

A man in a white shirt and tie stands in a field, looking towards the right. He is surrounded by a digital landscape of floating data points, binary code, and various charts like a globe and a bar chart. A large, glowing yellow ring is positioned in front of him.

# "We're Not in Kansas Anymore"

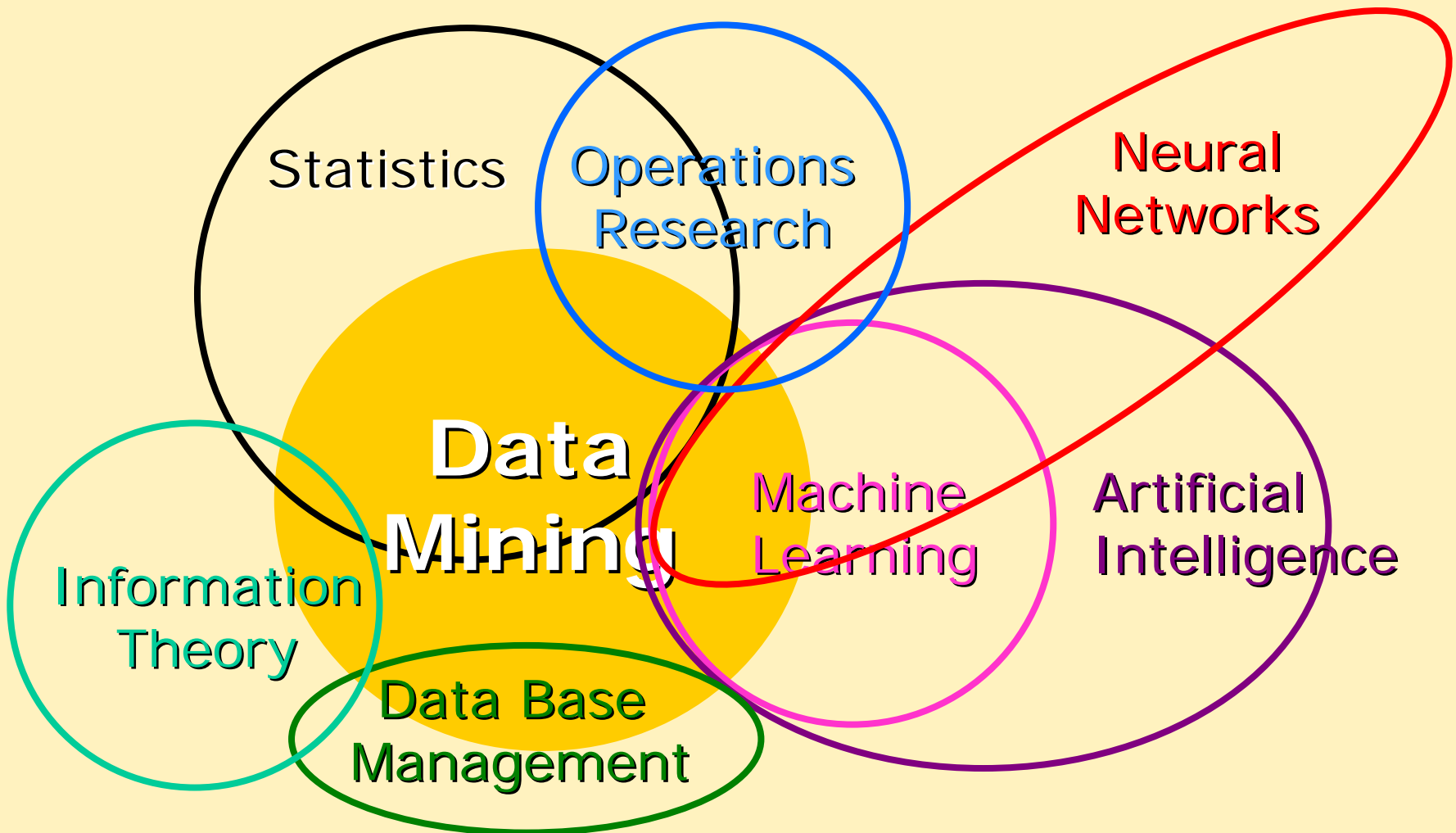
*Data Mining Users' Group*

*November 29, 2005*

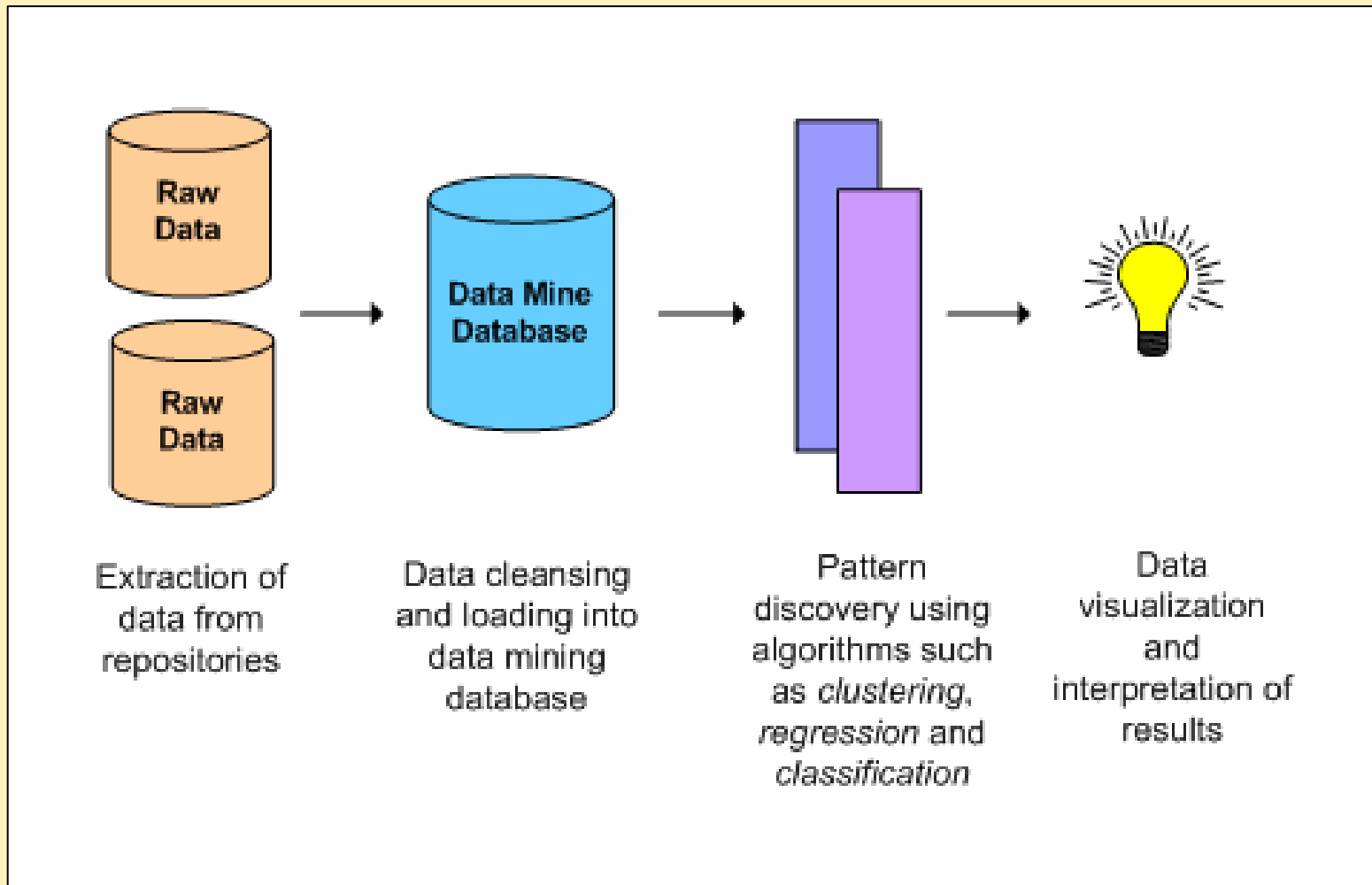
---

*David Yeo, Ph.D.*

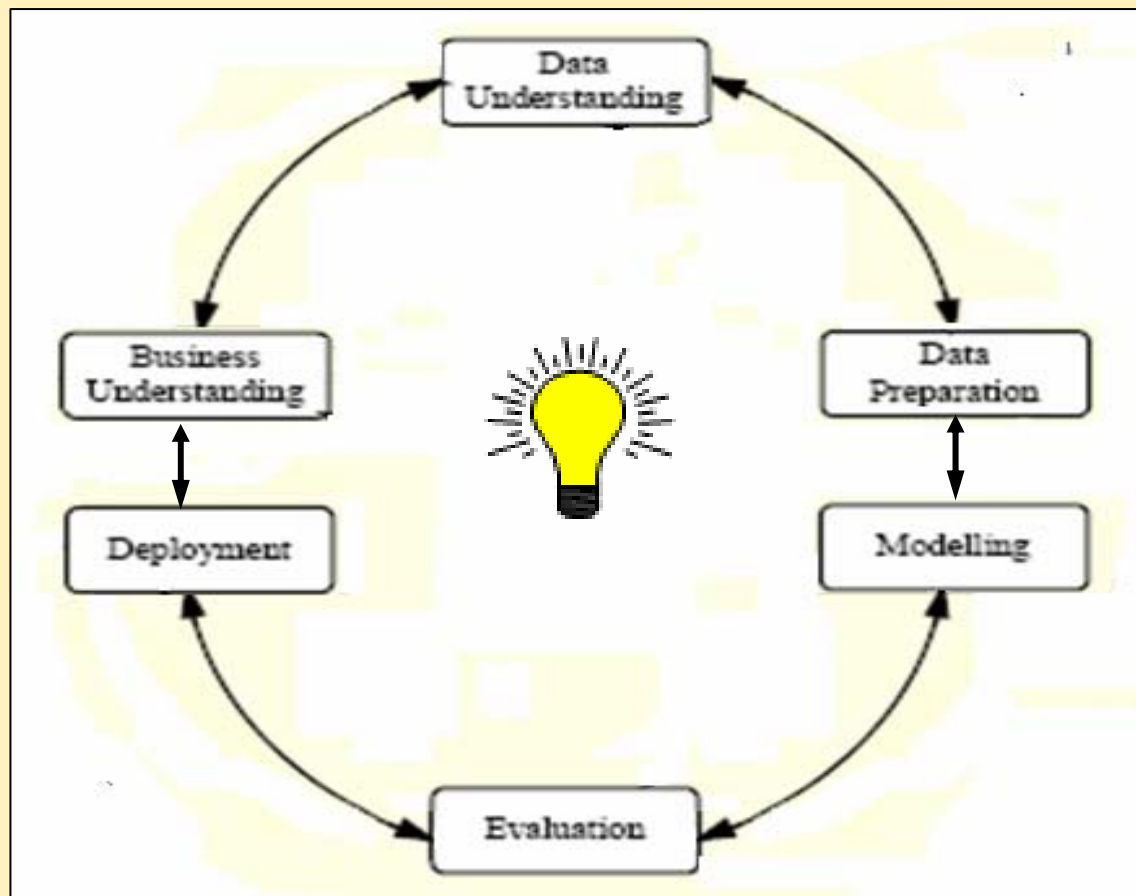
# Data Mining is NOT Just Statistics



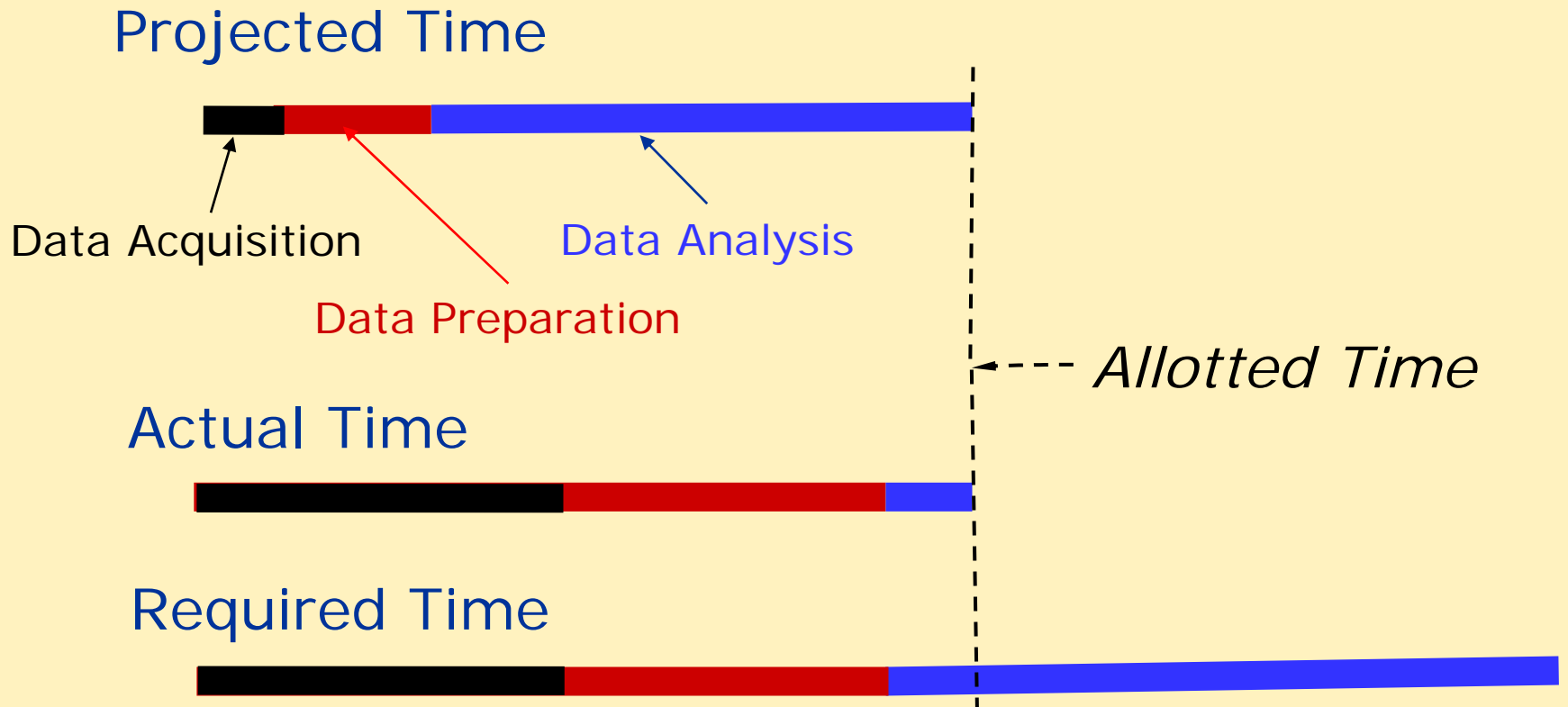
# Data Mining is a Process



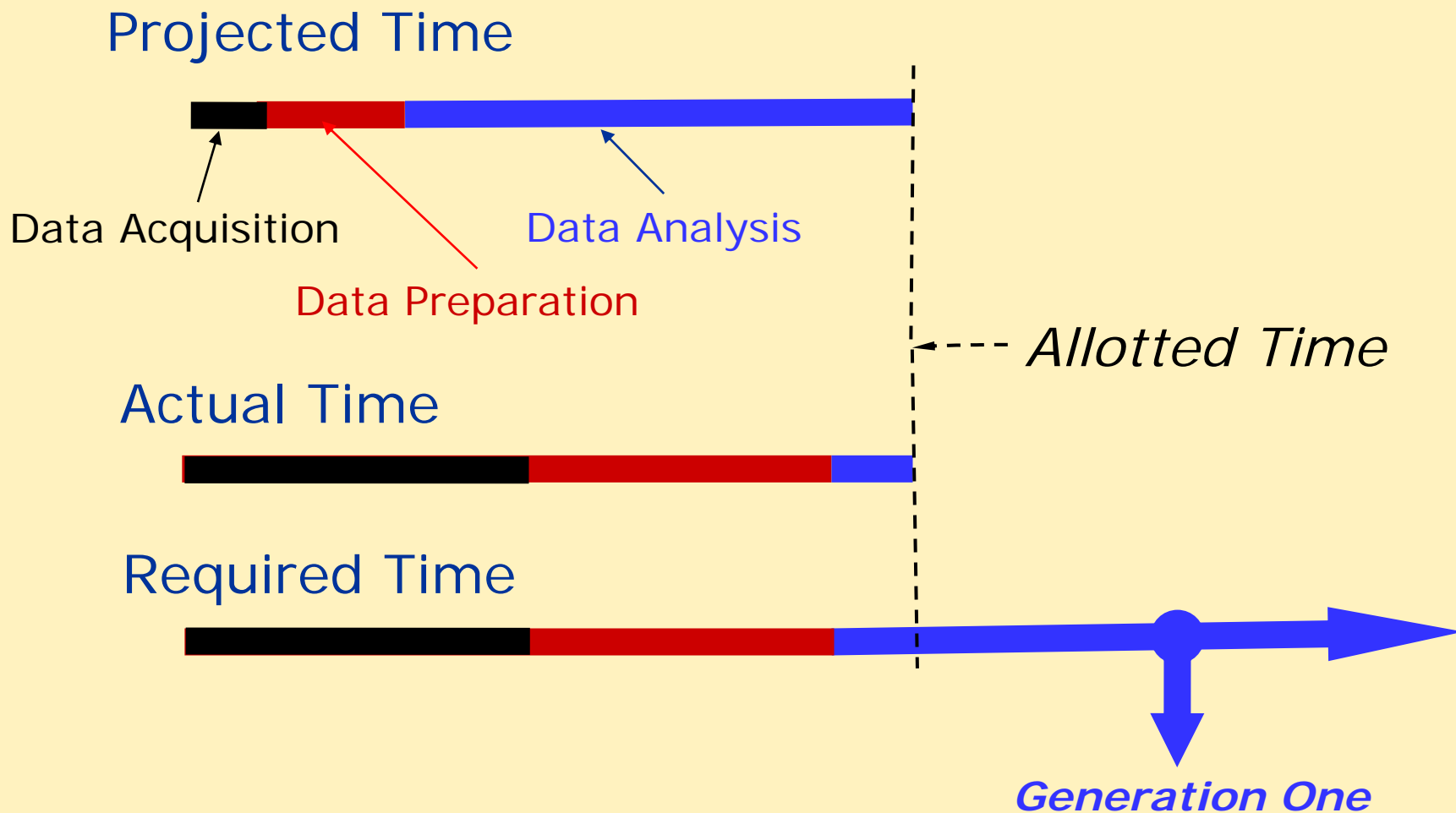
# Data Mining is an Iterative Process



# Managing the Data Mining Time Line



# Managing the Data Mining Time Line



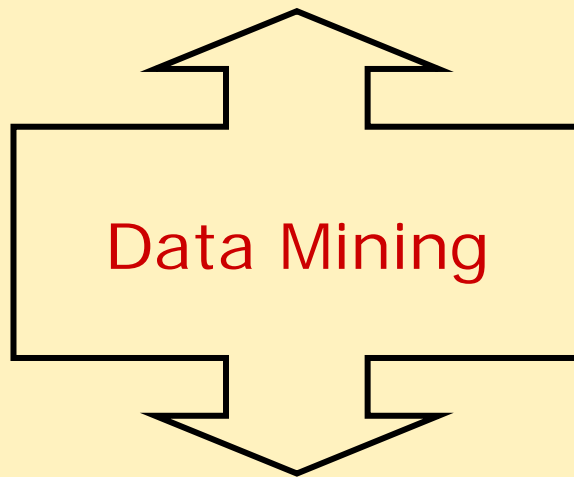
# The Data Mining Team

- The recommended data mining team consists of:
  1. a domain specialist,
  2. a data specialist,
  3. a data miner.
- The data miner is often expected to fulfill all three roles !!!
- This implies that the data miner produces models of the same quality as those produced by a multi-person team

# Multiple Personalities of Data Mining

## Research Analysis Tools

(e.g. decision trees, neural networks)

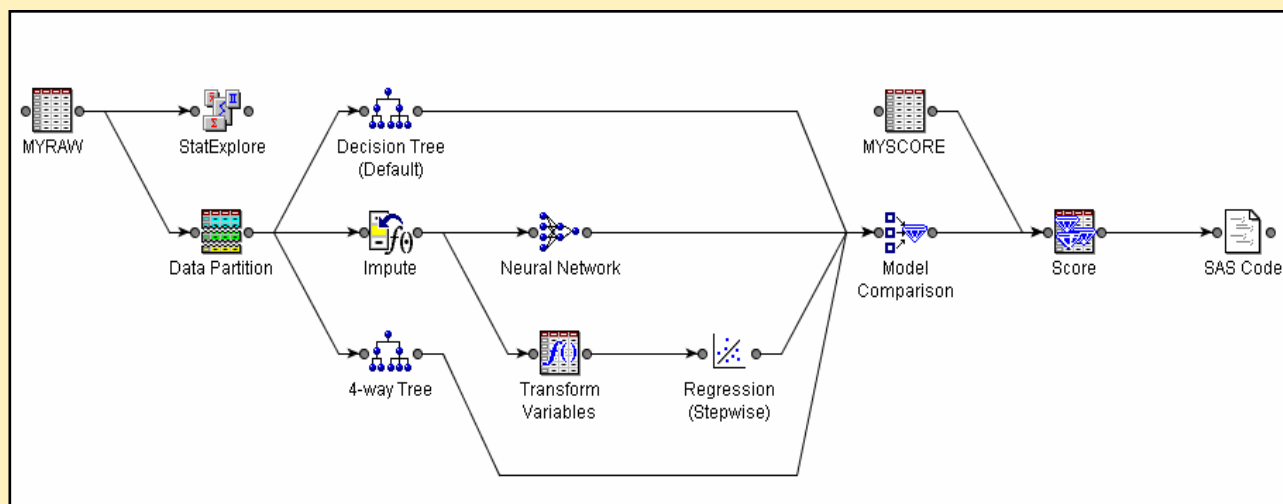


## Business Analysis Tools

(e.g. interactive visualization, dashboards)

# Research Analyst Tools

- Tools targeted primarily at the data miner/modeler:
  - Statistical : imputation, outlier detection, survival analysis, regression, clustering, etc.
  - Machine Learning : trees, neural nets, genetic algorithms, support vector machines, etc.
- Requires **extensive** (and **expensive**) knowledge to use



# Business Analysis Tools

- These tools are targeted primarily at the business end-user
- Business users don't want to create a decision tree or neural network, they want to solve a specific problem.
- Present decision support information in user-friendly ways, e.g. dashboard, campaign mgmt, sales automation

Measures							
	Name	Period	Actual	Target	Performance	Performance Range	
	Average cost of each product per quarter	JUN2005	3.65	3	82%		<input type="checkbox"/>
	Average cost per inspection per week	JUN2005	1.25	0.75	60%		<input type="checkbox"/>
	Average cost per package to move products to warehouse	JUN2005	2	1.25	63%		<input type="checkbox"/>
	Average job loss time due to preventable accidents per week/100 employees	JUN2005	7.5	6	80%		<input type="checkbox"/>
	Average number of customer complaints per week	JUN2005	9.5	8	84%		<input type="checkbox"/>
	Average number of inspections failed per week	JUN2005	12	9	75%		<input type="checkbox"/>
	Average number of package inquiries received by phone from customers per week	JUN2005	200	150	75%		<input type="checkbox"/>
	Average number of packages lost per week	JUN2005	4	2	50%		<input type="checkbox"/>
	Average number of packages tracked online by customers per week	JUN2005	800	1200	67%		<input type="checkbox"/>
	Average quarterly profit of each product	JUN2005	7.5	9	83%		<input type="checkbox"/>

Rows 1 - 10 of 13

## Why Now?

In 1963, machine limitations restricted the total number of variables which could be considered at one time to around **25** !!!

### Moore's Law

The information density on silicon-integrated circuits doubles every 18 to 24 months.

# Why Now?

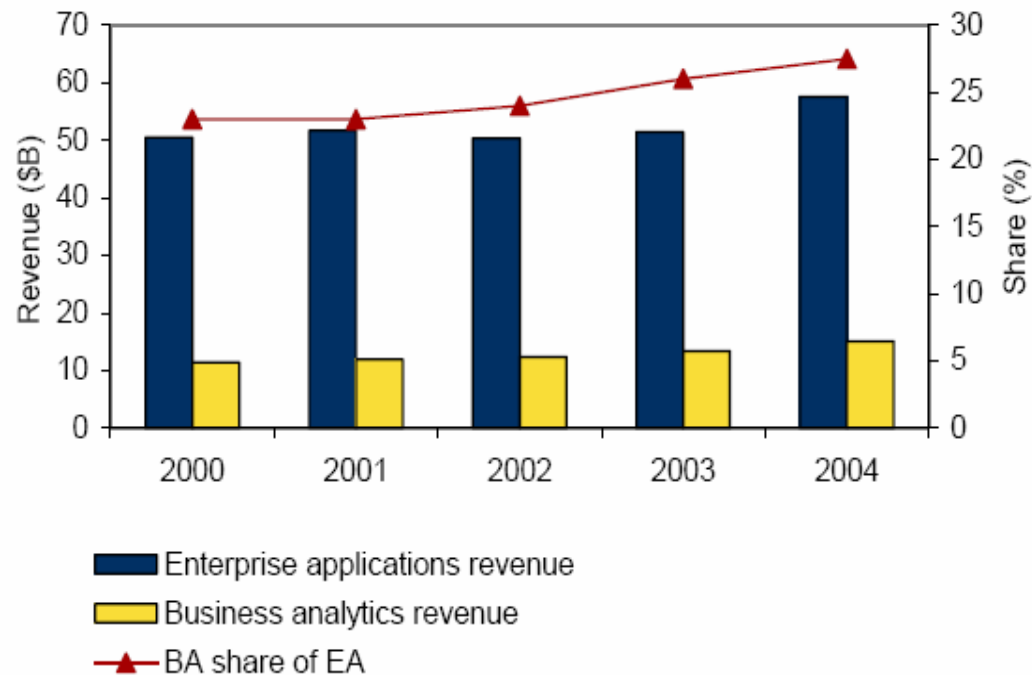
## Data Deluge

### Corollary to Moore's Law

The amount of information doubles every 18 to 24 months.

# Investment Trends

Worldwide Business Analytics and Enterprise Applications Software Revenue, 2000–2004



Note: Enterprise applications (EA) and business analytics (BA) are defined in the Market Definitions section.

Source: IDC, 2005

## Other Trends

- Increasing automation of the decision-making process, in particular specific-purpose mining applications
- Integration of structured and unstructured data (text mining)
- Emergence of parallel processing
- Embedding of data mining functionality in RDBMS
- New modeling algorithms (e.g. support vector machines)
- Increased use of operations research methods in mining

# Questions

