
THE DATA MINER
EUREKA! I FOUND A CORRELATION.
www.dilbert.com scottadams@aol.com
WHEN YOU'RE ON VACATION, ALL YOUR EMPLOYEES TELECOMMUTE.
THEY DO?
© 1999 United Feature Syndicate, Inc.
AND 100% OF ALL EXPENSE VOUCHERS ARE SIGNED WHEN YOU'RE OUT SICK.
WE HAVE VOUCHERS?

DM: Intro

What is Data?

Data is a set of facts/observations/measurements about objects/events/processes of interest

Higher Standards meijer Lower Prices

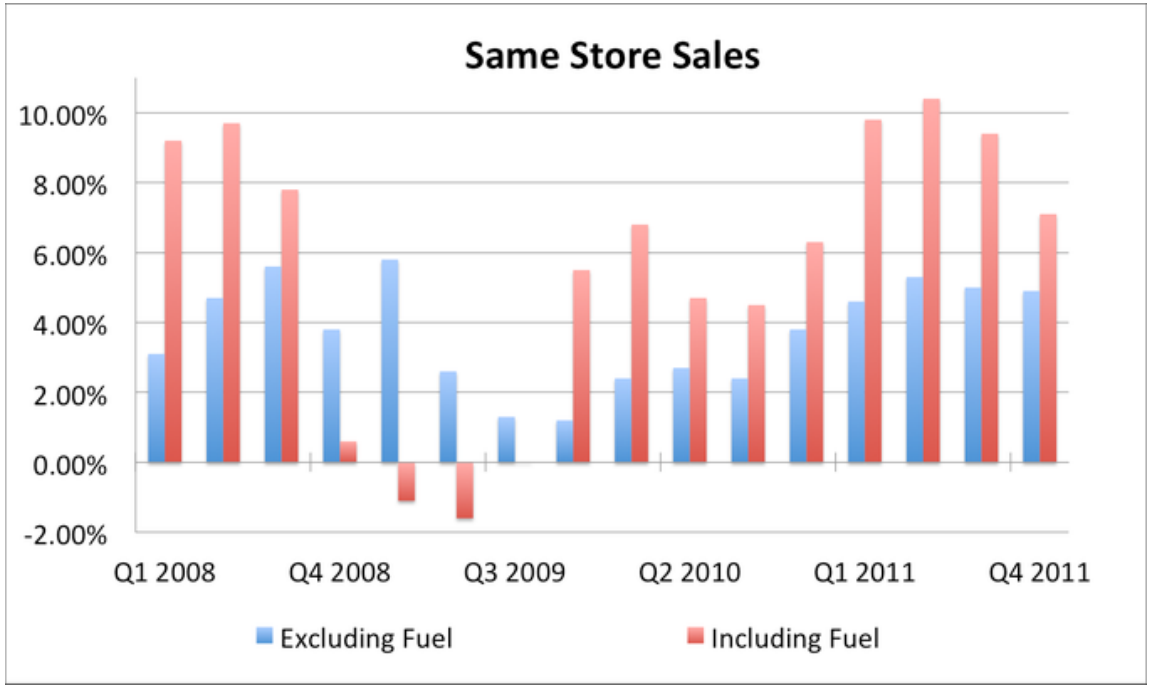
Elida Rd.
Lima, OH - # 110
(419) 331 6400 meijer.com

The Meijer Team appreciates your business
05/28/11
Your fast and friendly checkout was
provided by Fastlane107

```
****SAVINGS TODAY****
* TOTAL MEIJER PROMOTIONS          1.00 *
* TOTAL NON-COUPON SAVINGS         5.12 *
* TOTAL COUPON SAVINGS OF          4.49 *
*SAVINGS TOTAL  10.61*

GROCERY
*2670012911   FRENCH DIP
     was      1.69      now        .99  F
     mPerks Offer
 => 1.00 off                      -.99  F
    * Limit of .99 reached
*2670032200   DEANS DIP
     was      1.69      now        .99  F
     mPerks Offer
 => 1.00 off                      -.01  F
    * Limit of 1.00 reached
*3760028225   SALSA
     was      2.43      now        .99  F
*4335400790   TORTILLAS
     was      2.39      now       1.89  F
*1901401852   DOG FOOD
     2 @ 16.99
     was     35.76      now      33.98  T

COUPONS
              Vendor Coupon       -.50  F
              EXTRA COUPON        -.50  F
81101003760008 Vendor Coupon      -.50  F
81101003760008 EXTRA COUPON       -.49  F
81101001901403 Vendor Coupon     -1.00  N
81101001901403 Vendor Coupon     -1.00  N
81101002670002 Vendor Coupon      -.50R F

          Mperks # -- ********63
       Meijer Card -- 00900875
Monthly purch (up to 48 hr delay)   .00
TOTAL
          TOTAL TAX              2.21
          TOTAL                 35.56
```

What is Information?

Information is processed data that is useful in one way or the other, for example for decision making, communication etc.



While the data is fixed, information from it can differ based on needs

What is Knowledge?

Patterns of relationships in data and information that exhibit a high degree of certainty

What Is Data Mining?

Data mining is essentially a process of data-driven extraction of not so obvious but useful information from large databases. The entire process is interactive and iterative.

Data mining also goes under various other names such as:

Knowledge discovery in databases (KDD), knowledge extraction, data/pattern analysis, data analytics, business intelligence, etc.

What is data science?

Data science can be broken down into four essential parts.

Mining data

Collecting and formatting the information

Statistics

Information analysis

Interpret

Representation or visualization in the form of presentations, infographics, graphs or charts

Leverage

Implications of the data, application of the data, interaction using the data and predictions formed from studying it

What is Data Science?

The future belongs to the companies and people that turn data into products

Mike Loukides

Desperately Seeking Data Scientists

The role of data scientist is hotter than ever. American Express is looking for a whole team of them and many other companies are, too. Aren't internal staff worth retraining?

Posted June 21, 2012 to **Business Intelligence** | 1 Comment + Briefcase

I came across an intriguing job ad recently that shows just how important data scientists will be. The role of a data scientist isn't new but now it's hotter than ever. American Express is looking for a whole team of data scientists, including a senior leader for a role that sounds pretty darn critical to the future of the $32.3 billion financial services company.

Latest buzz word

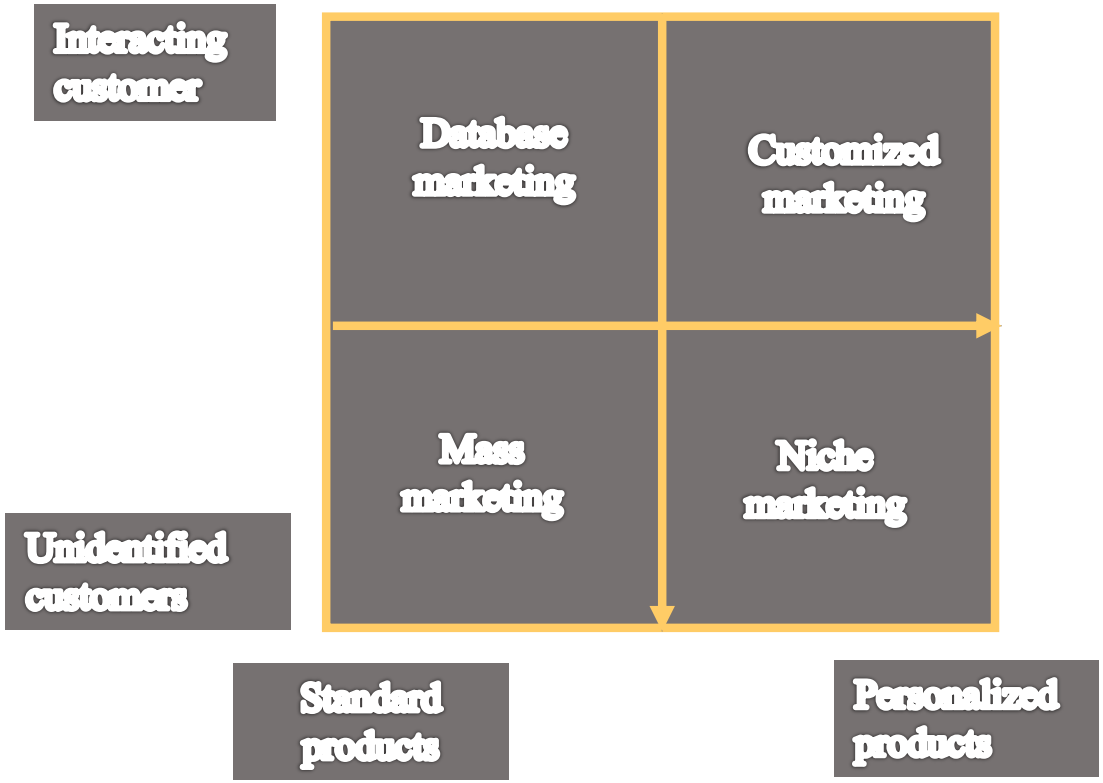
Why Data Mining?

- The Data Glut
  - Data rich but information poor businesses
  - Estimates of data doubling every 20 months
  - The average Fortune 500 company manages over a terabyte of data everyday
- Convergence of Technologies
- Competitive Edge
  - Mass marketing versus targeted marketing

"It's free, but they sell your information."

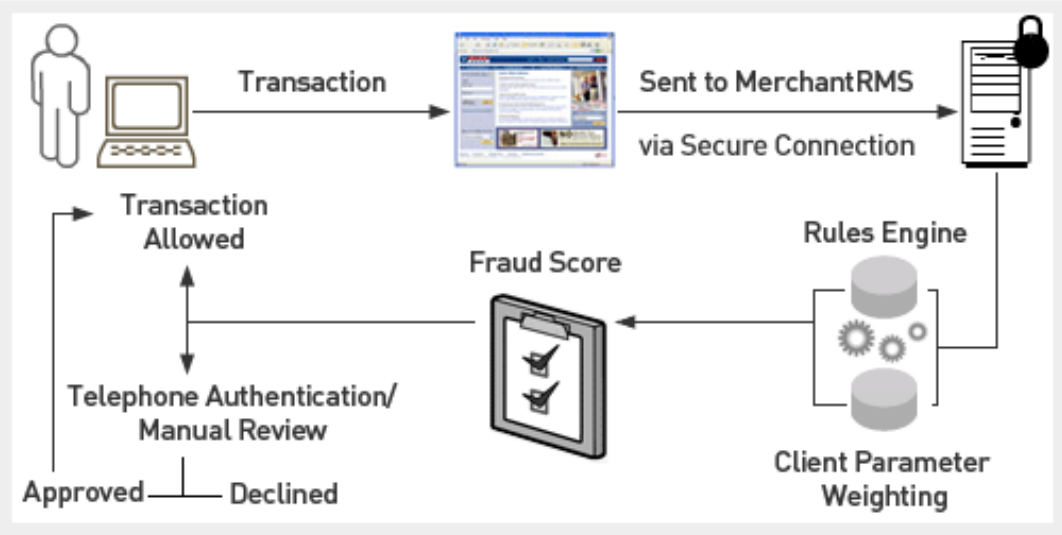

MARCH 21, 2011
A tale of two Libyas
Plus: Why the U.S. can't sit on the sidelines BY FAREED ZAKARIA
The GOP's misinformation campaign
BY JOE KLEIN
Could your baby be depressed?
THE CULTURE
Word up: A dictionary of slang
TIME
YOUR DATA FOR SALE
Everything about you is being tracked—get over it
BY JOEL STEIN
What data-mining companies think they know about Joel Stein
Household income: $100,000+
Owns a laptop
Age: 38-39
Likes: fashion
Major life-insurance holder
Likes: cooking & recipes
Likes: online news
Lives in New York City
Likes: Asian cuisi
Likes: online shopping
Dislikes: cars
Likes: green living
Favorite celebrities:
ZIP code: 1070
Wi-fi warrior
Politically active
Sister is a la
Religion: Jewish
Job: medical professional
Likes: parenting
Spent $180 on intimate app. & underwarments on Oct. 10, 2010
Male
Mother: Rosalind Bard
Likes: hiking
Household income: $150,000-$175,000
Owns a smart phone
Married
Dislikes: autos & vehicles
Likes: retail
Works at company with 5,000+ employees
Likes: newspapers
Likes: movies
Magazine subscriber
Smart-phone user
Sister: Lisa Stein Browning
Purchased house in month of November
Has used cocaine
Has had LASIK surgery
Likes: restaurants

Changing Business Direction



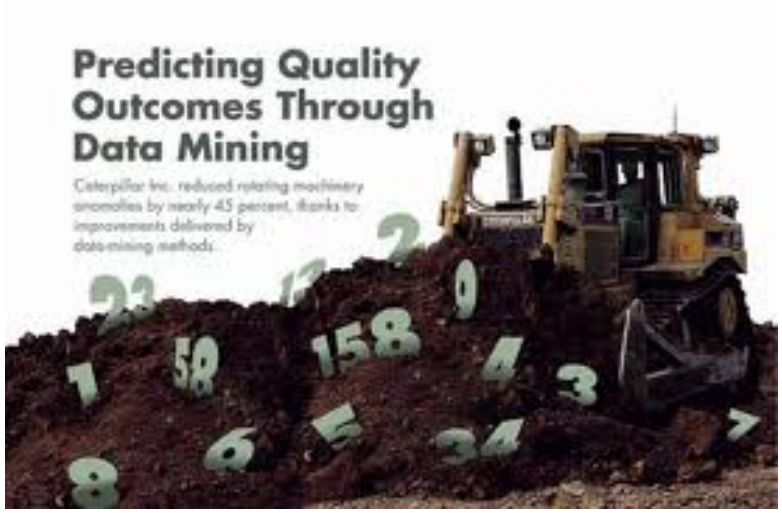
Typical Business Applications of Data Mining

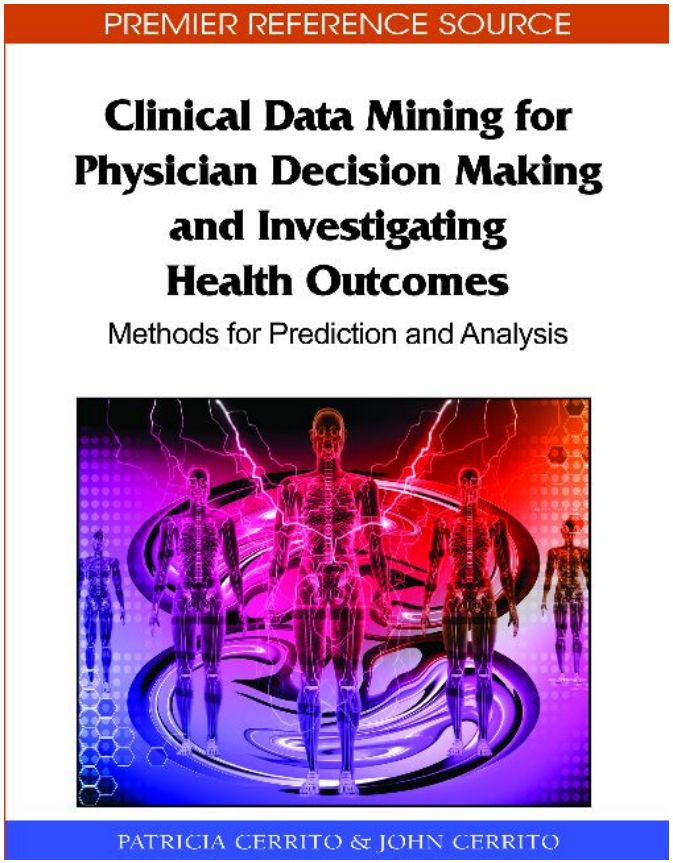
- Market Segmentation
- Customer Targeting and Retention
- Product Design and Placement
- Credit Card Fraud Detection
- Web Advertising
- Recommendation Systems



Many Applications of Data Mining

- Stock Market Trends
- Text and Multimedia Data Mining
- Sports Scouting
- Medical Outcomes Analysis
- Scientific Data Mining





Data Classification

- Structured Data
  - Data consisting of well-defined fields of numeric or alphanumeric values

ORDER FILE

Order Number	Order Date	Customer Number	Delivery Address	Concrete Type	Amount	Truck Number	Driver ID
100000	9/1/2004	1234	55 Smith Lane	1	8	111	123456789
100001	9/1/2004	3456	2122 E. Biscayne	1	3	222	785934444
100002	9/2/2004	1234	55 Smith Lane	5	6	222	435296657
100003	9/3/2004	4567	1333 Burr Ridge	2	4	333	435296657
100004	9/4/2004	4567	1333 Burr Ridge	2	8	222	785934444
100005	9/4/2004	5678	1222 Westminster	1	4	222	785934444
100006	9/5/2004	1234	222 East Hampton	1	4	111	123456789
100007	9/6/2004	2345	9 W. Palm Beach	2	5	333	785934444
100008	9/6/2004	6789	4532 Lane Circle	1	8	222	785934444
100009	9/7/2004	1234	987 Furlong	3	8	111	123456789
100010	9/9/2004	6789	4532 Lance Circle	2	7	222	435296657
100011	9/9/2004	4567	3500 Tomahawk	5	6	222	785934444

CUSTOMER FILE

Customer Number	Customer Name	Customer Phone	Customer Primary Contact
1234	Smelding Homes	3333333333	Bill Johnson
2345	Home Builders Superior	3334444444	Marcus Connolly
3456	Mark Akey	3335555555	Mark Akey
4567	Triple A Homes	3336666666	Janielle Smith
5678	Sheryl Williamson	3337777777	Sheryl Williamson
6789	Home Makers	3338888888	John Yu

EMPLOYEE FILE

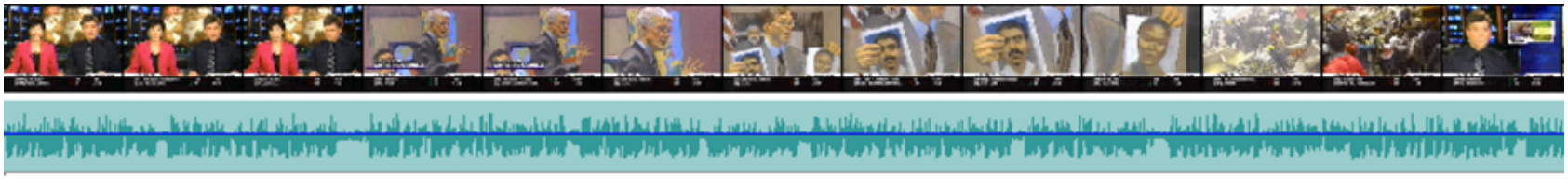
Employee ID	Employee Last Name	Employee First Name	Date of Hire
123456789	Johnson	Emilio	2/1/1985
435296657	Evaraz	Antonio	3/3/1992
785934444	Robertson	John	6/1/1999
984568756	Smithson	Allison	4/1/1997

TRUCK FILE

Truck Number	Truck Type	Date of Purchase
111	Ford	6/17/1999
222	Ford	12/24/2001
333	Chevy	1/1/2002

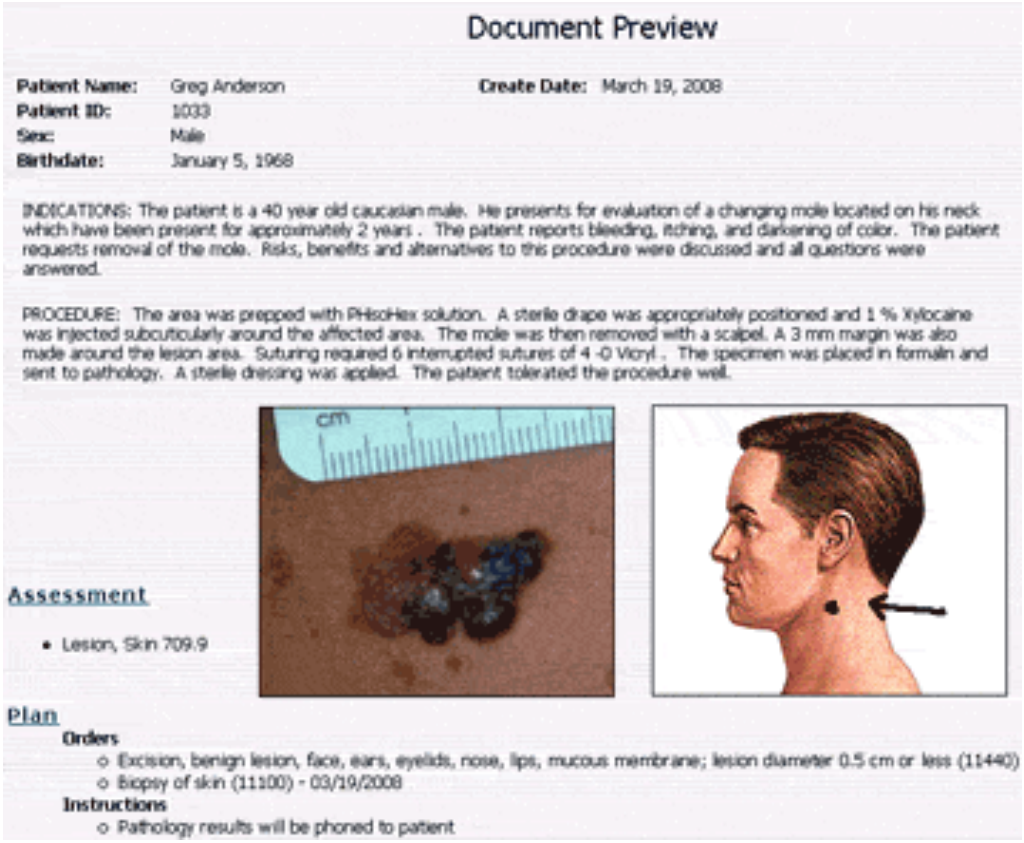
Data Classification

- Unstructured Data
  - No well-defined fields of information
  - Requires extensive processing to extract content information
  - Examples include blogs, news reports, images, videos, tweets etc.
  - Fastest growing data segment



Data Classification

- Semi-Structured Data
  - Data with partial structure (medical reports, executive summaries, interview scripts, web documents etc.)



Document Preview

Patient Name: Greg Anderson **Create Date:** March 19, 2008
Patient ID: 1033
Sex: Male
Birthdate: January 5, 1968

INDICATIONS: The patient is a 40 year old caucasian male. He presents for evaluation of a changing mole located on his neck which have been present for approximately 2 years . The patient reports bleeding, itching, and darkening of color. The patient requests removal of the mole. Risks, benefits and alternatives to this procedure were discussed and all questions were answered.

PROCEDURE: The area was prepped with PHisoHex solution. A sterile drape was appropriately positioned and 1 % Xylocaine was injected subcuticularly around the affected area. The mole was then removed with a scalpel. A 3 mm margin was also made around the lesion area. Suturing required 6 interrupted sutures of 4 -0 Vicryl . The specimen was placed in formalin and sent to pathology. A sterile dressing was applied. The patient tolerated the procedure well.

Assessment

- Lesion, Skin 709.9

Plan

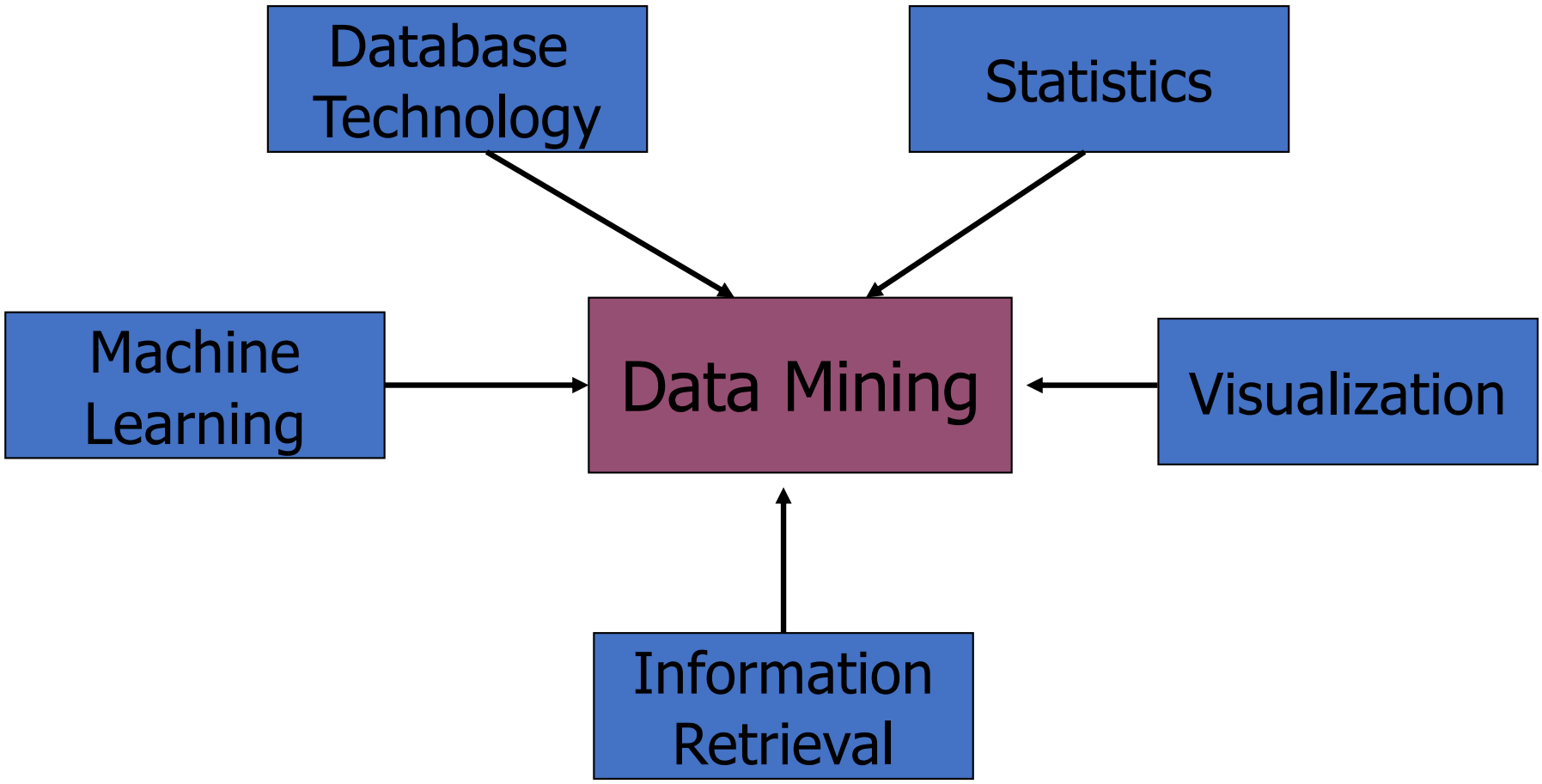
Orders

- Excision, benign lesion, face, ears, eyelids, nose, lips, mucous membrane; lesion diameter 0.5 cm or less (11440)
- Biopsy of skin (11100) - 03/19/2008

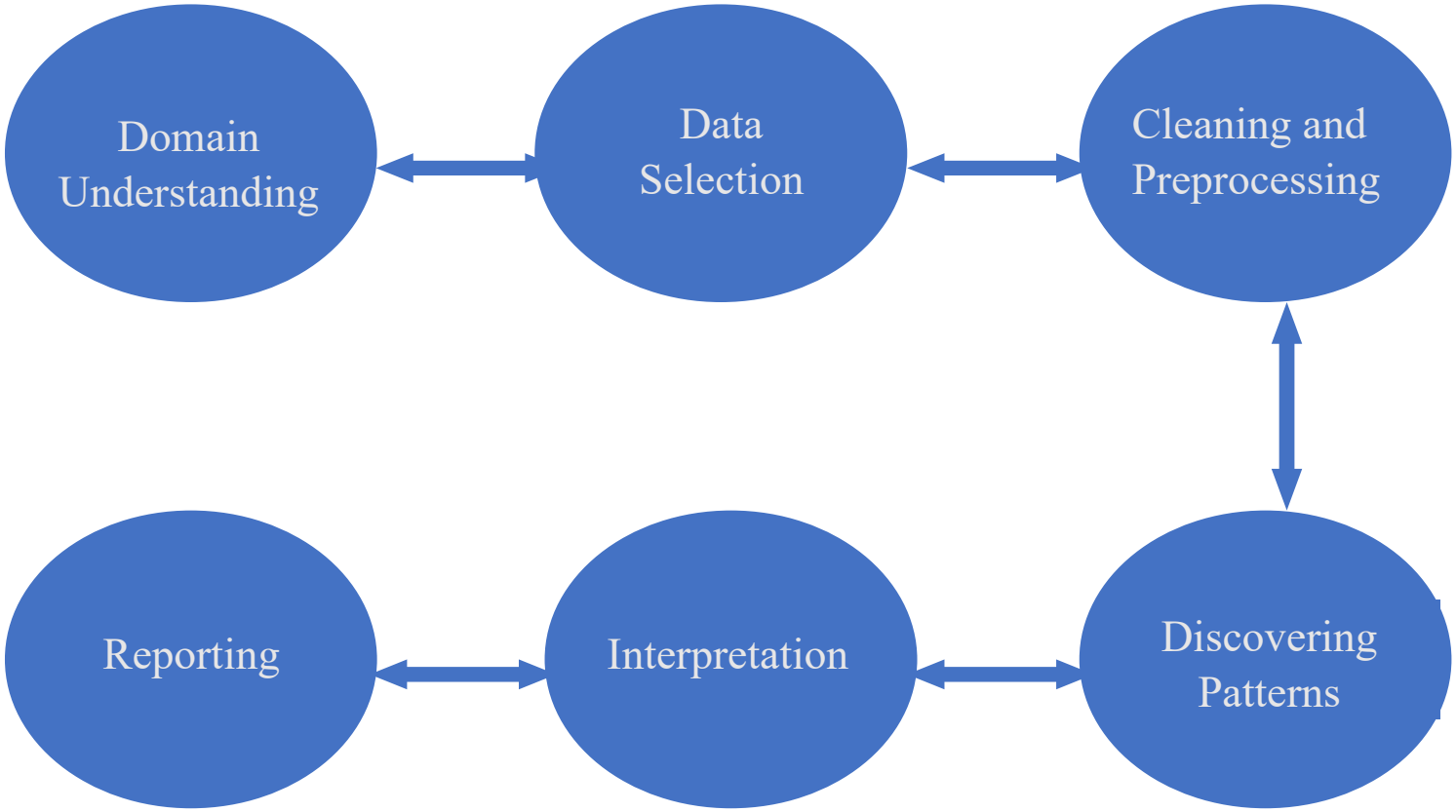
Instructions

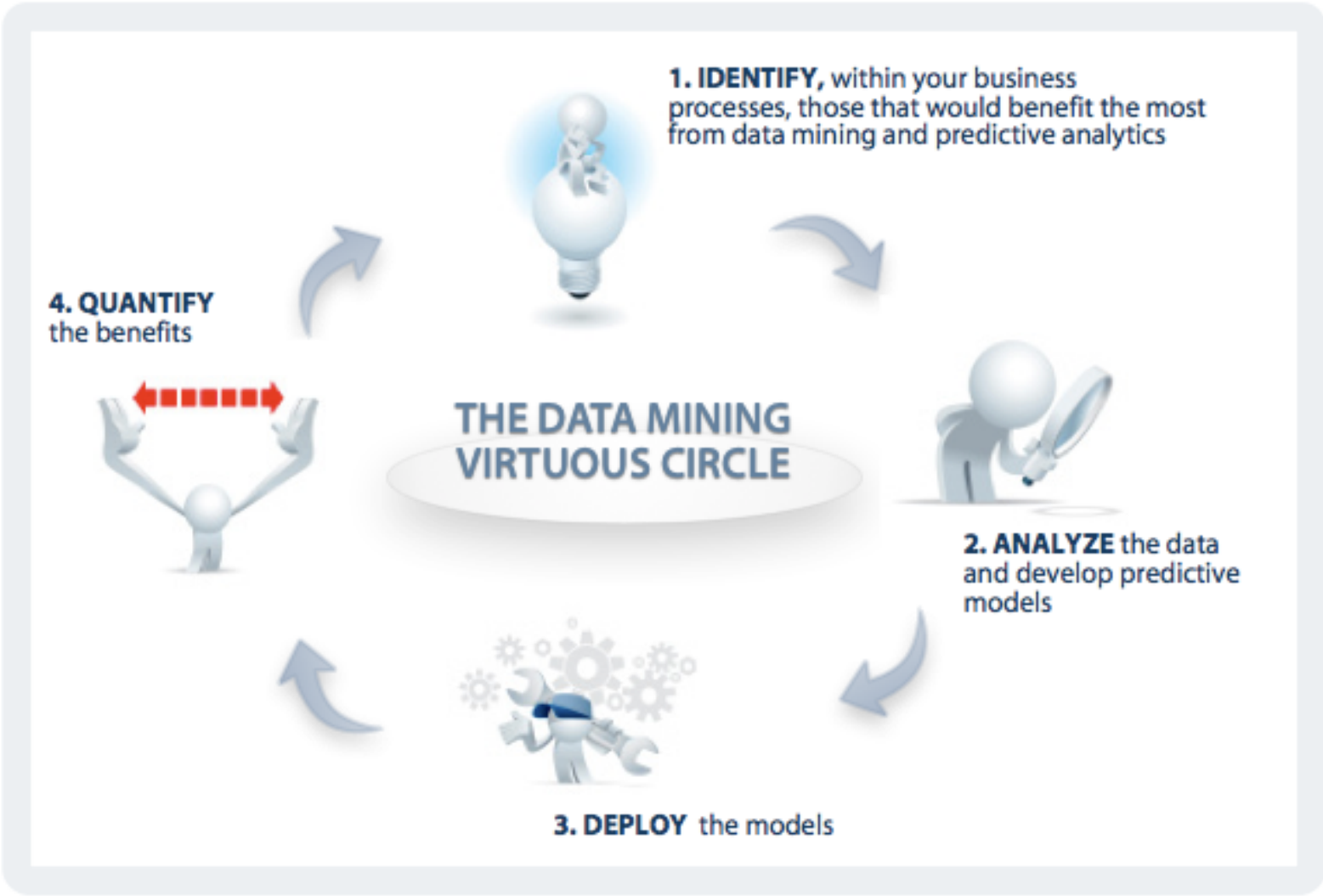
- Pathology results will be phoned to patient

Core Technologies for Data Mining



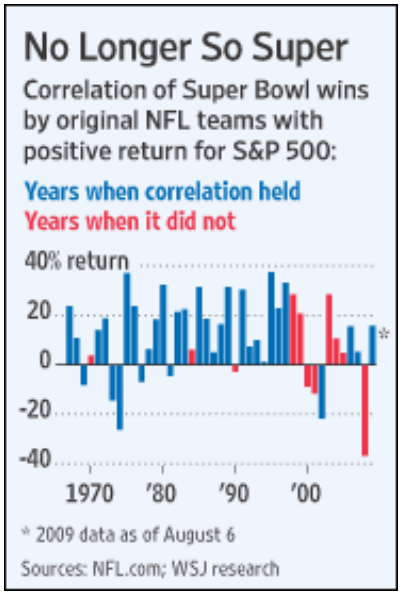
Data Mining Process

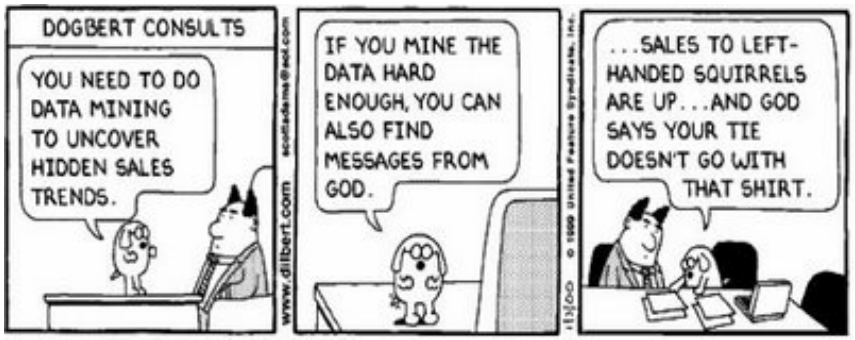



1. IDENTIFY, within your business processes, those that would benefit the most from data mining and predictive analytics
2. ANALYZE the data and develop predictive models
3. DEPLOY the models
4. QUANTIFY the benefits
THE DATA MINING VIRTUOUS CIRCLE

Domain Understanding Stage

- Learning the business goals
- Gathering relevant prior knowledge
- Best executed by a team of business and IT persons
- Good understanding of the domain avoids discovering irrelevant patterns or minimizes the chances for garbage in garbage out





Butter in Bangladesh Predicts the Stock Market

By Selena Maranjian | More Articles
September 20, 2007 | Comments (0)

Example Data Sources

- Point-of-sale data
- Credit card charge records
- Warranty claims
- Medical insurance claims
- Direct mail response data

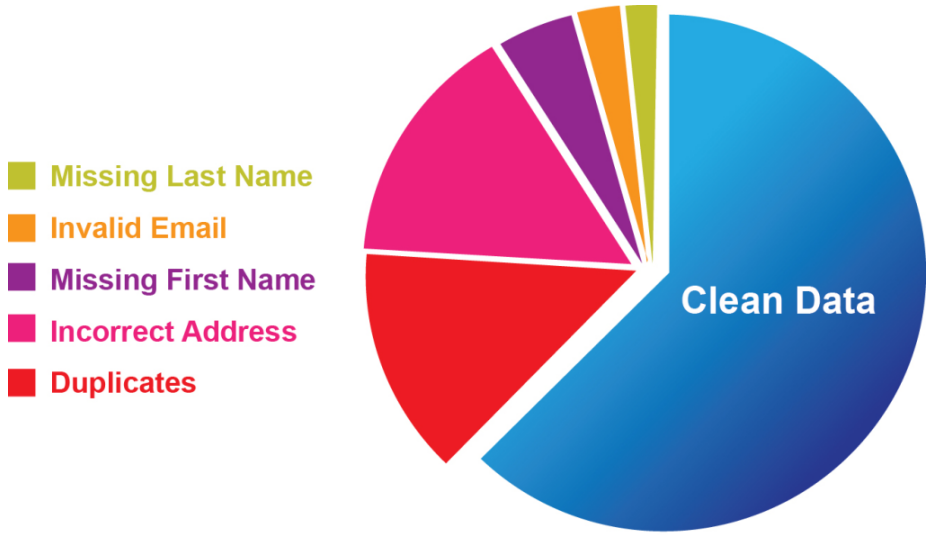
- Telephone call records
- Web activity data
- Economic data
- Utility charges
- Census returns
- Magazine subscription

Data Cleaning and Preprocessing Stage

- Data comes from many sources - internal and external
- Data comes in many forms and formats
  - Hierarchical databases, flat files, COBOL data sets
- Data is never clean
- Most important stage. Typically consumes about 60-80% of the total data mining effort

Business Data Corruption Examples

- Duplication - A common problem with direct mailers and credit card companies
- Missing and Confusing Data Fields
- Outliers - Generally present due to incorrect entry/coding of a data field.



Study above indicates that up to 30% of your CRM data could be wrong.

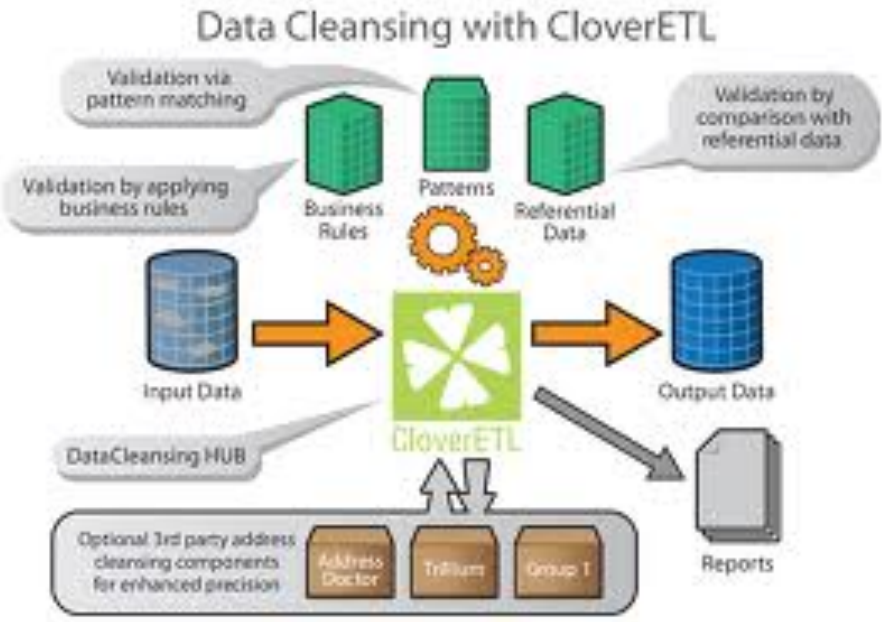
Data Preprocessing

This step is also known as *data transformation*. The aim here is to map data fields into representations suitable for the data discovery stage.

Examples:

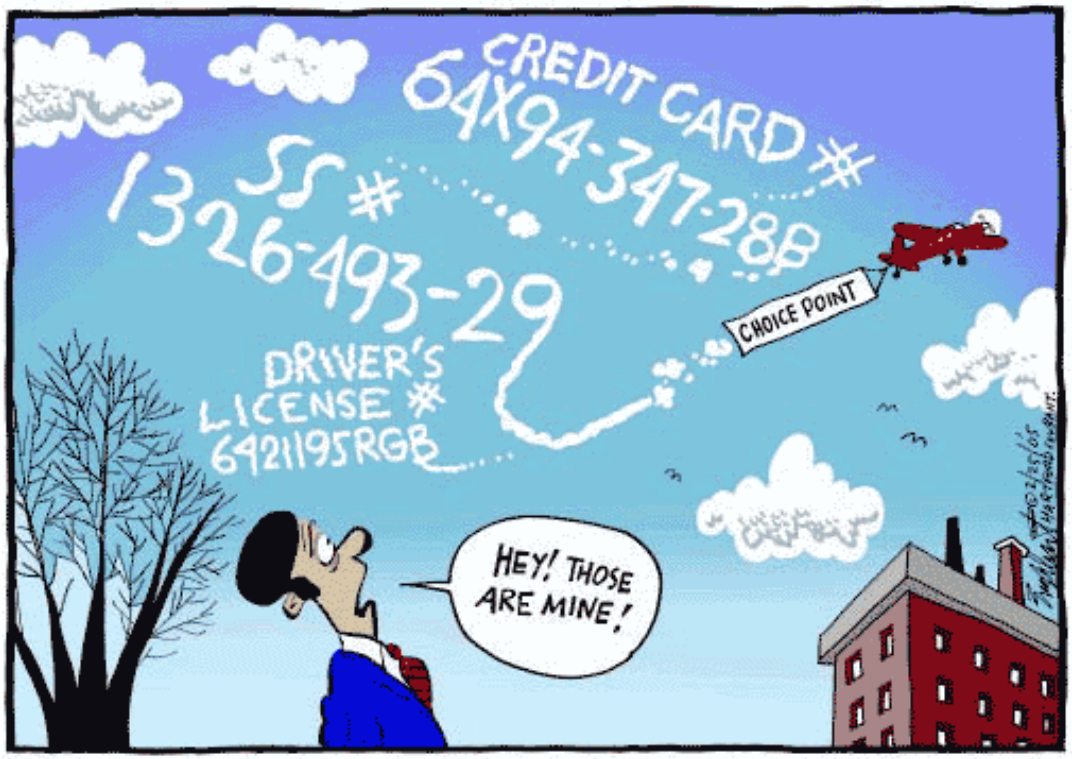
Month/Date/Year ===> Age Groups

Customer Address ===> Geographic Zone Code



Pattern Discovery Stage

- Discovery Model?
- Discovery Methodology?



Discovery Models

- Association Model
- Classification Model
- Clustering Model
- Regression Model
- Sequential Model
- Visual Model

"The computer predicted that I would have a sandwich for lunch. I ate cake!"

Association Model

90% of customers who subscribe to at least three premium channels also subscribe to pay-per view events

Also known as Market Basket Analysis

Association Model: Application

- Supermarket shelf management.

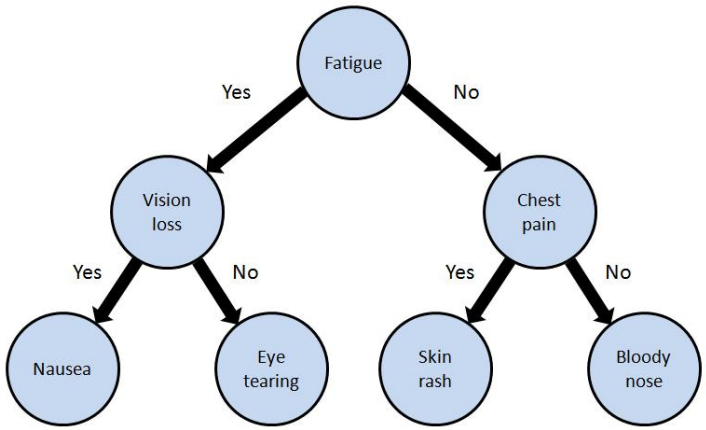
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Classification Model

- If *Annual_Income >40,000* **AND** *Home-Owner,* Then *Credit-Risk* ➔ *Medium*

- If $\frac{(Annual\_Income)^{1.2}}{(Avg\_Monthly\_CreditCardBalance + Mortgage)^{1.5}}$ *25,*

  Then *Loan-Approval* ➔ *Yes*



"I'd like to diversify my portfolio. For a change, why don't you get me a stock that's not in a group I like to classify as 'losers?'"

Classification: Application

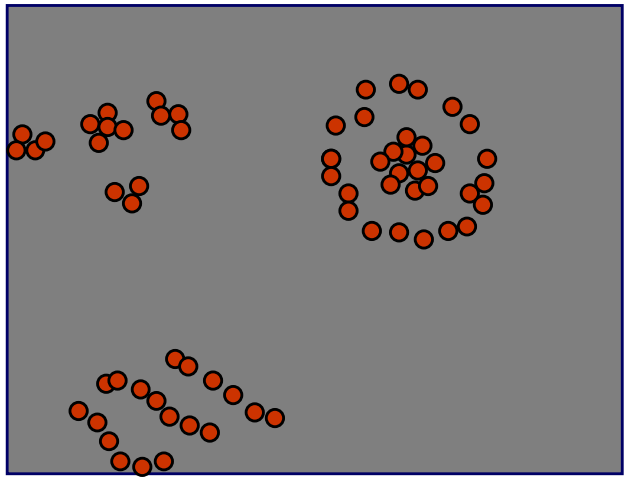
- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification: Application

- 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.

Clustering Model

Clustering models are similar to classification models except that no a-priori information is available for classes.

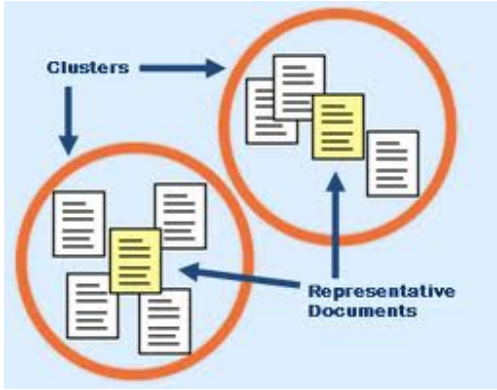


Clustering: Application

- 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.

Clustering: Application

- 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.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.



Regression Model

*Log(Peak_Load) = 300 + 1.6(|Temp - 65|)**2*

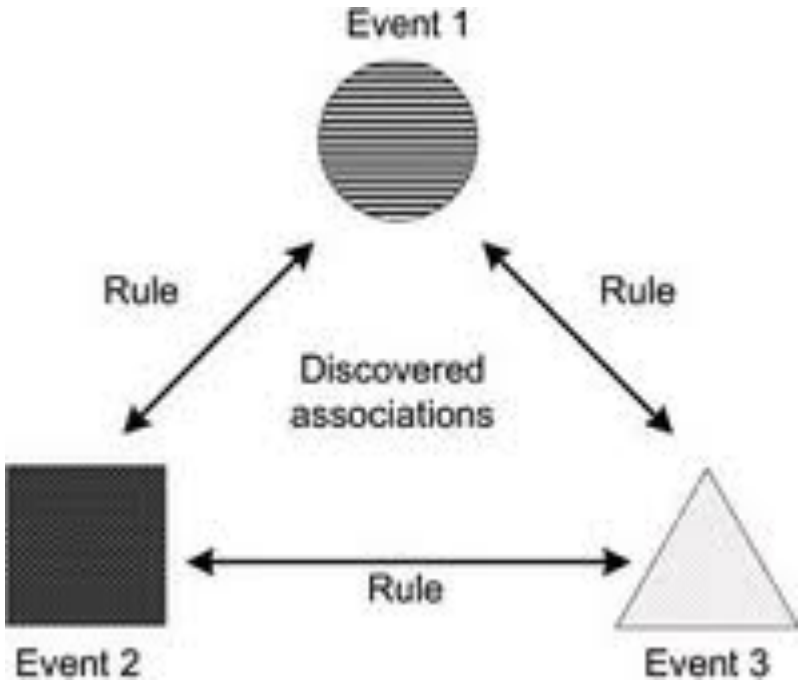
+ 2.4(Rel_Humidity - 80)

Unlike classification models that produce only discrete outcomes, a regression model generates a numerical score as its output.

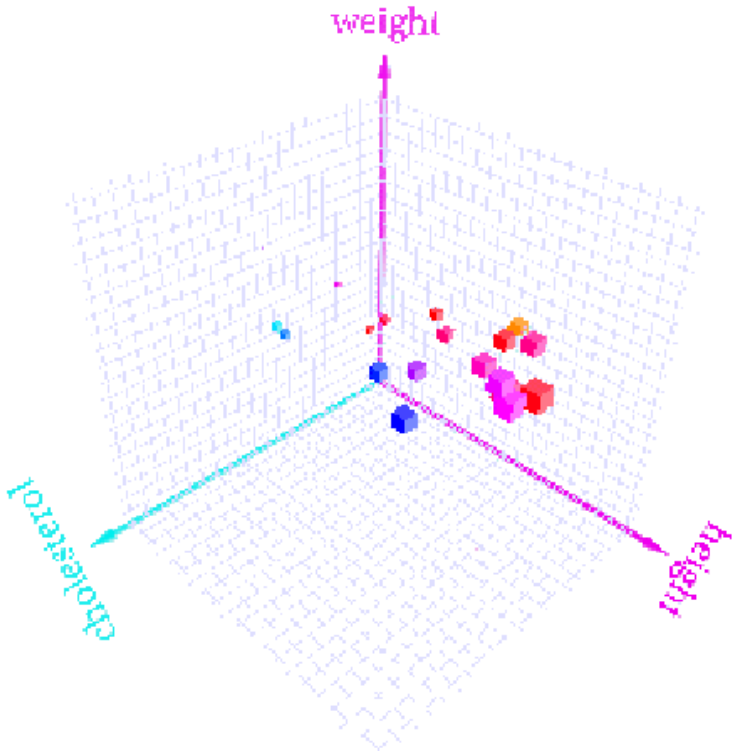
Sequential Model

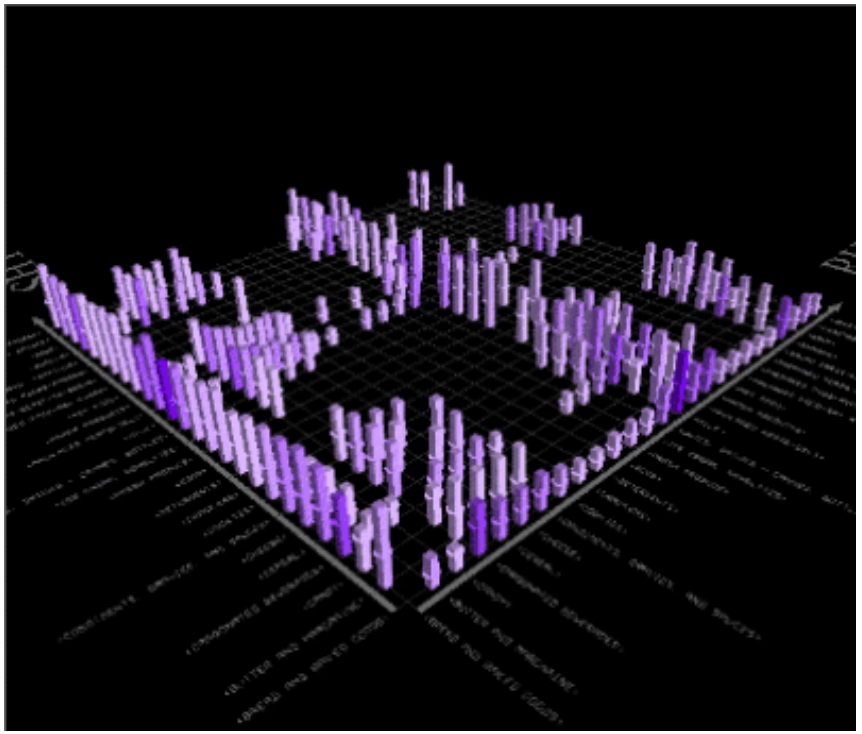
Similar to association models except that sequences of events are considered. For example:

“80% of customers who buy a product X are likely to buy product Y in next six months”



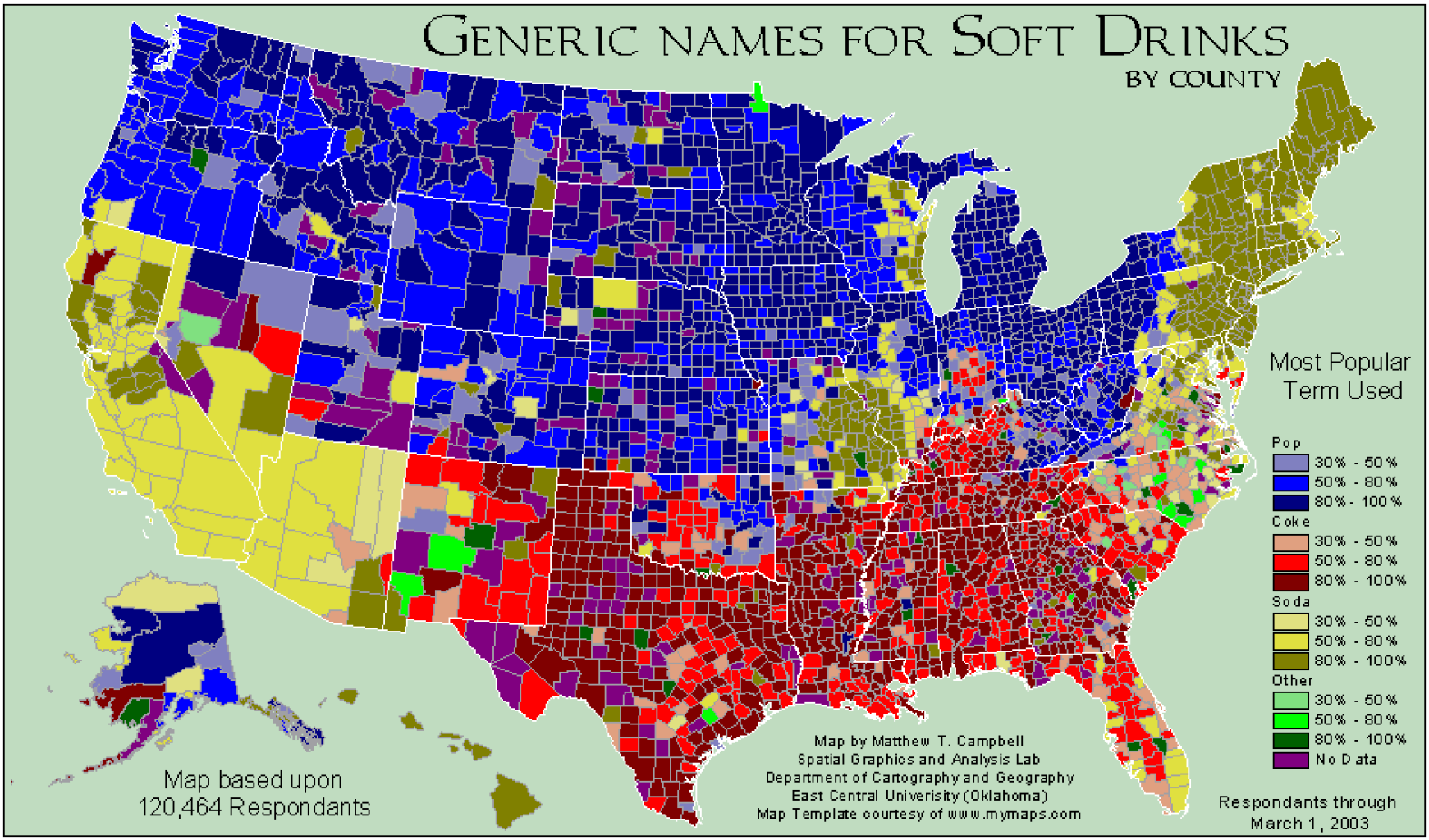
Visual Model





Geographical Patterns and

Map Visualization

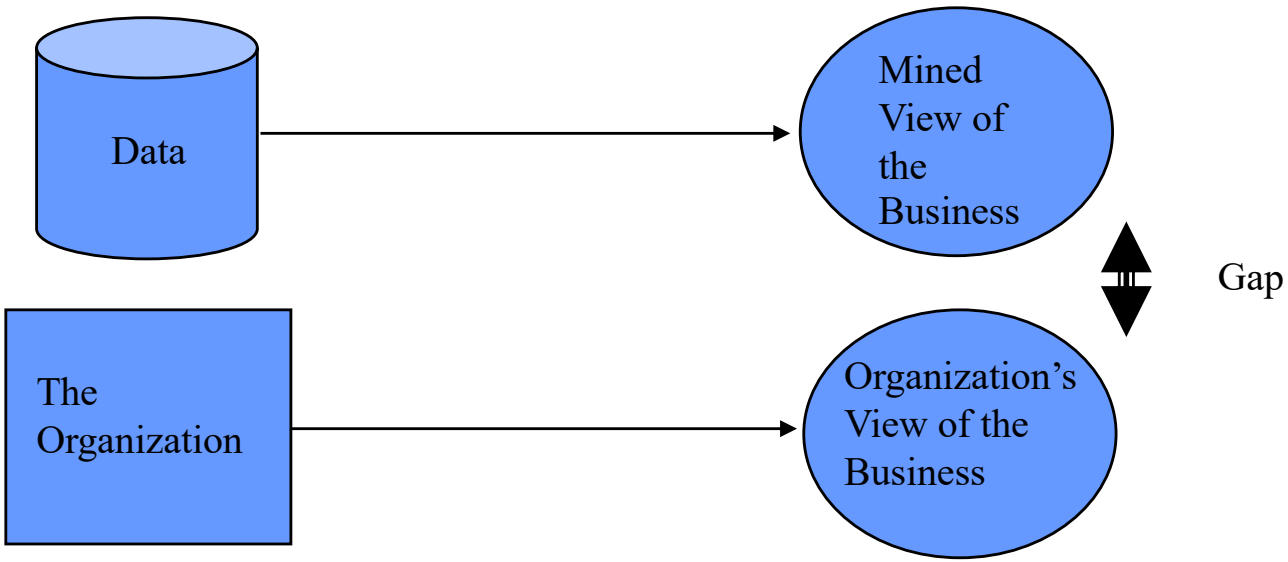


Interpretation Stage

- Evaluate the quality of the discovered patterns
- Determine the value of the discovery to the business

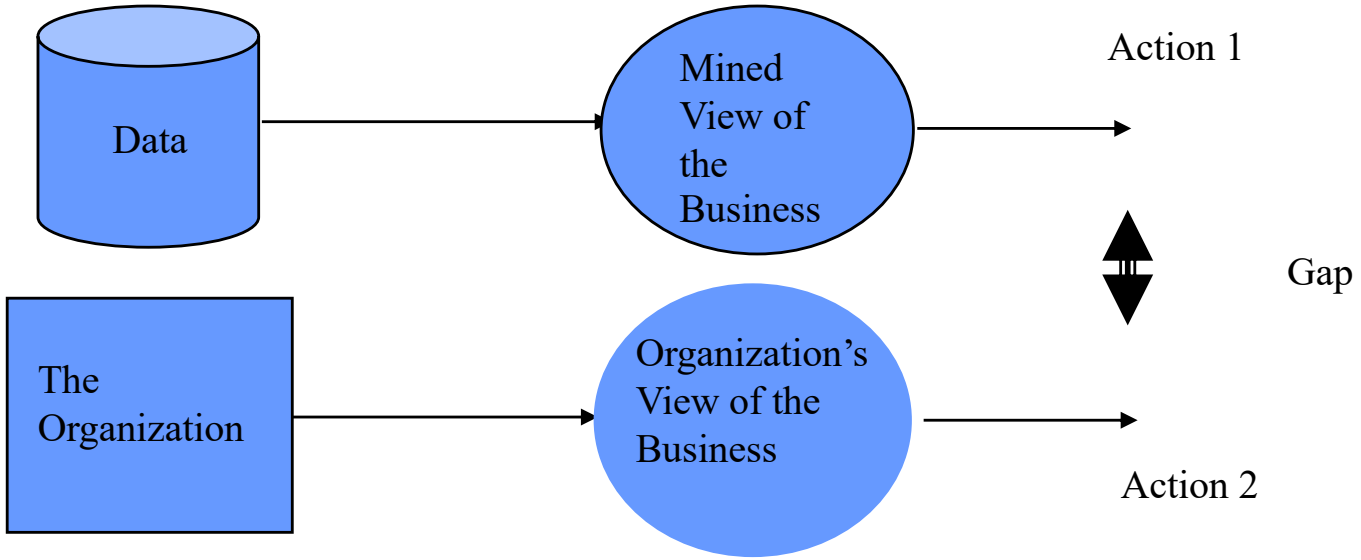
Value of Mined Information

- Perceptive gap



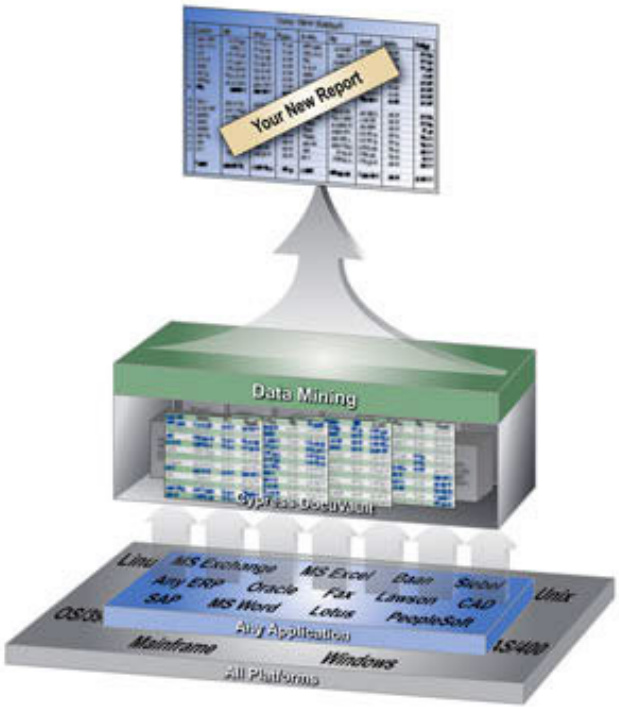
Value of Mined Information

- Dollar gap



Reporting Stage

- Reporting the discovery to higher management
- Transforming the discovery to new actions or products



Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation

Hey, wanna buy some data? Only slightly used! Heap of clicks left in it. Its last owner was a little old lady who only used it for shopping ...