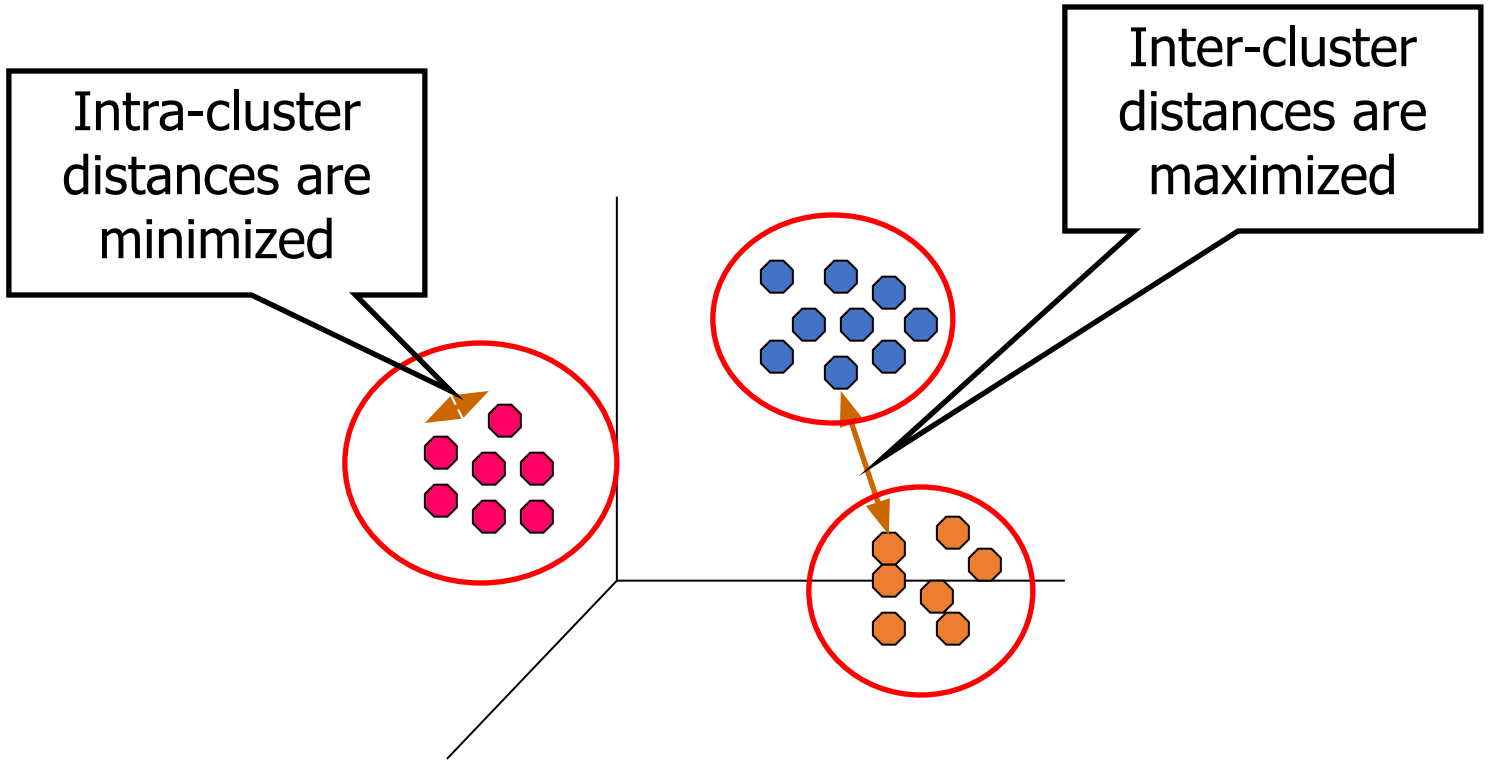
Classification: Application 2

Churn prediction for telephone customers

- **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.
  - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
  - Label the customers as loyal or disloyal.
  - Find a model for loyalty.

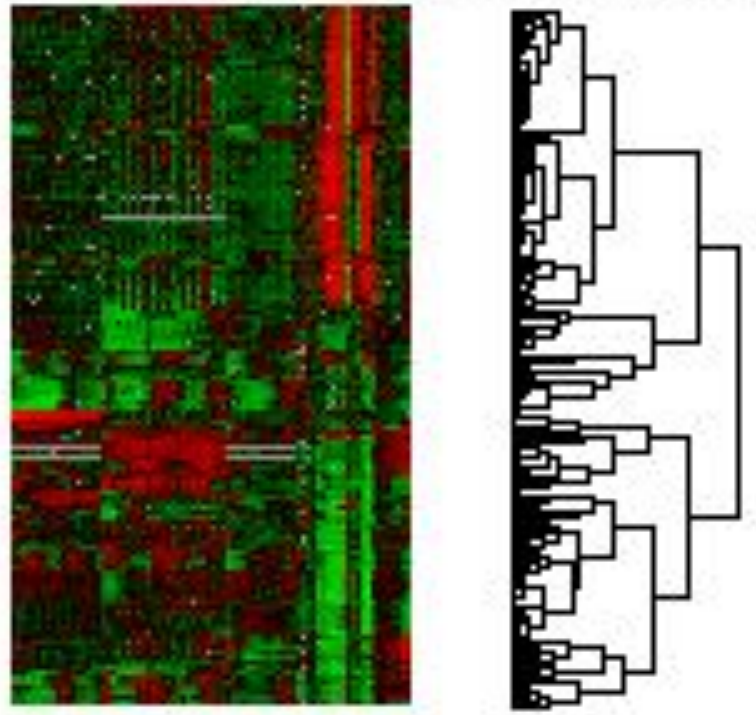
Clustering

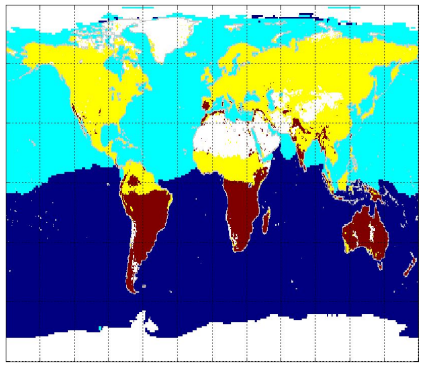
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.



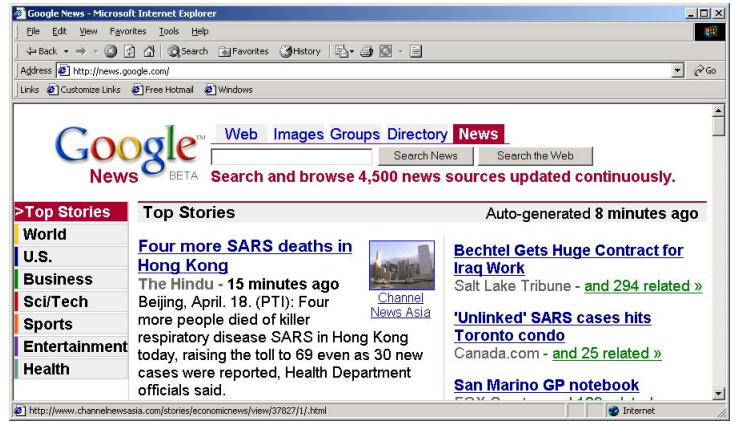
Applications of Cluster Analysi

- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets





Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Clustering: Application 1

Market Segmentation:

- **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- **Approach:**
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of **similar customers**.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

A Behavior Based Segmentation

Using unsupervised clustering segmentation for a grocery chain which would like better product assortment for its high profitable customers

Potential Inputs

Value	• Basket Size • Visit Frequency
Basket	• Spend by category • Type of category • Brand spend (i.e. private label)
Promotions	• % bought on targeted promotion • % bought from flyer
Time	• Time of day • Day of week
Location	• Store format • Area population density

Clustering approach

Deal Seeking Mom

Key Differentiators

- Full store shop
- High avg. basket size / # trips

- High % purchased on promotion
- Rewards seeker

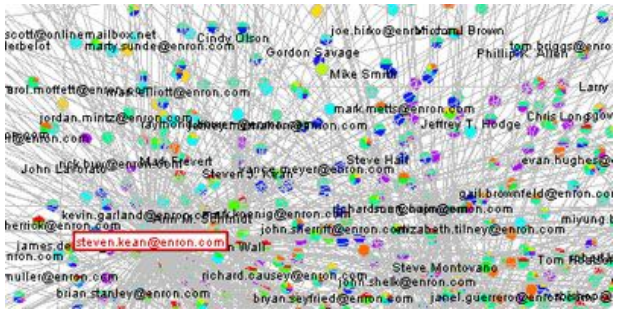
- High spend categories
  - Fresh produce
  - Organic food
  - Multipack juice, snack

Clustering: Application 2

Document Clustering:

- **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
- **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency 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

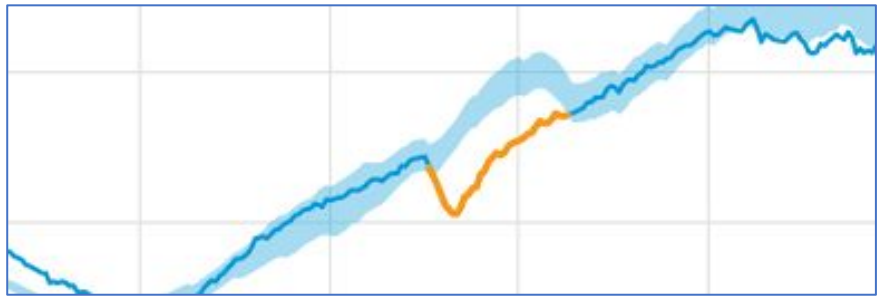
Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association Analysis: Applications

- **Market-basket analysis**
  - Rules are used for sales promotion, shelf management, and inventory management
- **Telecommunication alarm diagnosis**
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- **Medical Informatics**
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior



- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.

Motivating Challenges

Statistic techniques may be unsuitable due to some challenges:

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis

CRoss-Industry Standard Process for Data Mining

Why Should There be a Standard Process?

- The data mining process must be *reliable* and *repeatable* by people with little data mining background.
- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- “Comfort factor” for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on “stars”

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

Overview


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

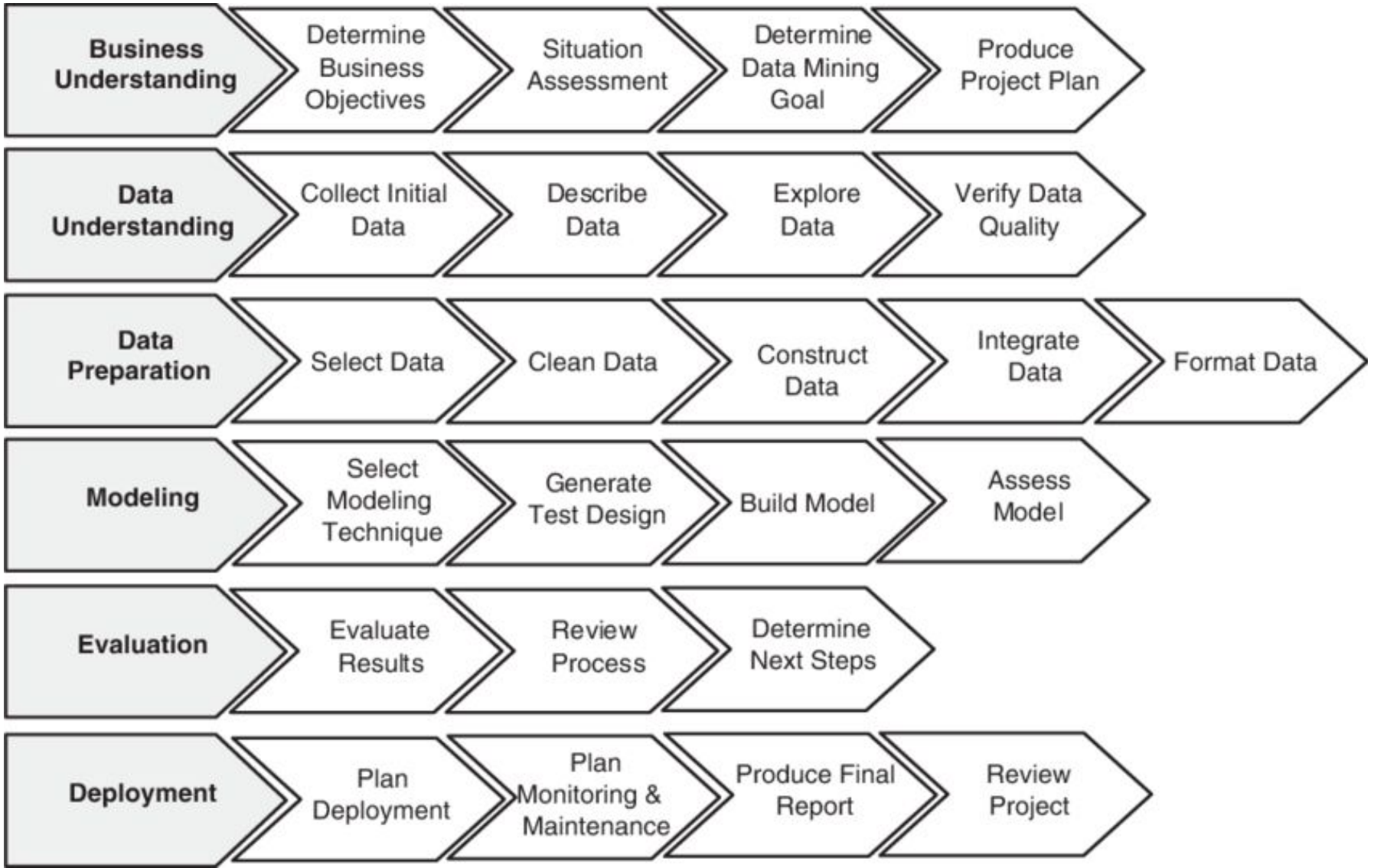
Phases

- **Business Understanding**
  - Project objectives and requirements understanding, Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization, Data quality problems identification
- **Data Preparation**
  - Table, record and attribute selection, Data transformation and cleaning
- **Modeling**
  - Modeling techniques selection and application, Parameters calibration
- **Evaluation**
  - Business objectives & issues achievement evaluation
- **Deployment**
  - Result model deployment, Repeatable data mining process implementation

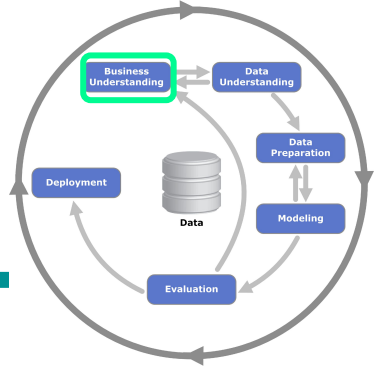
Phases

- **Business Understanding**
  - Project objectives and requirements understanding, Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization, Data quality problems identification
- **Data Preparation**
  - Table, record and attribute selection, Data transformation and cleaning
- **Modeling**
  - Modeling techniques selection and application, Parameters calibration
- **Evaluation**
  - Business objectives & issues achievement evaluation
- **Deployment**
  - Result model deployment, Repeatable data mining process implementation

Phases and Tasks

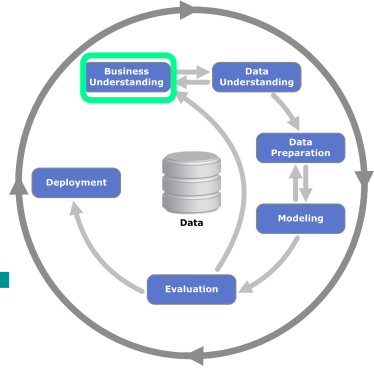


Business Understanding



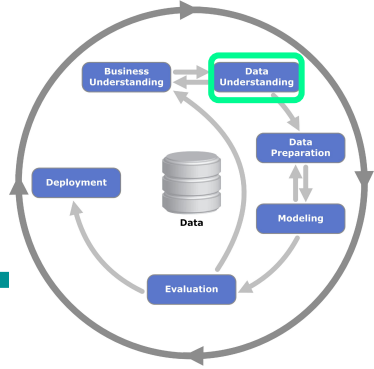
- **Determine business objectives**
  - thoroughly understand, from a business perspective, what the client really wants to accomplish
  - uncover important factors, at the beginning, that can influence the outcome of the project
  - neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions
- **Assess situation**
  - more detailed fact-finding about all of the resources, constraints, assumptions and other factors that should be considered
  - flesh out the details

Business Understanding



- **Determine data mining goals**
  - a business goal states objectives in business terminology
  - a data mining goal states objectives in technical terms
    - A business goal: “Increase catalog sales to existing customers.”
    - A data mining goal: “Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city) and the price of the item.”
- **Produce project plan**
  - describe the intended plan for achieving the data mining goals and the business goals
  - the plan should specify the anticipated set of steps to be performed during the rest of the project including an initial selection of tools and techniques

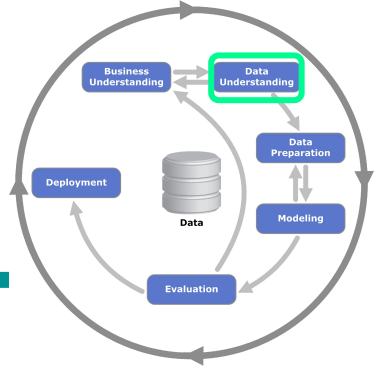
Data Understanding



- **Explore the Data**
- **Verify the Quality**
- **Find Outliers**

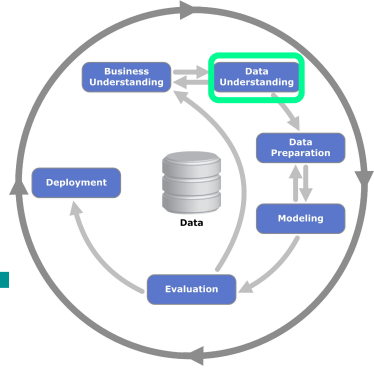
Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data Understanding



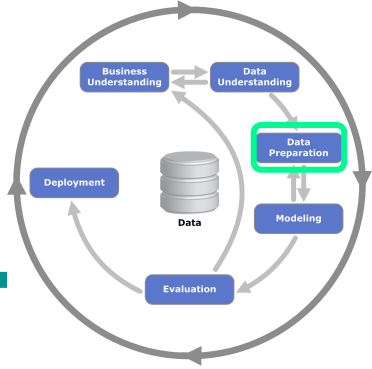
- **Collect initial data**
  - acquire within the project the data listed in the project resources
  - includes data loading if necessary for data understanding
  - possibly leads to initial data preparation steps
  - if acquiring multiple data sources, integration is an additional issue, either here or in the later data preparation phase
- **Describe data**
  - examine the “gross” or “surface” properties of the acquired data
  - report on the results

Data Understanding



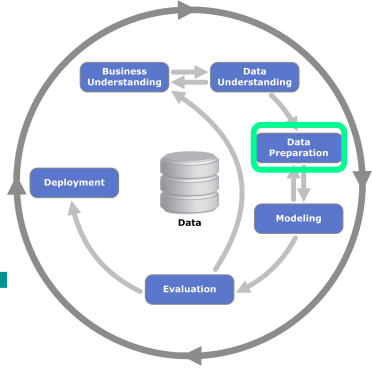
- **Explore data**
  - tackles the data mining questions, which can be addressed using querying, visualization and reporting including:
    - distribution of key attributes, through aggregations
    - relations between pairs of attributes
    - properties of significant sub-populations
  - may address directly the data mining goals
  - may contribute to data description and quality reports
  - may feed into the transformation and other data preparation needed
- **Verify data quality**
  - examine the quality of the data, addressing questions such as: “Is the data complete?”, Are there missing values in the data?”

Data Preparation



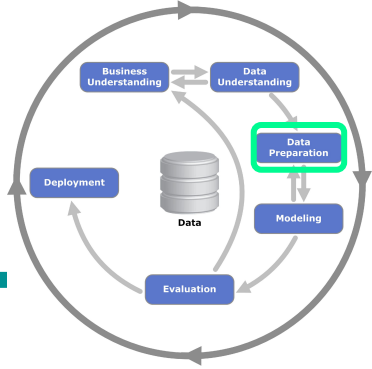
- Takes usually over 90% of the time
  - **Collection**
  - **Assessment**
  - **Consolidation and Cleaning**
  - **Data selection**
  - **Transformations**
- Covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Data Preparation



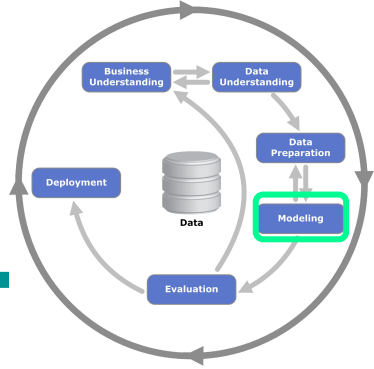
- **Select data**
  - decide on the data to be used for analysis
  - criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types
  - covers selection of attributes as well as selection of records in a table
- **Clean data**
  - raise the data quality to the level required by the selected analysis techniques
  - may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling

Data Preparation



- **Construct data**
  - constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes
- **Integrate data**
  - methods whereby information is combined from multiple tables or records to create new records or values
- **Format data**
  - formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool

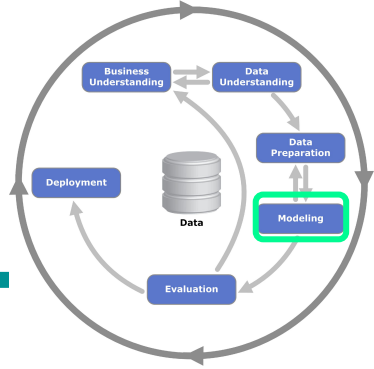
Modeling



- **Select the modeling technique**
  - (based upon data mining objectives)
- **Build model**
  - (Parameter settings)
- **Assess model**
  - (rank the models)

Various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Some techniques have specific requirements on the form of data. Therefore, *stepping back to the data preparation phase is often necessary*.

Modeling

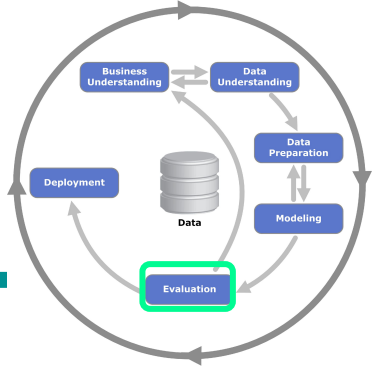


- **Select modeling technique**
  - select the actual modeling technique that is to be used ex) decision tree, neural network
  - if multiple techniques are applied, perform this task for each technique separately
- **Generate test design**
  - before actually building a model, generate a procedure or mechanism to test the model's quality and validity ex) In classification, it is common to use error rates as quality measures for data mining models. Therefore, typically separate the dataset into train and test set, build the model on the train set and estimate its quality on the separate test set

Modeling

- **Build model**
  - run the modeling tool on the prepared dataset to create one or more models
- **Assess model**
  - interprets the models according to his domain knowledge, the data mining success criteria and the desired test design
  - judges the success of the application of modeling and discovery techniques more technically
  - contacts business analysts and domain experts later in order to discuss the data mining results in the business context
  - only consider models whereas the evaluation phase also takes into account all other results that were produced in the course of the project

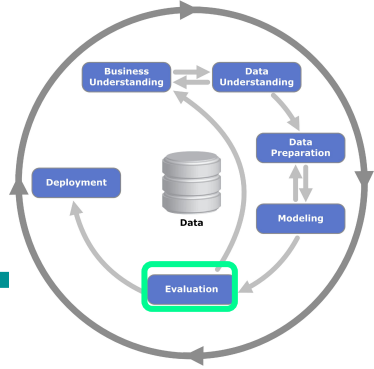
Evaluation



- **Evaluation of model**
  - how well it performed on test data
- **Methods and criteria**
  - depend on model type
- **Interpretation of model**
  - importance and hardness depend on the algorithm

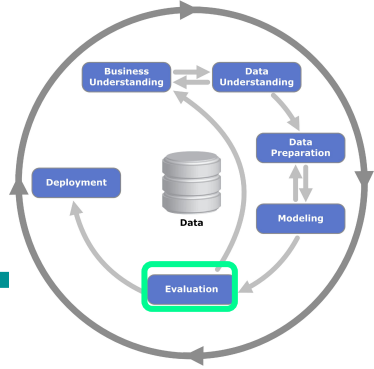
Thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is *to determine if there is some important business issue that has not been sufficiently considered*. At the end of this phase, a decision on the use of the data mining results should be reached

Evaluation



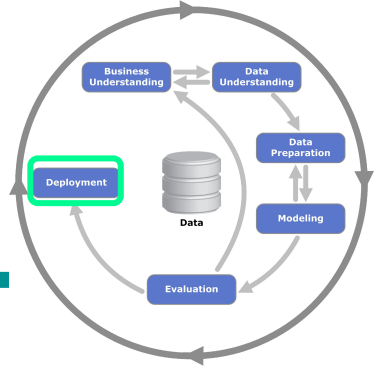
- **Evaluate results**
  - assesses the degree to which the model meets the business objectives
  - seeks to determine if there is some business reason why this model is deficient
  - test the model(s) on test applications in the real application if time and budget constraints permit
  - also assesses other data mining results generated
  - unveil additional challenges, information or hints for future directions

Evaluation



- **Review process**
  - do a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked
  - review the quality assurance issues ex) “Did we correctly build the model?”
- **Determine next steps**
  - decides how to proceed at this stage
  - decides whether to finish the project and move on to deployment if appropriate or whether to initiate further iterations or set up new data mining projects
  - include analyses of remaining resources and budget that influences the decisions

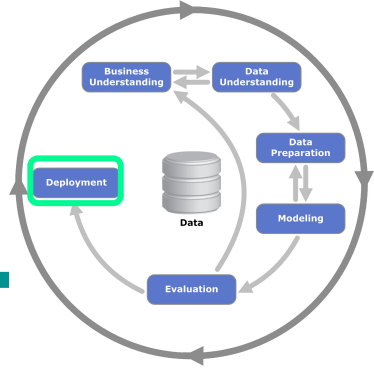
Deployment



- Determine **how** the results need to be utilized
- **Who** needs to use them?
- **How often** do they need to be used
- Deploy Data Mining results

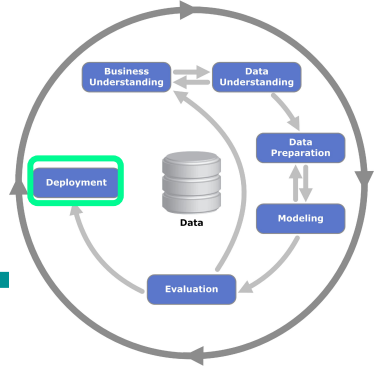
The knowledge gained will need to *be organized and presented in a way that the customer can use it*. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Deployment



- **Plan deployment**
  - in order to deploy the data mining result(s) into the business, takes the evaluation results and concludes a strategy for deployment
  - document the procedure for later deployment
- **Plan monitoring and maintenance**
  - important if the data mining results become part of the day-to-day business and it environment
  - helps to avoid unnecessarily long periods of incorrect usage of data mining results
  - needs a detailed on monitoring process
  - takes into account the specific type of deployment

Deployment



- **Produce final report**
  - the project leader and his team write up a final report
  - may be only a summary of the project and its experiences
  - may be a final and comprehensive presentation of the data mining result(s)

- **Review project**
  - assess what went right and what went wrong, what was done well and what needs to be improved

Summary

Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills
- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation
- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

References

- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz, (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) “CRISP-DM 1.0 -Step-by-step data mining guide”

- Websites
  - http://www.crisp-dm.org/
  - http://www.spss.com/
  - http://www.kdnuggets.com/