



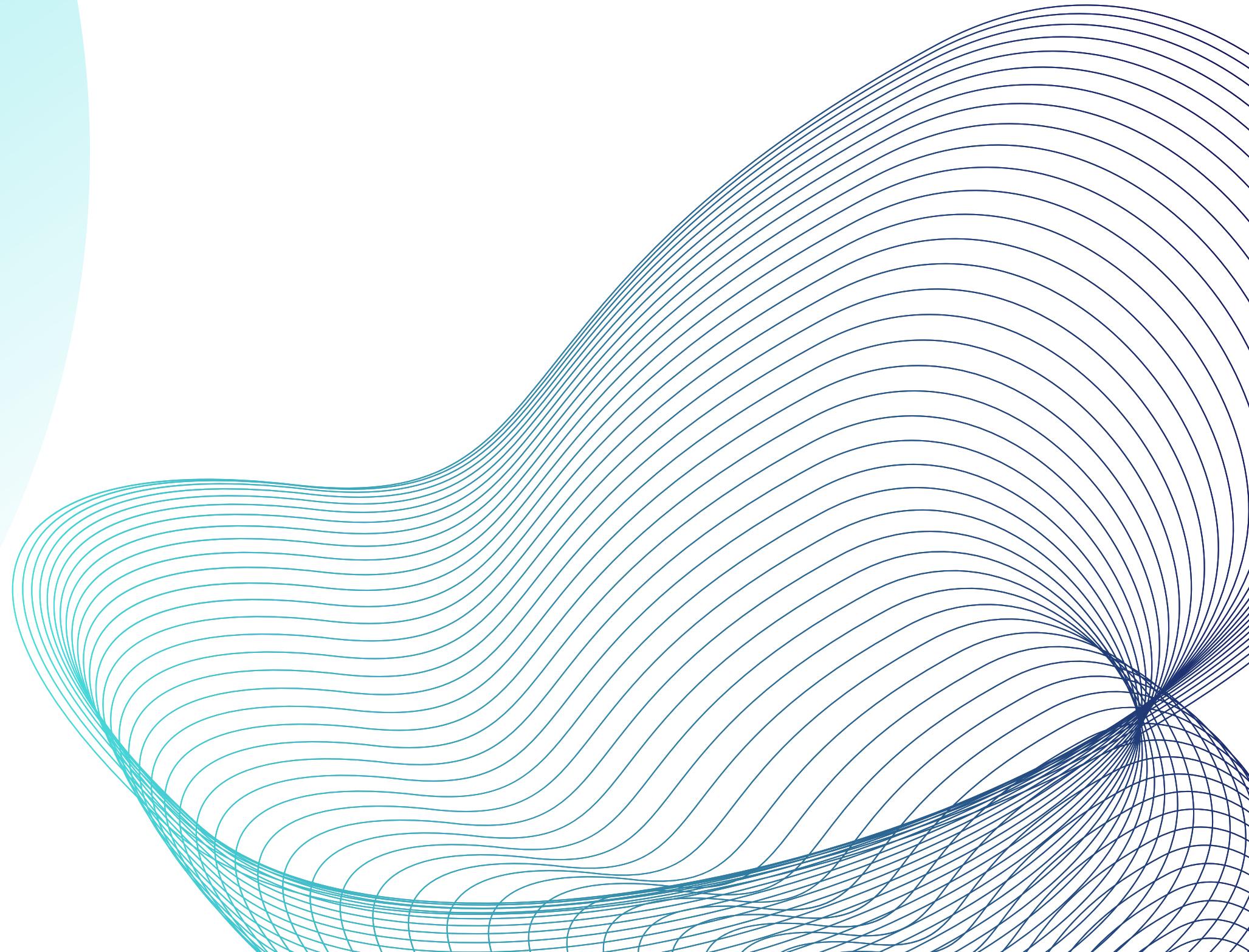
INTRODUCTION TO ARTIFICIAL INTELLIGENCE

PGS. TS. Nguyễn Thanh Bình



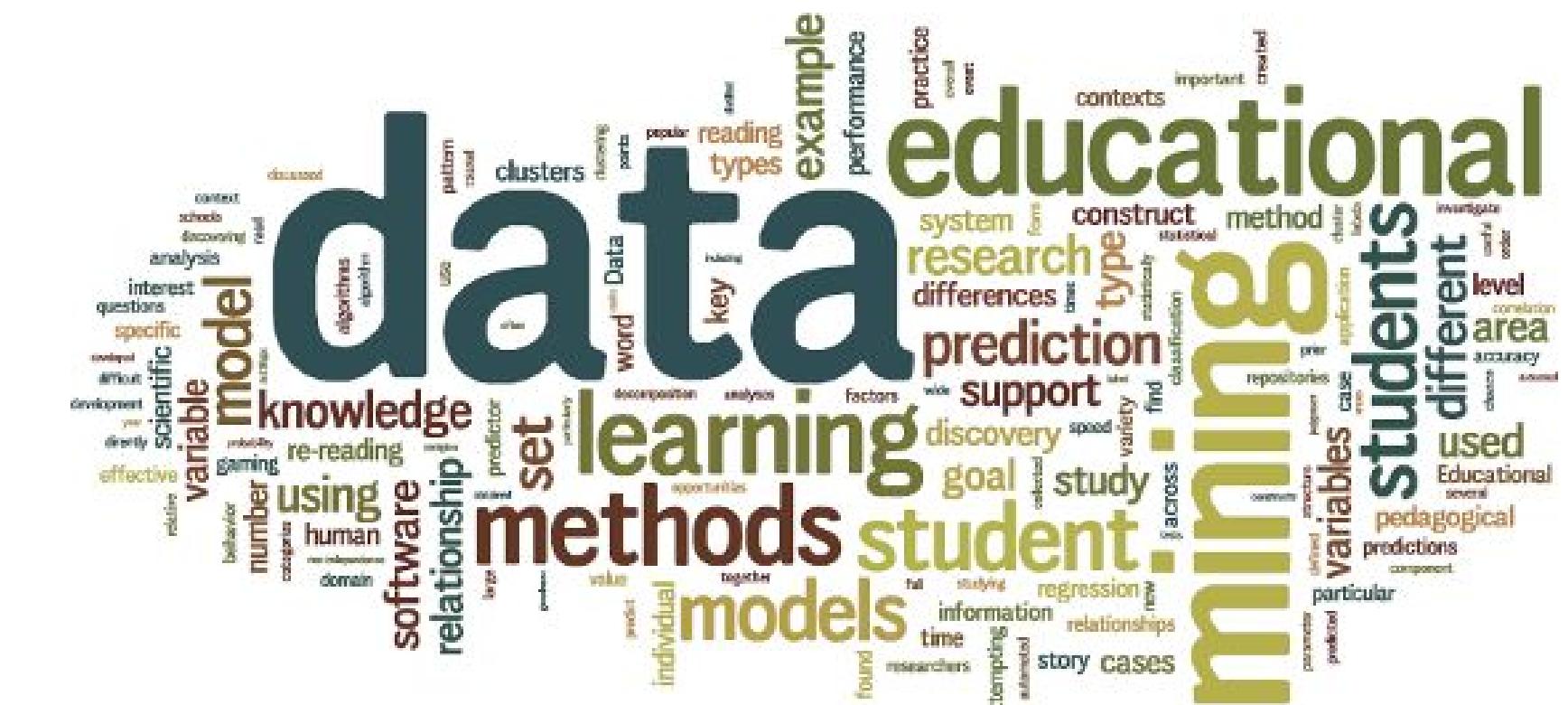


DATA-MINING PROCESS



WHAT IS DATA-MINING?

- (Wikipedia) Extract information from a dataset and transform it into an understandable structure for further use.
 - To discover the patterns and relationships in the data in order to help make better business decisions



WHAT IS DATA-MINING?

- Databases today can range in size into the terabyte



EBay

- 532 nodes cluster (8 * 532 cores, 5.3PB).
- Heavy usage of Java MapReduce, Pig, Hive, HBase
- Using it for Search optimization and Research.



shine_production 16.1806640625 GB

5.3 PB = 5300 TB = 5300.000 GB



Facebook

- We use Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and research.
- Currently we have 2 major clusters:
 - A 1100-machine cluster with 8800 cores and about 12 PB raw storage.
 - A 300-machine cluster with 2400 cores and about 3 PB raw storage.
 - Each (commodity) node has 8 cores and 12 TB of storage.
 - We are heavy users of both streaming as well as the Java APIs. We have built a higher level data warehousing framework (<http://hadoop.apache.org/hive/>). We have also developed a FUSE implementation over HDFS.



Spotify

- We use Hadoop for content generation, data aggregation, reporting and analysis (see more: [Hadoop at Spotify](#))
- 690 node cluster = 8280 physical cores, 38TB RAM, 28 PB storage
- +7,500 daily Hadoop jobs (scheduled by Luigi, our home-grown and recently open-sourced job scheduler - [code](#) and [video](#))

WHAT IS DATA-MINING?

- Within these mass of data lies lots of crucial and hidden information.



EBay

- 532 nodes cluster (8 * 532 cores, 5.3PB).
- Heavy usage of Java MapReduce, Pig, Hive, HBase
- Using it for Search optimization and Research.



shine_production 16.1806640625 GB

5.3 PB = 5300 TB = 5300.000 GB



Facebook

- We use Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and research.
- Currently we have 2 major clusters:
 - A 1100-machine cluster with 8800 cores and about 12 PB raw storage.
 - A 300-machine cluster with 2400 cores and about 3 PB raw storage.
 - Each (commodity) node has 8 cores and 12 TB of storage.
 - We are heavy users of both streaming as well as the Java APIs. We have built a higher level data warehousing framework (<http://hadoop.apache.org/hive/>). We have also developed a FUSE implementation over HDFS.



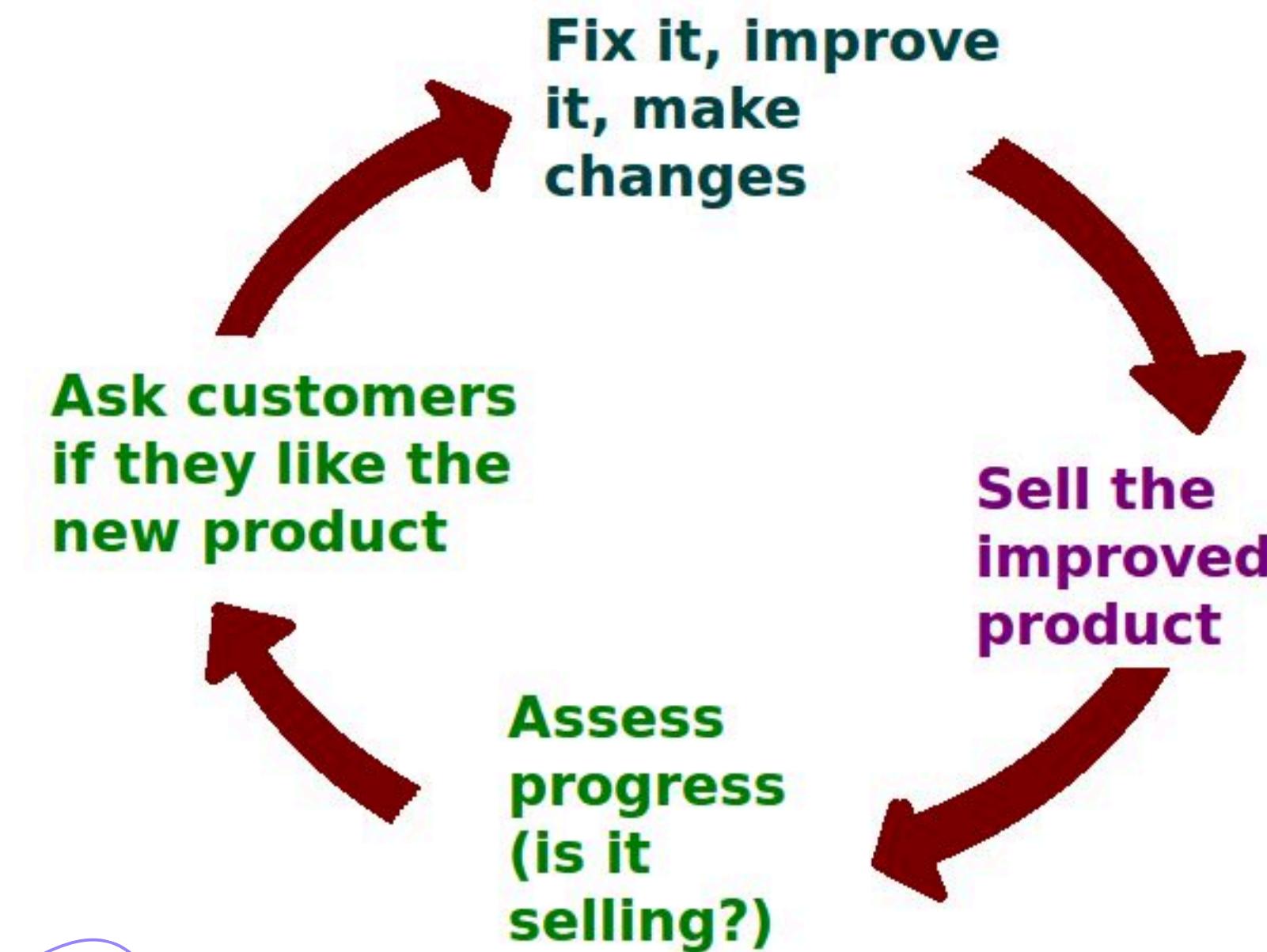
Spotify

- We use Hadoop for content generation, data aggregation, reporting and analysis (see more: [Hadoop at Spotify](#))
- 690 node cluster = 8280 physical cores, 38TB RAM, 28 PB storage
- +7,500 daily Hadoop jobs (scheduled by Luigi, our home-grown and recently open-sourced job scheduler - [code](#) and [video](#))

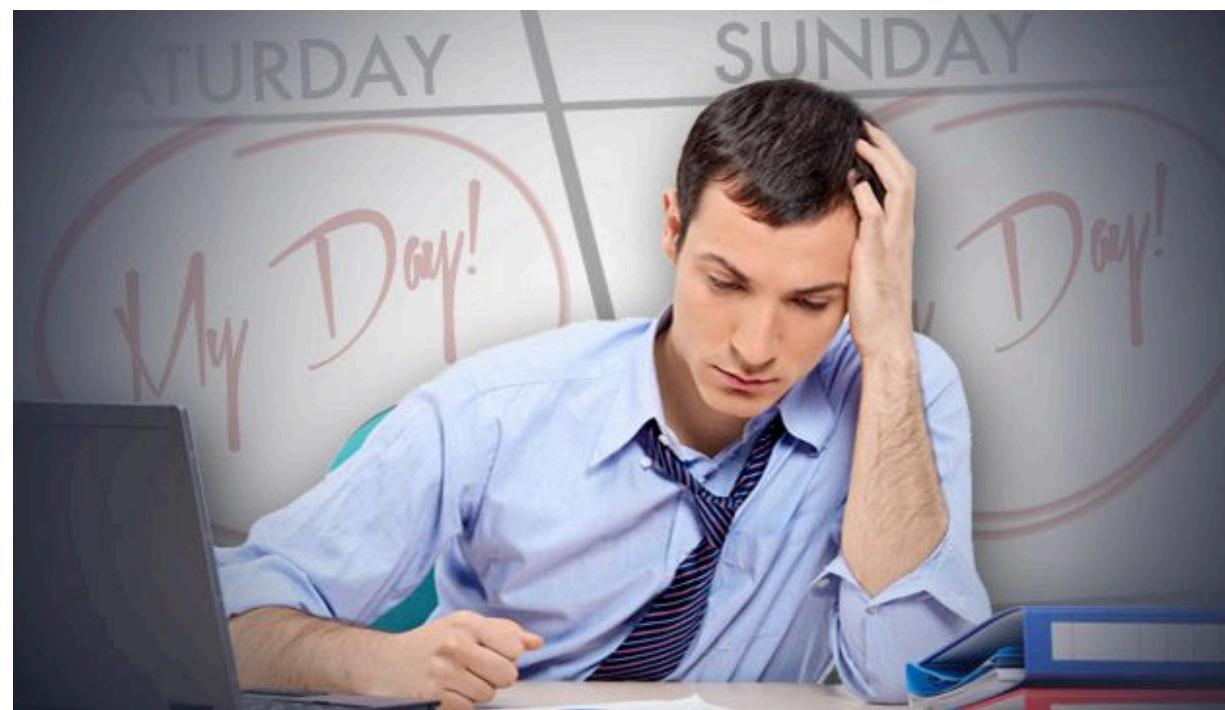
WHAT IS DATA-MINING?



WHAT IS DATA-MINING?

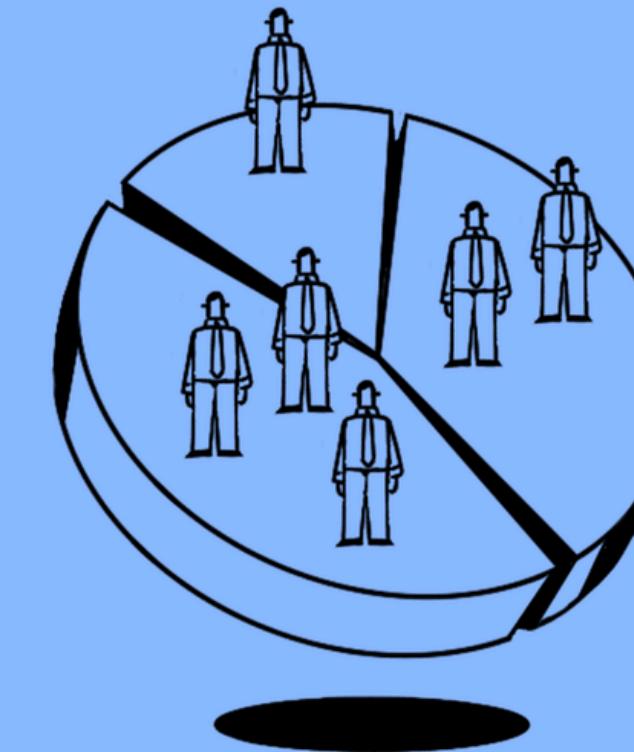


WHAT IS DATA-MINING?



WHAT CAN DATA MINING DO?

- **Market segmentation:** identify the common characteristics of customers who buy the same products from our company.



Market Segmentation

[mär-ket ,seg-men'-shen]

A marketing term that refers to aggregating prospective buyers into groups or segments with common needs and who respond similarly to a marketing action.

WHAT CAN DATA MINING DO?

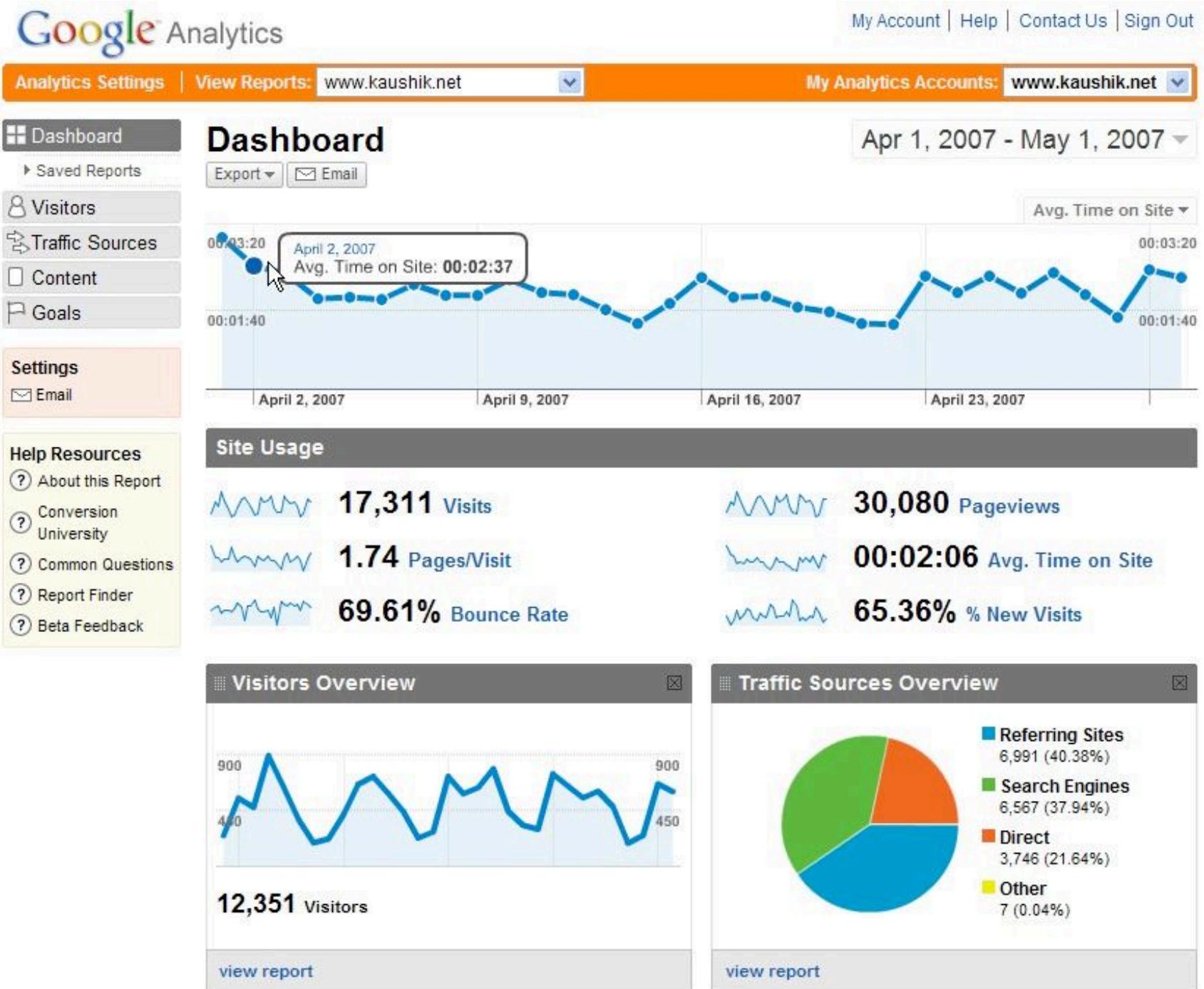
- **Customer churn:** predict which customers are likely to leave our company and go to the competitor





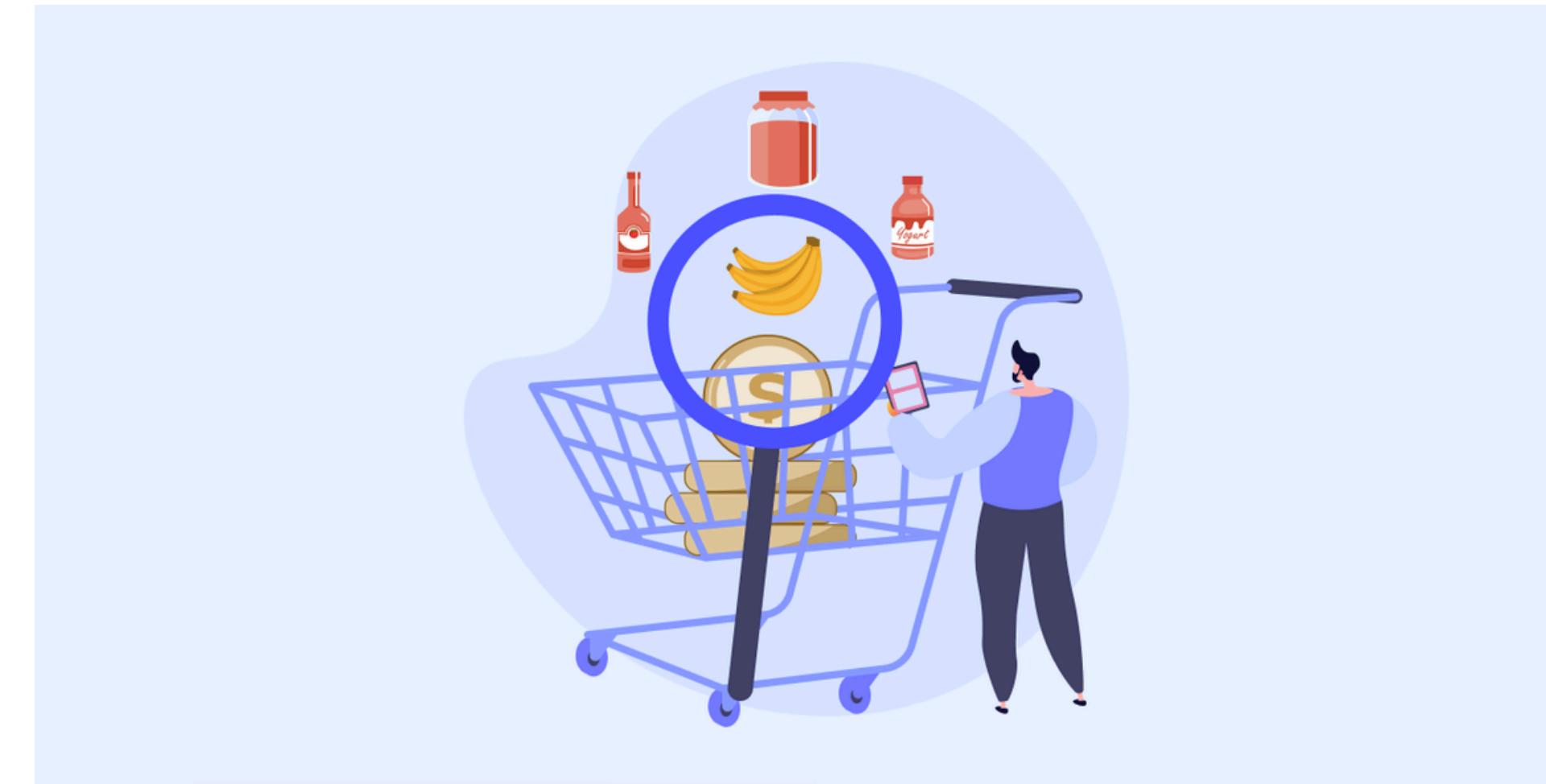
WHAT CAN DATA MINING DO?

- **Interactive marketing:** predict what each individual accessing our Website is mostly likely interested in seeing



WHAT CAN DATA MINING DO?

- **Market basket analysis:** understand what products or services are commonly purchased together.



WHAT CAN DATA MINING DO?

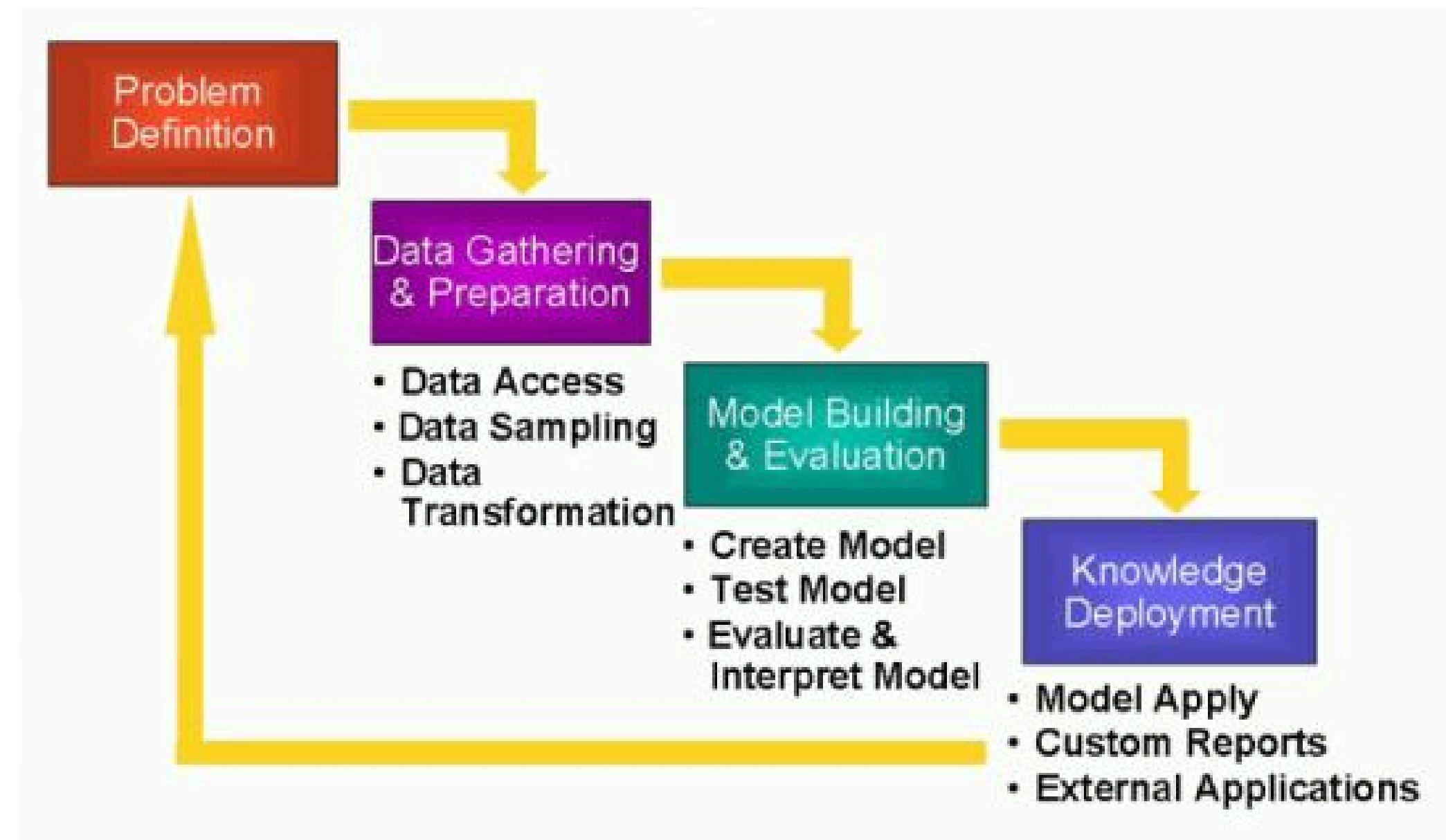
- **Automated prediction of trends and behaviors:** data mining automates the process of finding predictive information in a large database



WHAT CAN DATA MINING DO?

- **Automated discovery of previously unknown patterns:** data mining tools sweep through databases and identify previously hidden patterns.
 - For example: analyze the sales data in an company to identify seemingly unrelated products that are often purchased together.

DATA-MINING PROCESS

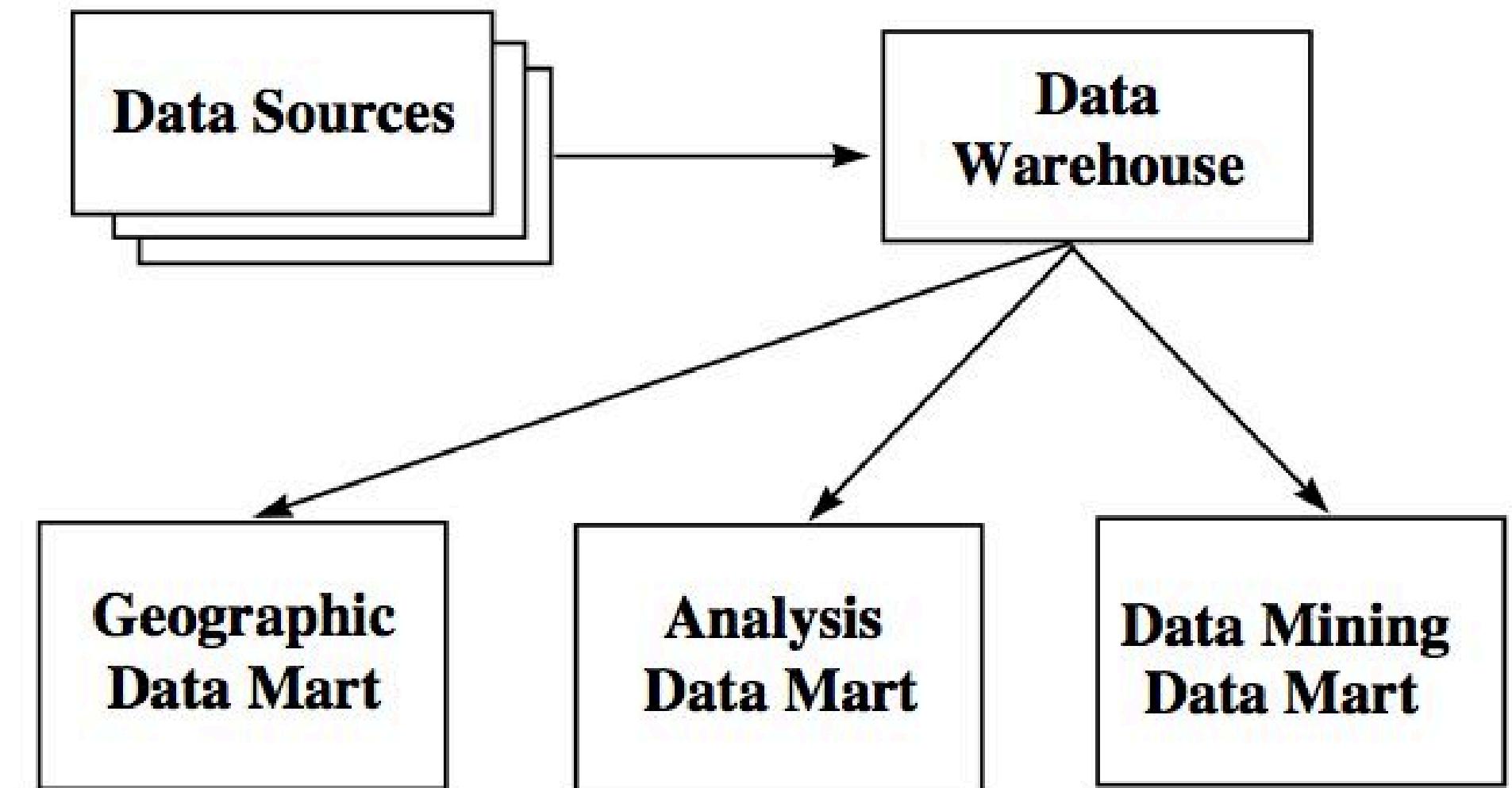


DEFINE BUSINESS PROBLEM

- Understand our data and business
- Clear statements of our objectives.
- Examples:
 - **Problem:** increase the response of a direct mail campaign.
 - Model 1: “Increasing the response rate”
 - Model 2: “Increase the value of a response”

BUILD DATA-MINING DATABASE

- Usually take anywhere from 50% to 90% time and efforts of the entire knowledge discovery process



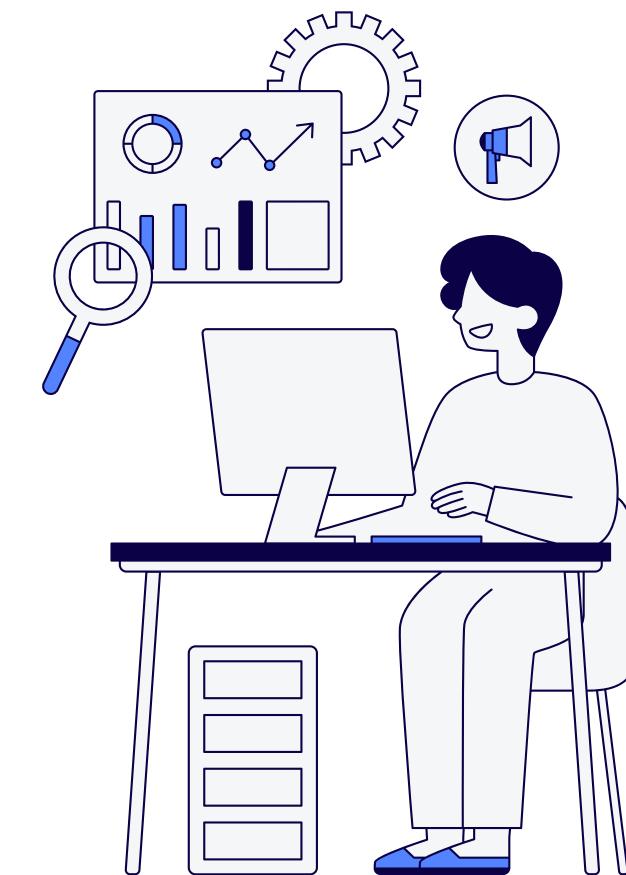
EXPLORE THE DATA

- Identify the most important fields in predicting an outcome and determine which derived values may be useful.



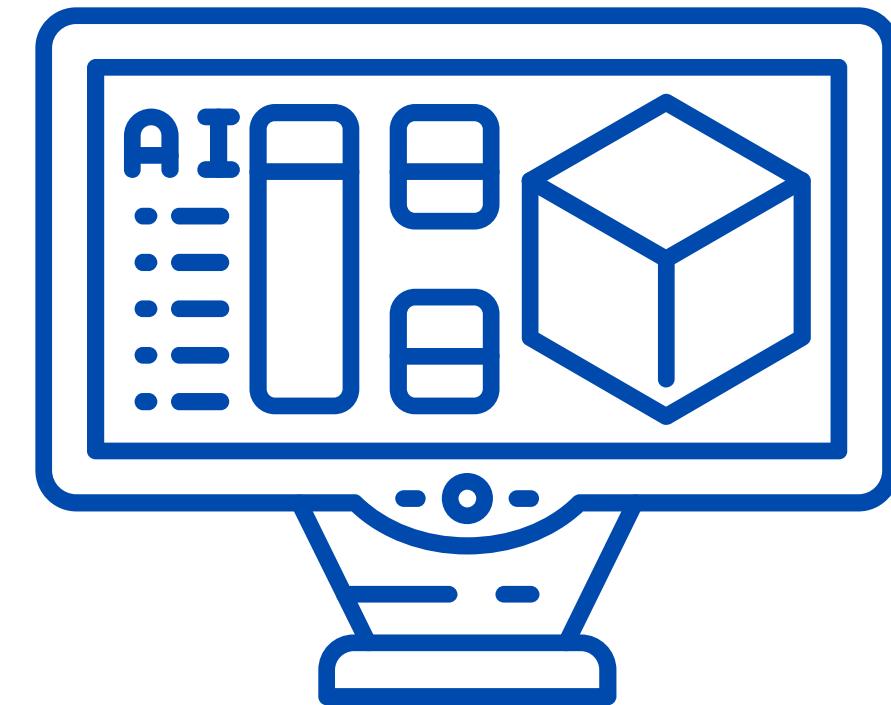
PREPARE DATA FOR MODELING

- This is the final data preparation step before building models.
- Have four main parts:
 - Select variables.
 - Select rows.
 - Construct new variables.
 - Transform variables.



DATA MINING MODEL BUILDING

- Explore alternative models to find the best one in solving our business problem.
- Choose a model type for making the prediction.
- Require a good training and validation protocol (supervised learning) for accurate and robust predictions



FINAL STEPS

- Evaluate the model and interpret the significance of its results.
- The accuracy rate applies only to which the model was built.
- When a data mining model was built and validated, it can be used to recommend actions or to apply the model to various data sets.



FINAL STEPS

- Evaluate the model and interpret the significance of its results.
- The accuracy rate applies only to which the model was built.
- When a data mining model was built and validated, it can be used to recommend actions or to apply the model to various data sets.

