

# DM: Intro

# What is Data?

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



Elida Rd.  
Lima, OH - # 110  
(419) 331-6400 [meijer.com](http://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

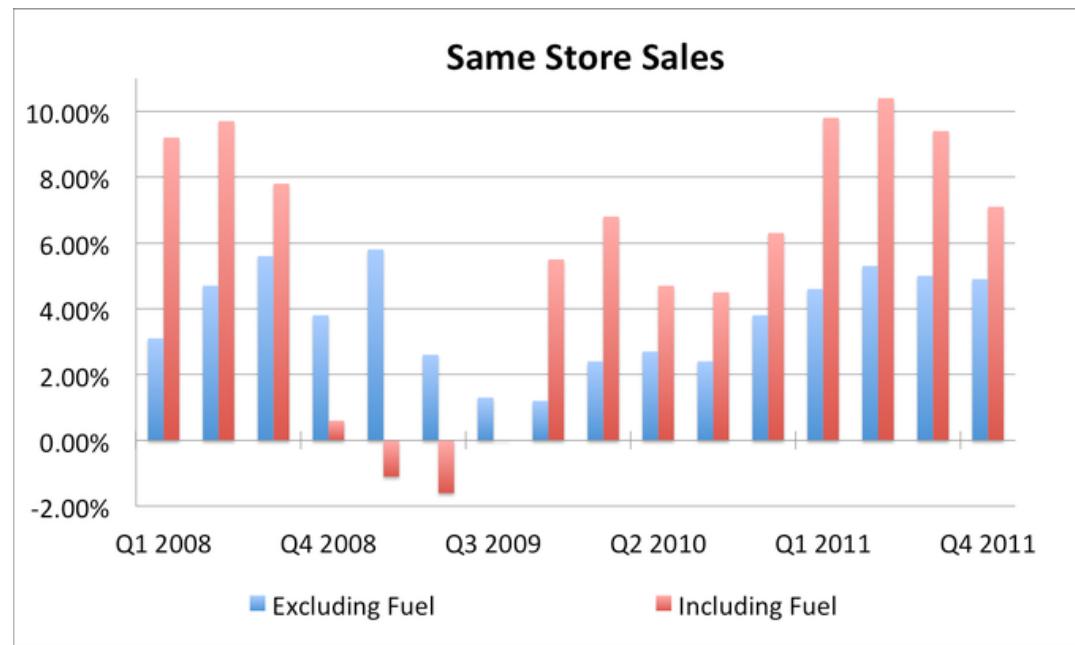
#### COUPONS

Vendor Coupon	-.50	F
EXTRA COUPON	-.50	F
81101003760008 Vendor Coupon	-.50	F
81101003760008 EXTRA COUPON	-.49	F
81101001901403 Vendor Coupon	-.100	N
81101001901403 Vendor Coupon	-.100	N
81101002670002 Vendor Coupon	-.50R	F

Moerks # -- \*\*\*\*\*63  
Meijer Card -- 00900675  
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

## Frequently Bought Together

LOOK INSIDE! + LOOK INSIDE! + LOOK INSIDE!

**Price For All Three: \$31.25**

Add all three to Cart    Add all three to Wish List

[Show availability and shipping details](#)

This item: **The Inheritance of Loss** by Kiran Desai Paperback \$10.17

**The White Tiger: A Novel** by Aravind Adiga Paperback \$10.20

**The God of Small Things: A Novel** by Arundhati Roy Paperback \$10.88

## Customers Who Bought This Item Also Bought



**The God of Small Things:**  
A Novel  
► Arundhati Roy  
 (943)  
Paperback  
**\$10.88**



**The White Tiger: A Novel**  
► Aravind Adiga  
 (410)  
Paperback  
**\$10.20**



**Interpreter of Maladies**  
► Jhumpa Lahiri  
 (525)  
Paperback  
**\$10.17**



**The Killing Zone: The United States Wages ...**  
► Stephen G. Rabe  
 (1)  
Paperback  
**\$17.16**



**History of How the Spaniards Arrived in ...**  
► Diego De Castro Titu Cusi...  
 (2)  
Paperback  
**\$18.00**

# 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.



Copyright © 2000 United Feature Syndicate, Inc.  
Redistribution in whole or in part prohibited

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?

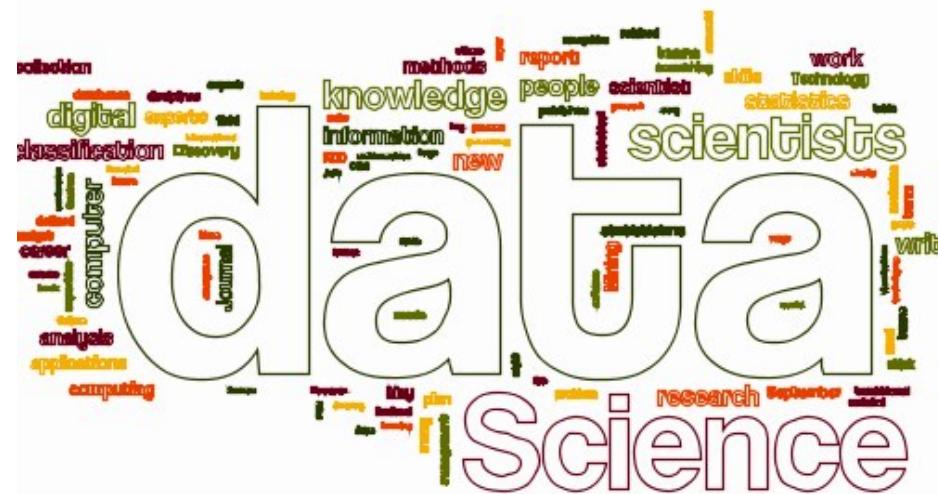
Posted June 21, 2012 to Business Intelligence | 1 Comment



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

Owns a laptop

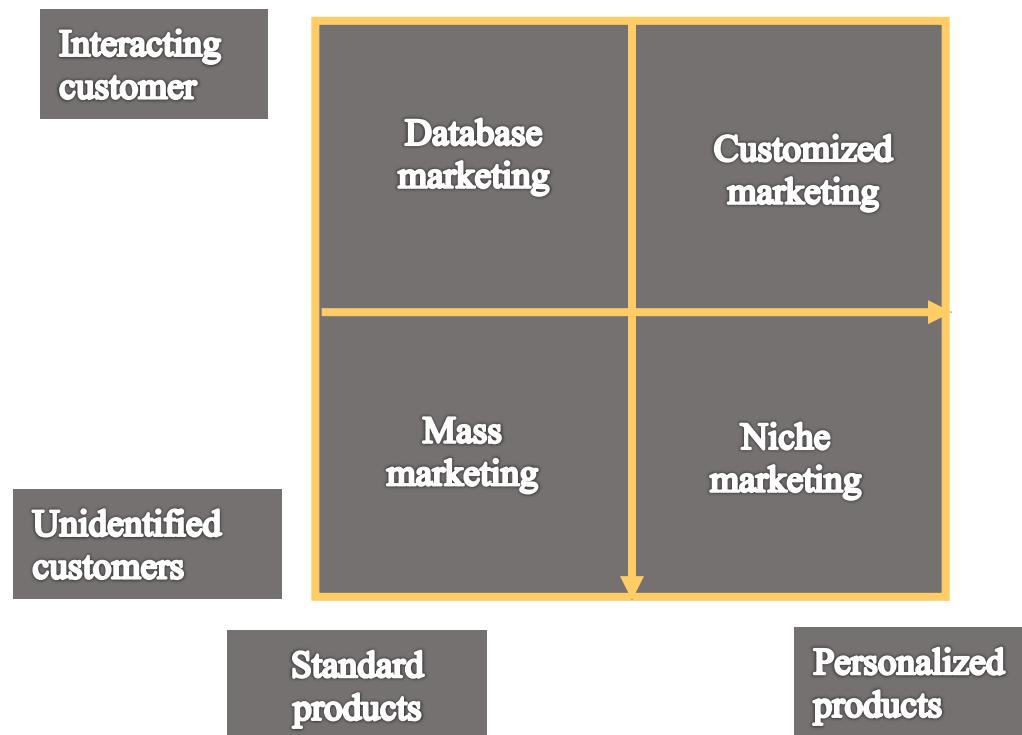
Household income: \$100,000+  
Age: 38-39 Likes: fashion Major life-insurance holder Age: 36-45  
Likes: online news Likes: wife/wife's hobbies: cooking & recipes Likes: no kids  
Lives in Los Angeles Likes: Asian cuisine Likes: online shopping Lives in New York City  
Likes: green living Likes: cars Likes: celebrities & actresses Likes: textile designer  
Purchased house six years ago Likes: food Likes: fashion Likes: travel at same address for four years  
Favorite celebrities: Penelope Cruz, Robert De Niro ZIP code: 10701 Profession: teacher  
ZIP code: 10701 Property: 1,250 sq. ft. Likes: Robert Goulet Likes: home & garden  
Wi-fi warrior Age: 35-39 Likes: business & finance Likes: Politically active Likes: music  
Sister is a lawyer Likes: parenting Likes: RV Likes: hockey  
Frequent purchaser: apparel Likes: Spent \$180 on intimate app. & undies on Oct. 10, 2010 Likes: hiking Household income: \$150,000-\$175,000  
Recently traveled to House Likes: medical professional Likes: male Mother: Rosalind Band Likes: hiking Household income: \$150,000-\$175,000  
Likes: parenting Likes: No handles Likes: smart phone Likes: music Likes: retail Likes: music  
Likes: coffee & tea Likes: Smart-phone user Likes: movies Magazine subscriber Likes: finance Likes: newspapers  
Has had LASIK surgery Likes: restaurants Likes: less magazines Likes: travel & hotel Likes: retail Likes: music  
Likes: less discount Likes: coffee & tea Likes: Has used cocaine Likes: small-business owner Likes: news  
Likes: chocolate Likes: TV subscriber Likes: more discounts Likes: family

## YOUR DATA FOR SALE

Everything about you  
is being tracked—  
get over it  
BY JOEL STEIN

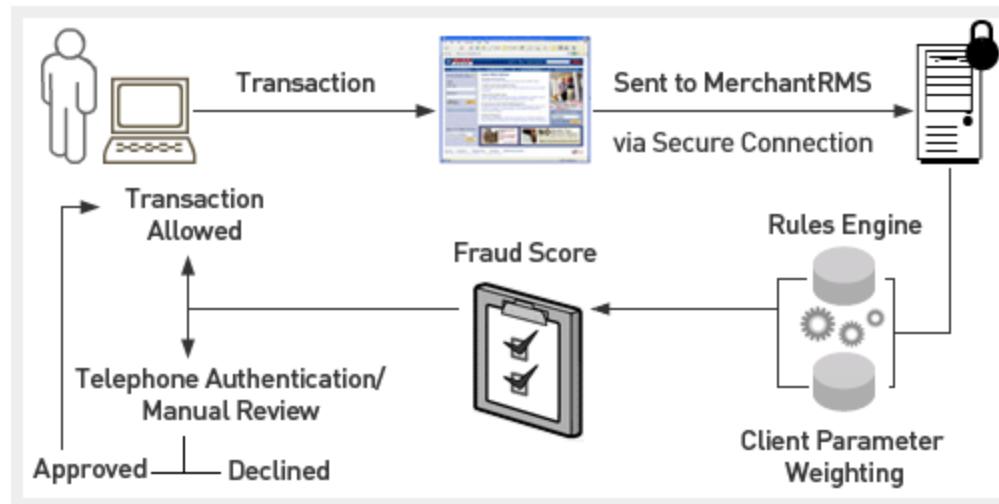
What data-mining  
companies think  
they know about  
Joel Stein

# 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



PREMIER REFERENCE SOURCE

## Clinical Data Mining for Physician Decision Making and Investigating Health Outcomes

Methods for Prediction and Analysis



PATRICIA CERRITO & JOHN CERRITO

# 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	9W. 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	3500Tomahawk	5	6	222	785934444

CUSTOMER FILE				EMPLOYEE FILE			
Customer Number	Customer Name	Customer Phone	Customer Primary Contact	Employee ID	Employee Last Name	Employee First Name	Date of Hire
1234	Smelding Homes	3333333333	Bill Johnson	123456789	Johnson	Emilio	2/1/1985
2345	Home Builders Superior	3334444444	Marcus Connolly	435296657	Evaraz	Antonio	3/3/1992
3456	Mark Akey	3335555555	Mark Akey	785934444	Robertson	John	6/1/1999
4567	Triple A Homes	3336666666	Janielle Smith	984568756	Smithson	Allison	4/1/1997
5678	Sheryl Williamson	3337777777	Sheryl Williamson				
6789	Home Makers	3338888888	John Yu				

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

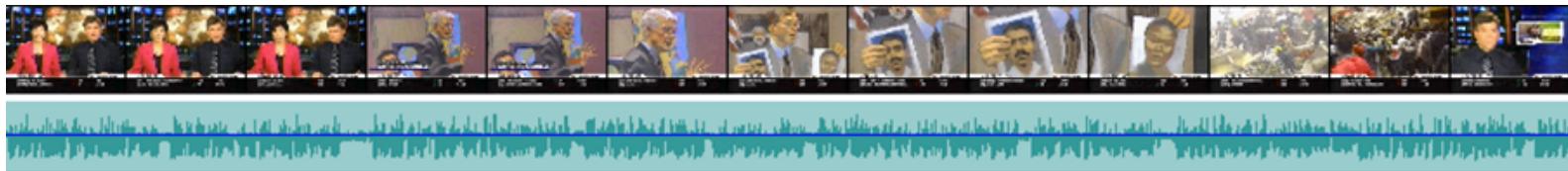
20 new Tweets

 **Strata Conference** @strataconf 6h  
Got dirty **data**? Find the right tool for the job [oreil.ly/RfbV9k](http://oreil.ly/RfbV9k) #strataconf tutorial w/ speakers from @thenyworld #datajournalism  
↗ Promoted by Strata Conference  
Expand

 **Kade ellis** @kade\_ellis 9h  
You guys there is actually something called Patriarch and it's a **data** management and corporate **data mining** client for the FBI.  
#nodads  
Expand

 **Mark Ginnebaugh** @markginnebaugh 1h  
Introduction to Microsoft **Data Mining** > Sept 6 in San Francisco  
[tmbir.co/Zyb2qvRx-aFy](http://tmbir.co/Zyb2qvRx-aFy)  
Expand

 **A Wiki for CFPs** @WikiCPP 2h  
[CFP] SDM 2013 : SIAM International Conference on **Data Mining**  
[bit.ly/SogMoC](http://bit.ly/SogMoC)  
Expand



# 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**

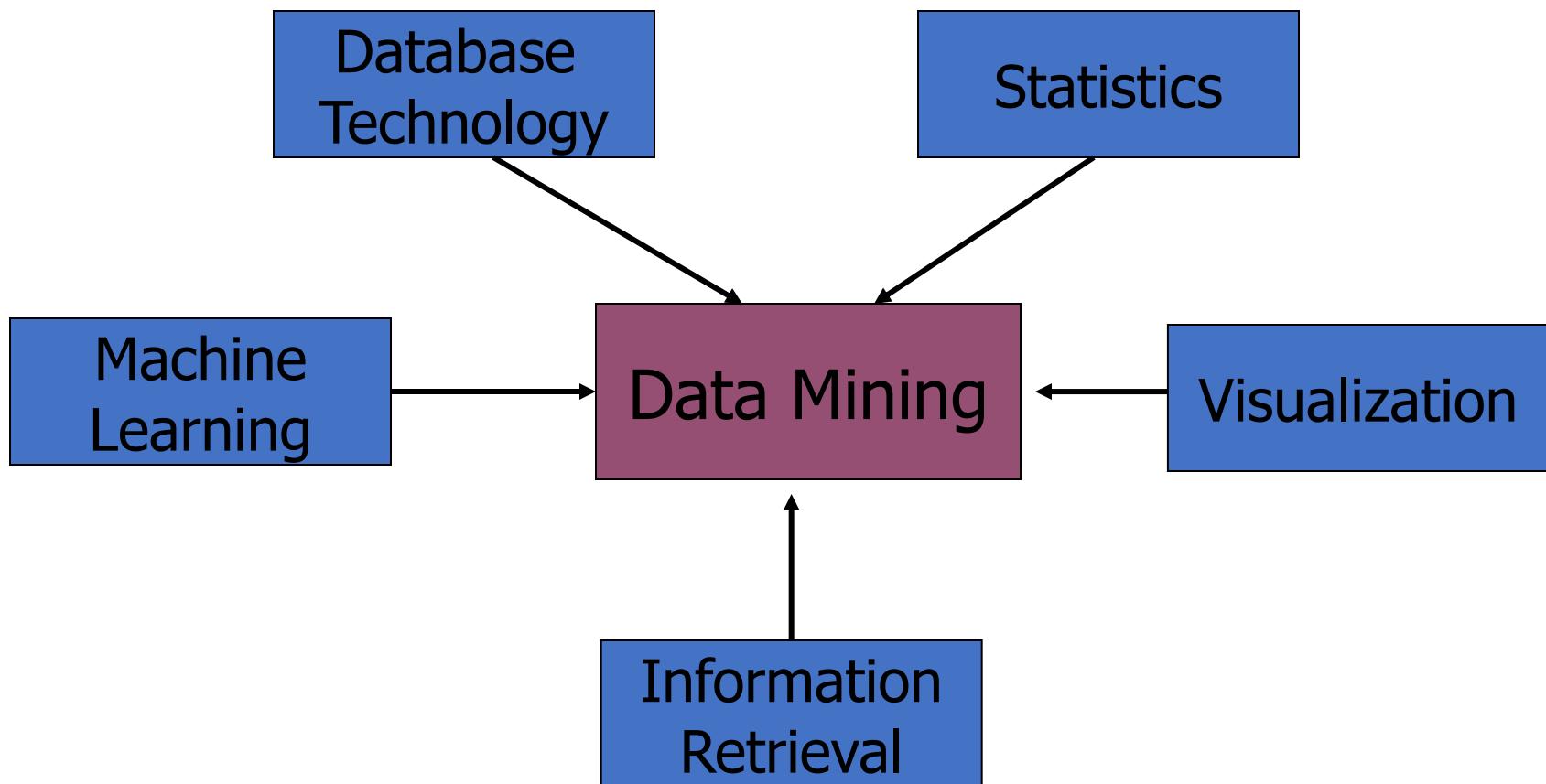
- o Excision, benign lesion, face, ears, eyelids, nose, lips, mucous membrane; lesion diameter 0.5 cm or less (11440)
- o Biopsy of skin (11100) - 03/19/2008

**Instructions**

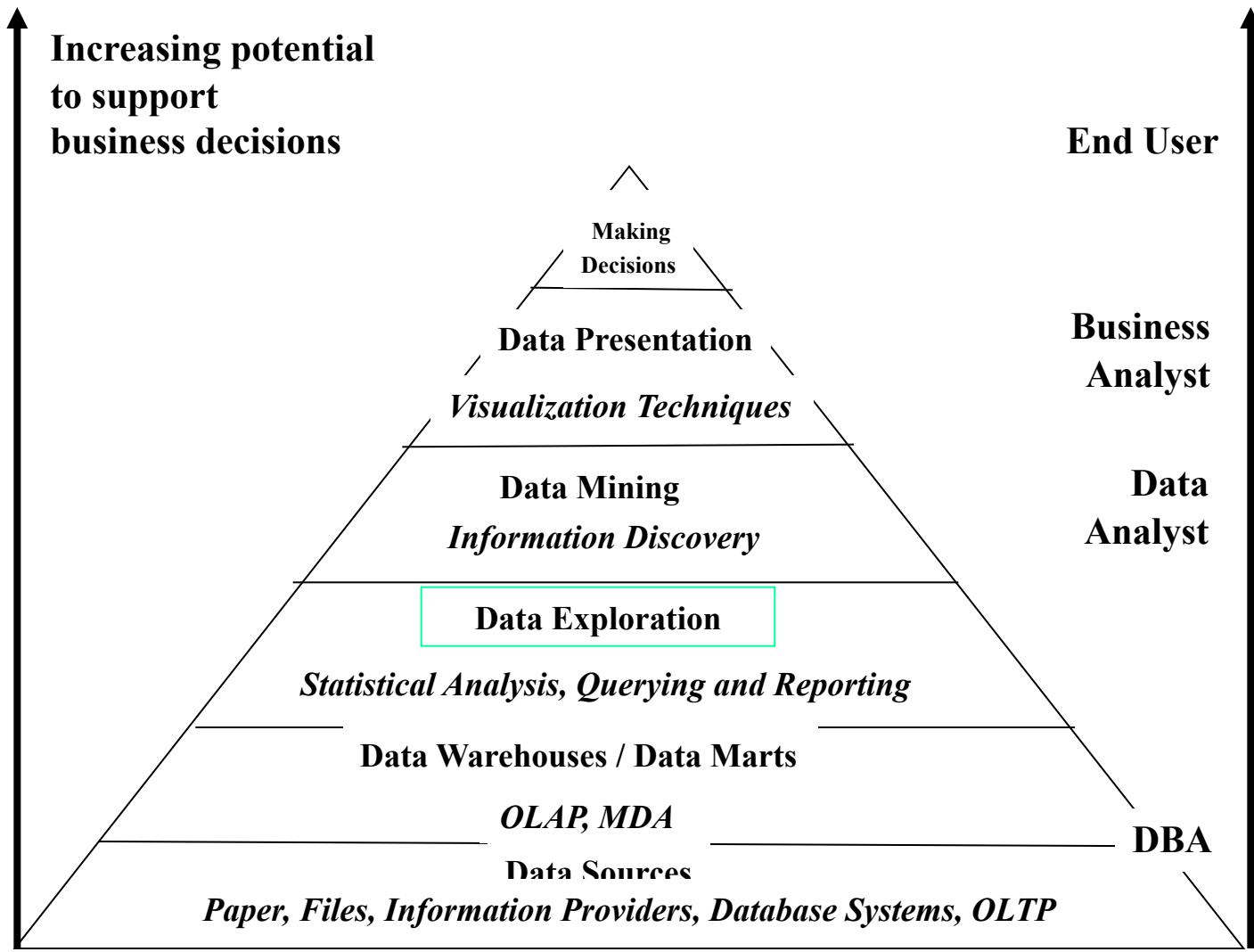
- o Pathology results will be phoned to patient.

The image contains two photographs. On the left, a close-up view of a dark, irregular skin lesion is shown next to a ruler marked in centimeters (cm), indicating its size. On the right, a profile view of a man's head and neck is shown, with a small black dot and an arrow pointing to the location of the lesion on the left side of his neck.

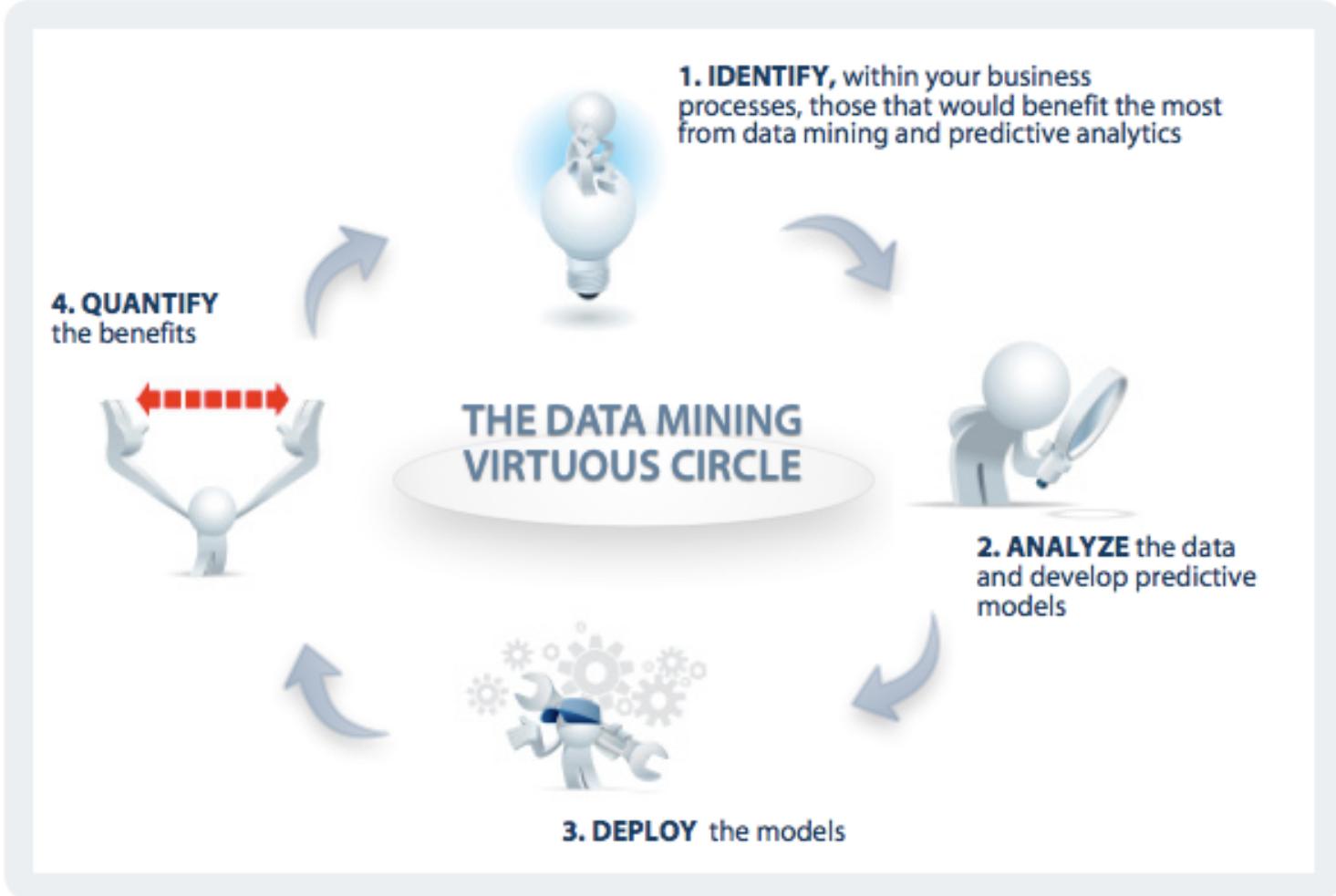
# Core Technologies for Data Mining



# Data Mining and Business Intelligence

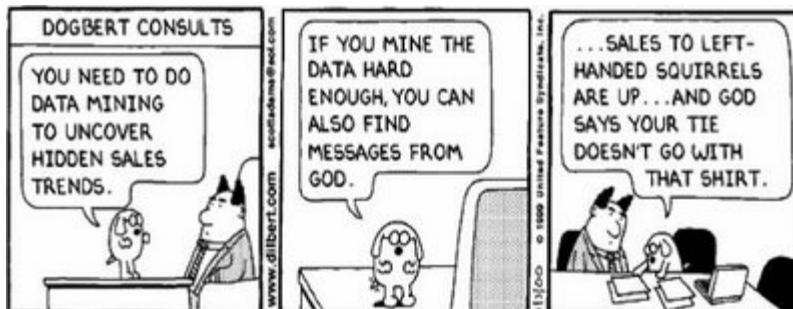
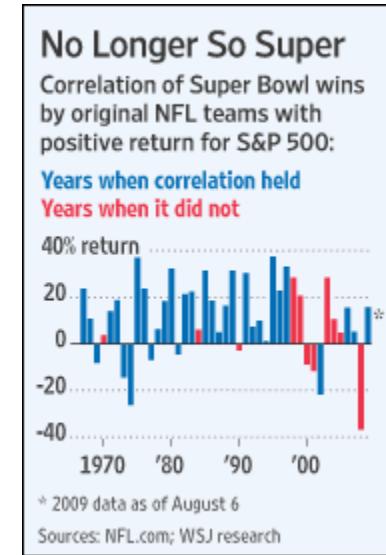


# Data Mining Process



# 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



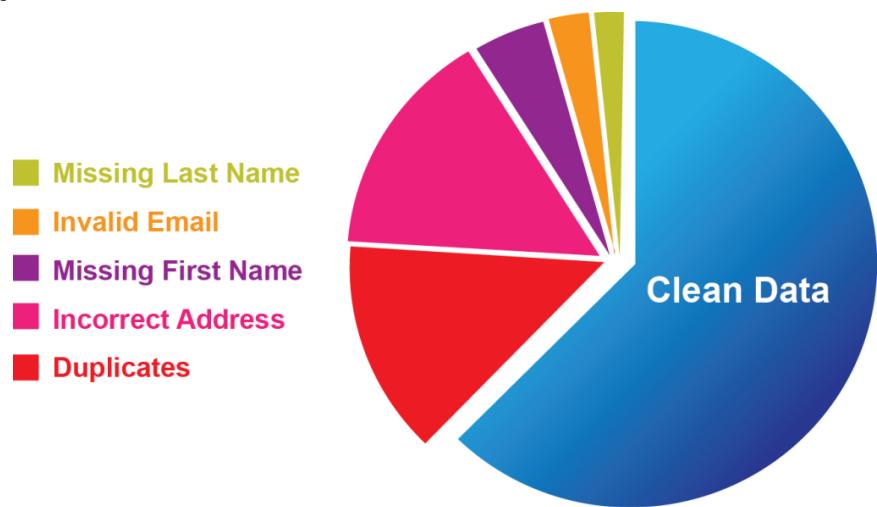
# Data Cleaning and Preprocessing Stage

- 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.

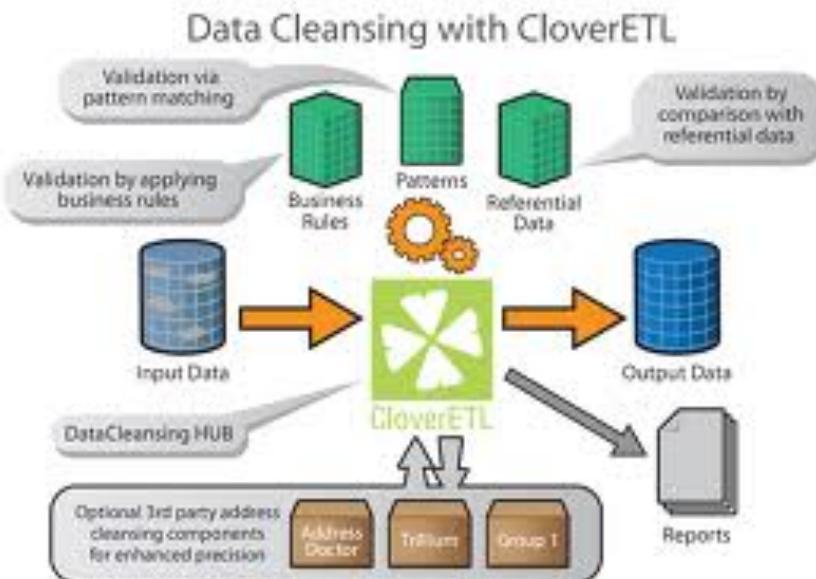
# Data Preprocessing

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:

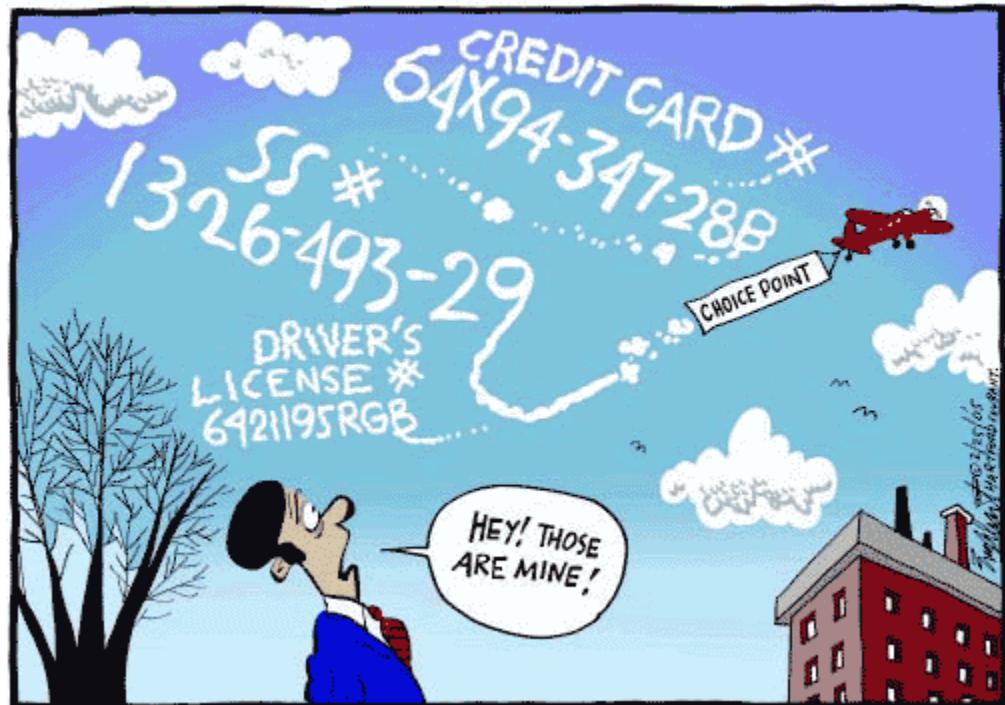
Month/Date/Year    ==> Age Groups

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.



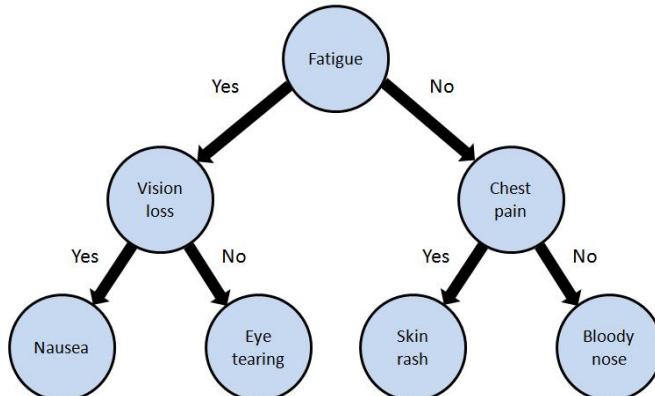
- Goal: To identify items that are bought together by sufficiently many customers.
- Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
- A classic rule --
  - If a customer buys diaper and milk, then he is very likely to buy beer.
  - So, don't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

# Classification Model

- If  $\text{Annual\_Income} > 40,000 \text{ AND Home-Owner}$ , Then  $\text{Credit-Risk} \rightarrow \text{Medium}$

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

Then  $\text{Loan-Approval} \rightarrow \text{Yes}$



*"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.



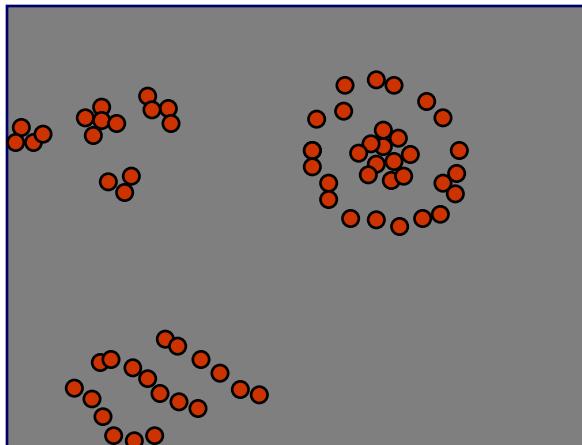
# Classification: Application

- 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

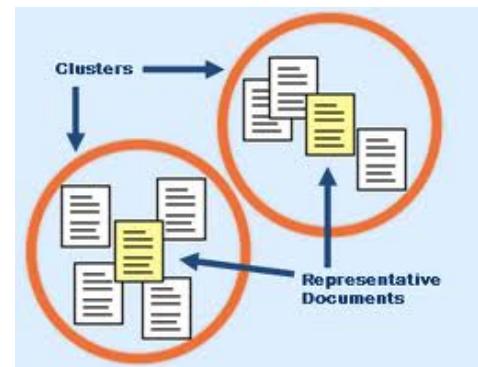
- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related 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

$$\begin{aligned} \text{Log}(Peak\_Load) = & 300 + 1.6(|Temp - 65|)^{**2} \\ & + 2.4(ReL\_Humidity - 80) \end{aligned}$$

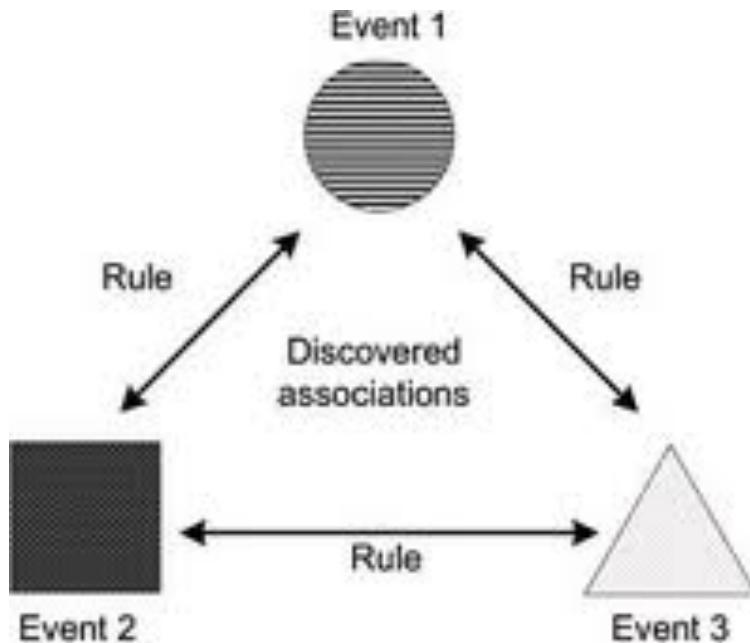
Unlike classification models that produce only discrete outcomes, a regression model generates a numerical score as its output.



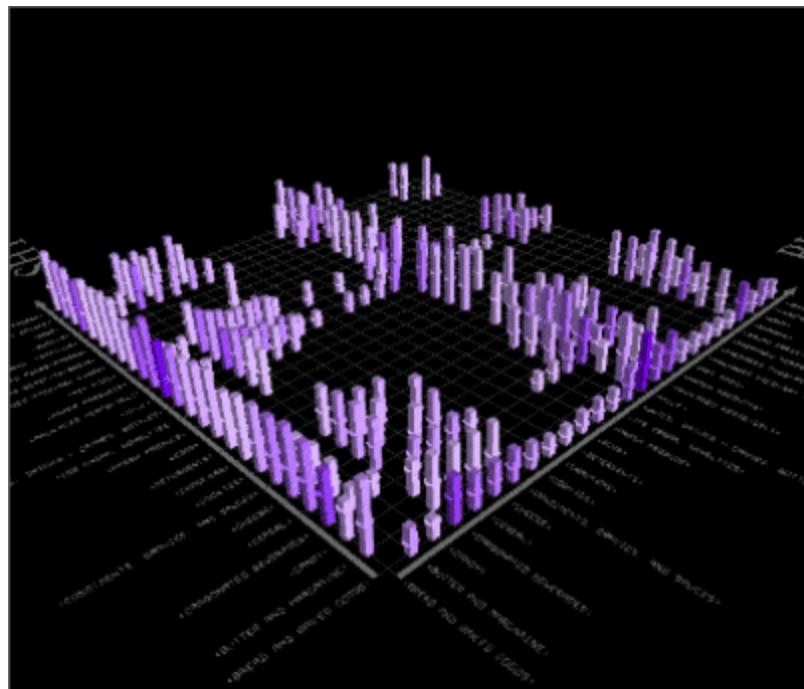
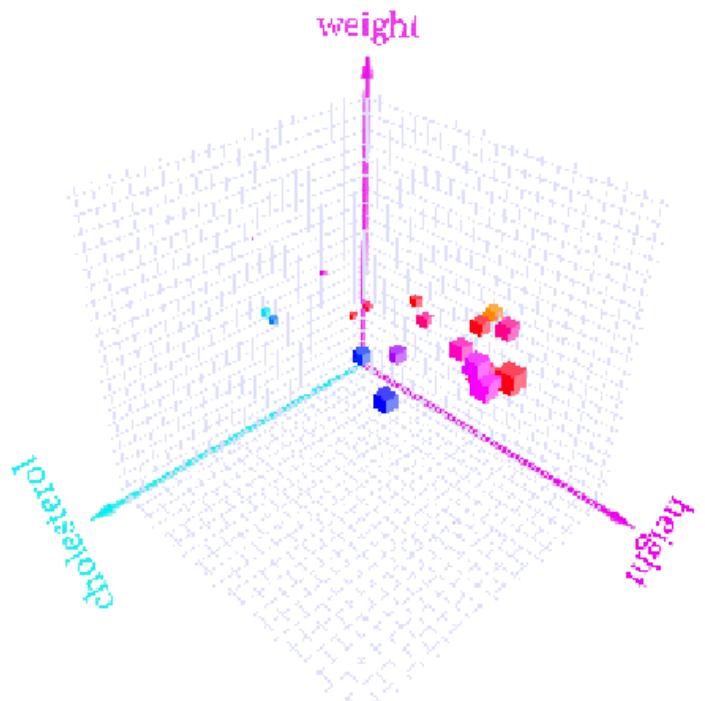
# Sequential Model

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

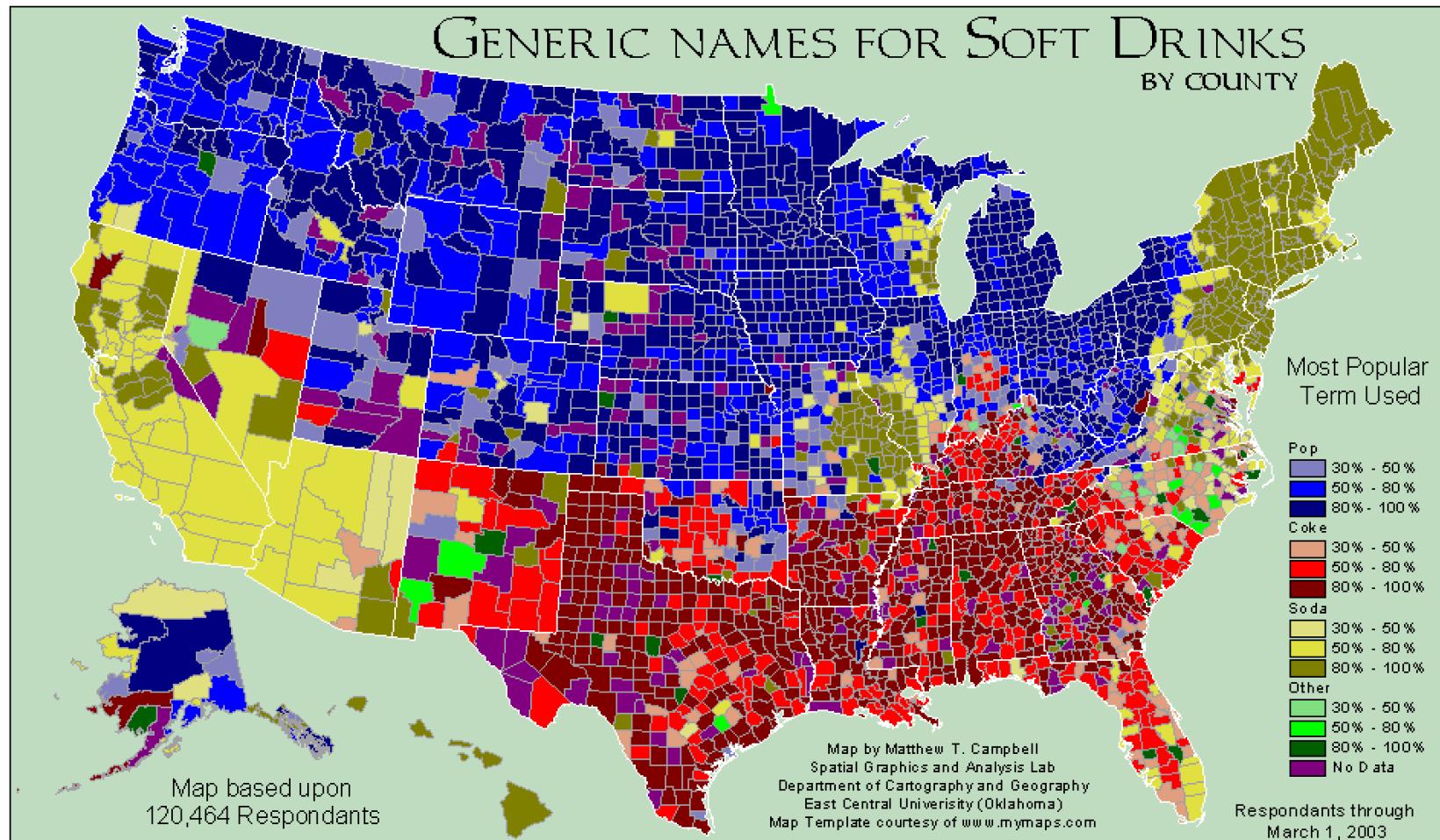
“80% of customers who buy a product X are likely to buy product Y in next six months”



# Visual Model



# Geographical Patterns and Map Visualization



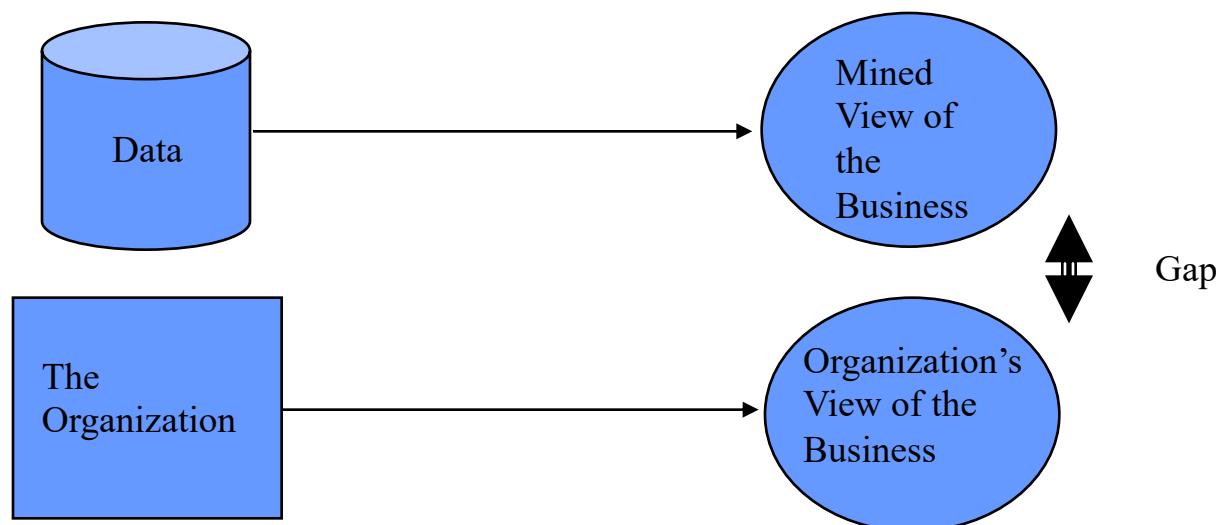
# Interpretation Stage

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



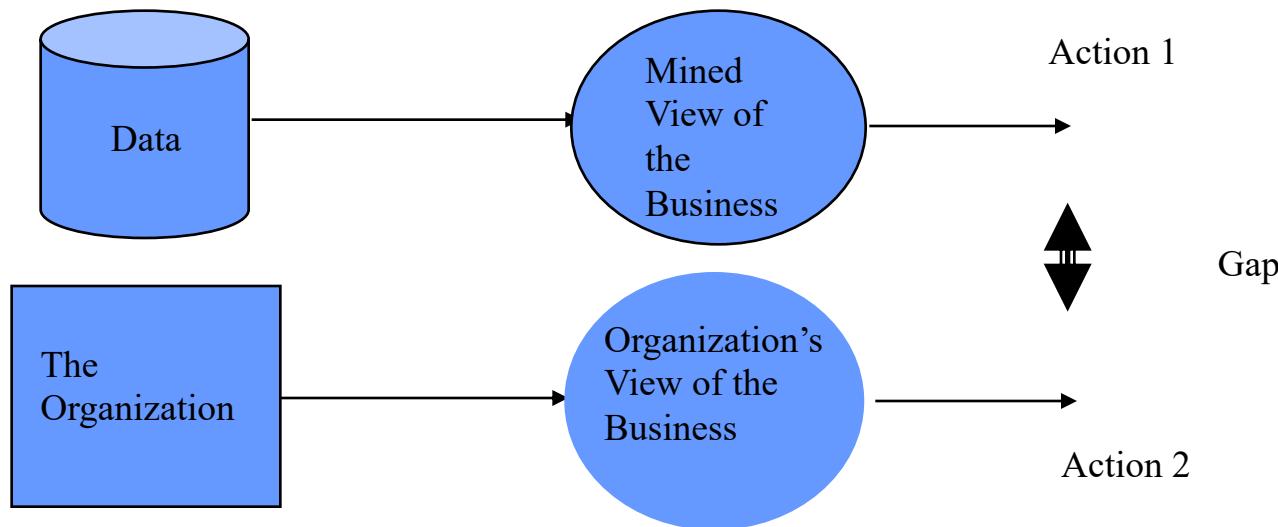
# Value of Mined Information

- Perceptive gap



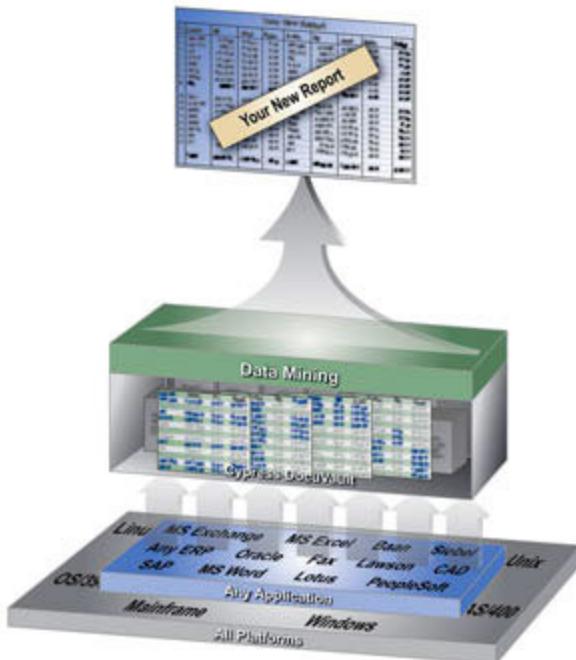
# Value of Mined Information

- 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 ...