What is Data Mining?

Some Definitions:

-"The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Piatetsky-Shapiro)

- "...the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, ... or data streams." (Han)

cuu duong than cong. com

- "...the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful..." (Witten)

- "...finding hidden information in a database." (Dunham)

cuu duong than cong. com

–"...the process of employing one or more computer learning techniques to automatically analyse and extract knowledge from data contained within a database." (Roiger)

CuuDuongThanCong.com https://fb.com/tailieudientucntt

What is Data Mining?

Keywords from each definition:

-"The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Piatetsky-Shapiro)

- "...the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, ... or data streams." (Han)

cuu duong than cong. com

- "...the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful..." (Witten)

- "...finding hidden information in a database." (Dunham)

cuu duong than cong. com

–"...the process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data contained within a database." (Roiger)

CuuDuongThanCong.com https://fb.com/tailieudientucntt

KDD: Knowledge Discovery in Databases

Many texts treat KDD and Data Mining as the same process, but it is also possible to think of Data Mining as the discovery part of KDD.

Dunham:

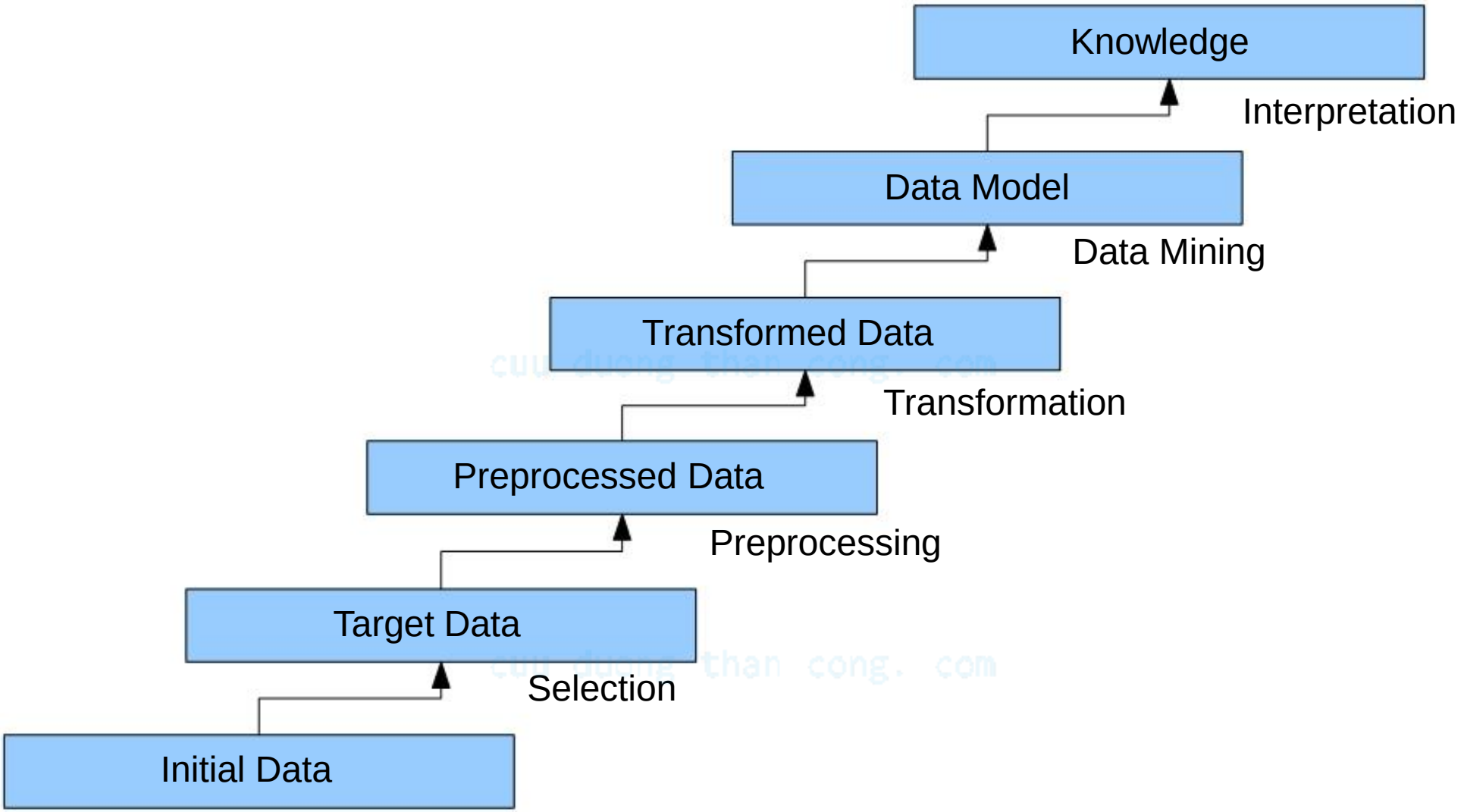
KDD is the process of finding useful information and patterns in data.

Data Mining is the use of algorithms to extract information and patterns derived by the KDD process.

For this course, we will discuss the entire process (KDD) but focus mostly on the algorithms used for discovery.

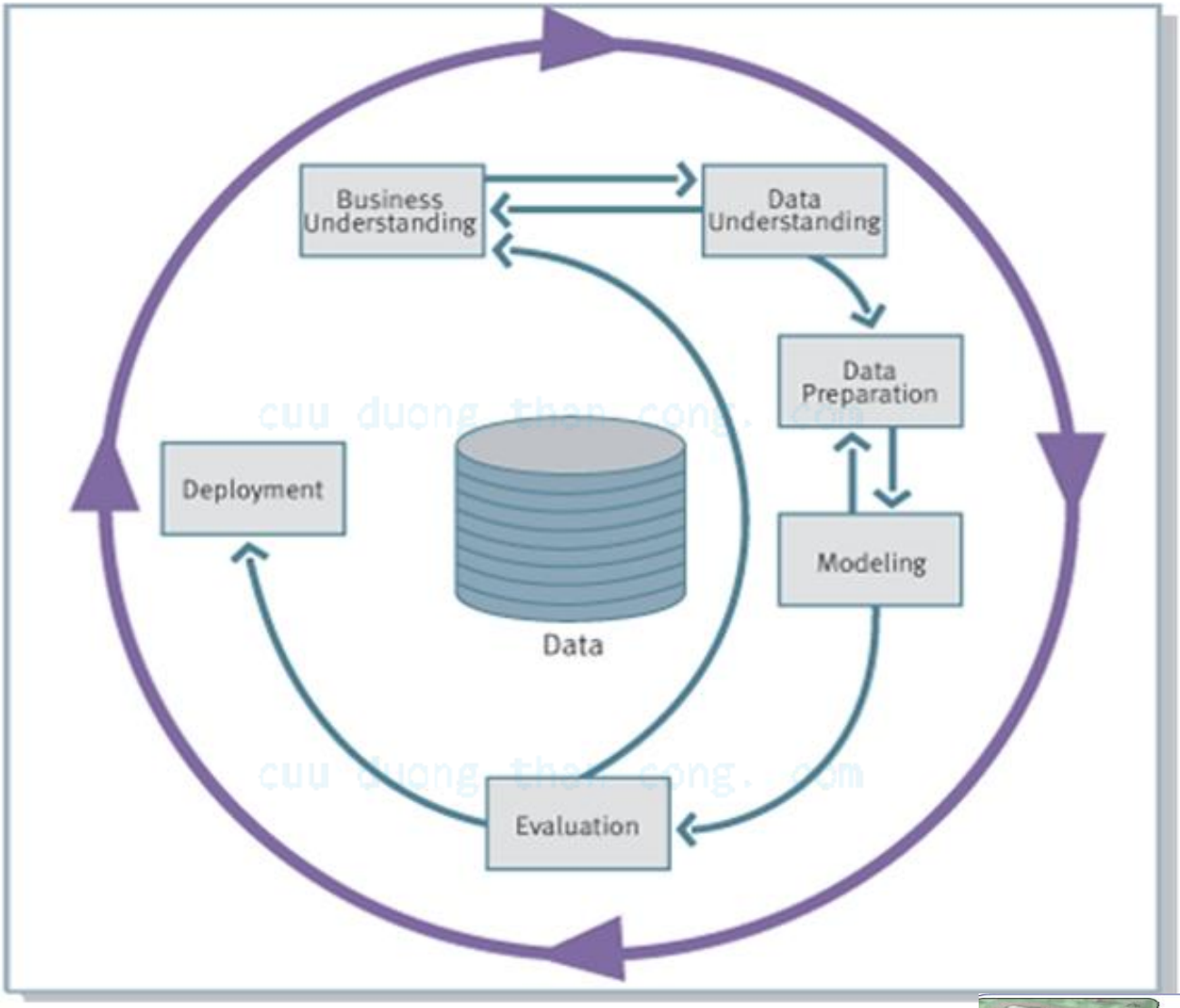
CuuDuongThanCong.com https://fb.com/tailieudientucntt

Piatetsky-Shapiro View



CuuDuongThanCong.com
https://fb.com/tailieudientucntt

CRISP-DM View (Cross Industry Standard Process for Data Mining)



CuuDuongThanCong.com
https://fb.com/tailieudientucntt

Data Mining Functions

All Data Mining functions can be thought of as attempting to find a model to fit the data.

Each function needs Criteria to create one model over another.

Each function needs a technique to Compare the data.

cuu duong than cong. com

Two types of model:

- Predictive models predict unknown values based on known data
- Descriptive models identify patterns in data

cuu duong than cong. com

Each type has several sub-categories, each of which has many algorithms. We won't have time to look at ALL of them in detail.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Data Mining Functions

Data Mining

- Predictive
  - Classification: Maps data into predefined classes
  - Regression: Maps data into a function
  - Prediction: Predict future data states
  - Time Series Analysis: Analyze data over time

  (Supervised Learning)
- Descriptive
  - Clustering: Find groups of similar items
  - Association Rules: Find relationships between items
  - Characterisation: Derive representative information
  - Sequence Discovery: Find sequential patterns

  (Unsupervised Learning)

cuu duong than cong. com

CuuDuongThanCong.com
https://fb.com/tailieudientucntt

Classification

The aim of classification is to create a model that can predict the 'type' or some category for a data instance that doesn't have one.

Two phases:

1. Given labelled data instances, learn model for how to predict the class label for them. (Training)
2. Given an unlabelled, unseen instance, use the model to predict the class label. (Prediction)

Some algorithms predict only a binary split (yes/no), some can predict 1 of N classes, some give probabilities for each of N classes.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Clustering

The aim of clustering is similar to classification, but without predefined classes.

Clustering attempts to find clusters of data instances which are more similar to each other than to instances outside of the cluster.

cuu duong than cong. com

Unsupervised Learning: learning by observation, rather than by example.

Some algorithms must be told how many clusters to find, others try to find an 'appropriate' number of clusters.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Association Rule Mining

The aim of association rule mining is to find patterns that occur in the data set frequently enough to be interesting. Hence the association or correlation of data attributes within instances, rather than between instances.

These correlations are then expressed as rules - if X and Y appear in an instance, then Z also appears.

Most algorithms are extensions of a single base algorithm known as 'A Priori', however a few others also exist.

cuu duong than cong. com

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Why?

That all sounds ... complicated. Why should I learn about Data Mining?

What's wrong with just a relational database? Why would I want to go through these extra [complicated] steps?

Isn't it expensive? It sounds like it takes a lot of skill, programming, computational time and storage space. Where's the benefit?

Data Mining isn't just a cute academic exercise, it has very profitable real world uses. Practically all large companies and many governments perform data mining as part of their planning and analysis.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

The Data Explosion

The rate of data creation is accelerating each year. In 2003, UC Berkeley estimated that the previous year generated 5 exabytes of data, of which 92% was stored on electronically accessible media.

Mega < Giga < Tera < Peta < Exa ... All the data in all the books in the US Library of Congress is ~136 Terabytes. So 37,000 New Libraries of Congress in 2002.

VLBI Telescopes produce 16 Gigabytes of data every second.

Each engine of each plane of each company produces ~1 Gigabyte of data every trans-atlantic length journey. Google searches 18 billion+ accessible web pages.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Data Explosion Implications

As the amount of data increases, the proportion of information decreases.

As more and more data is generated automatically, we need to find automatic solutions to turn those stored raw results into information.

cuu duong than cong. com

Companies need to turn stored data into profit ... otherwise why are they storing it?

Let's look at some real world examples.

cuu duong than cong. com

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Classification

The data generated by airplane engines can be used to determine when it needs to be serviced. By discovering the patterns that are indicative of problems, companies can service working engines less often (increasing profit) and discover faults before they materialise (increasing safety).

cuu duong than cong. com

Loan companies can “give you results in minutes” by classifying you into a good credit risk or a bad risk, based on your personal information and a large supply of previous, similar customers.

cuu duong than cong. com

Cell phone companies can classify customers into those likely to leave, and hence need enticement, and those that are likely to stay regardless.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Clustering

Discover previously unknown groups of customers/items. By finding clusters of customers, companies can then determine how best to handle that particular cluster.

For example, this could be used for targeted advertising, special offers, transferring information gathered by association rule mining to other members of the cluster, and so forth.

The concept of 'Similarity' is often used for determining other items that you might be interested in, eg 'More Like This' links.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Association Rule Mining

By finding association rules from shopping baskets, supermarkets can use this information for many things, including:

- Product placement in the store
- What to put on sale
- What to create as 'joint special offers'
- What to offer the customer in terms of coupons
- What to advertise together

It shouldn't be surprising that your Tesco coupons are for things that you sometimes buy, rather than things you always or never buy.

Wal-Mart in the US records every transaction at every store -- petabytes of information to sift through. (TeraData)

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Data/Information/Knowledge/Wisdom

Note well that data mining applications have no wisdom. They cannot apply the knowledge that they discover appropriately.

For example, a data mining application may tell you that there is a correlation between buying music magazines and beer, but it doesn't tell you how to use that knowledge. Should you put the two close together to reinforce the tendency, or should you put them far apart as people will buy them anyway and thus stay in the store longer?

cuu duong than cong. com

Data mining can help managers plan strategies for a company, it does not give them the strategies.

CuuDuongThanCong.com https://fb.com/tailieudientucntt