# Enterprise Data Warehousing in Support of Data Mining for Fraud and Abuse Detection

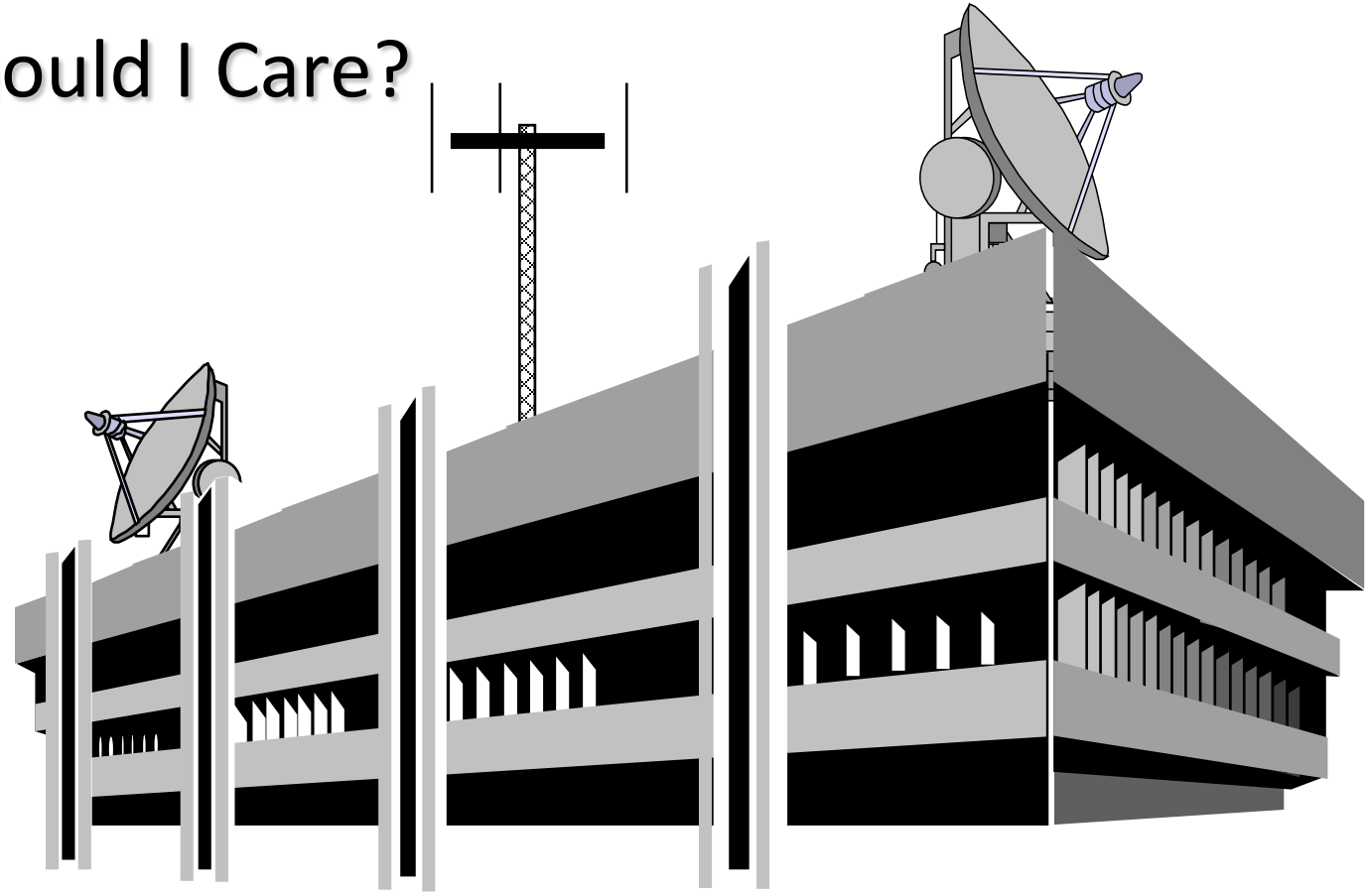
State of New Jersey
Office of Information Technology
Data Management Services
Dan Paolini, Deputy CTO

# Dan Paolini

- Deputy CTO for Information Architecture and Data Management Services – State of NJ OIT
- Certified Data Management Professional, Master's Level, with ICCP/DAMA
- Certified Business Intelligence Professional, Master's Level, with ICCP/TDWI
- Certified Data Protection Specialist, The Data Management Institute
- Former Certified Management Accountant, National Association of Accountants
- Former Vice President for Standards, DAMA Foundation
- Former Moderator, DAMA Data Architecture Professional Group
- Former Board Member, New Jersey Chapter of DAMA
- Author of the award-winning programmer's toolkit, ***PALADIN***
- Member of development team for the award-winning editor, ***PLAYRIGHT PRO***
- Contributing Editor, 1992 - 1995, for the technical magazine, ***Paradox Informant***
- Technical editor for three books on database analysis and queries
- Frequent speaker at more than sixty technology events in North America, Asia, and Europe, including keynote speaker at ten technology events in North America and Europe
- May 2002 Leadership Award recipient from the "Government without Boundaries" program of the United States Office of Management and Budget
- Volunteer Firefighter/EMS since 1972 including chief officer in four different organizations
- NFHS Soccer and Lacrosse Official, USFF Referee, USSF Referee, Assigner, and Associate Assessor
- Keyboard/Sax/Vocals for D*Luxe ♫

# Today's Session

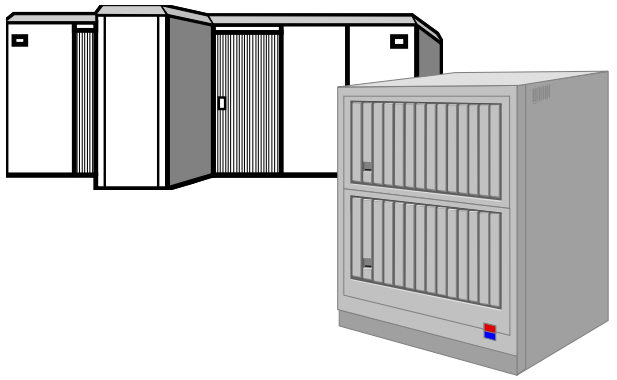
- What is Data Warehousing?
- What is Data Mining?
- Why Should I Care?



# The Business Problem:

***On-line Transactional Processing Systems:***

- Tuned for transactions, not decision support
- Unable to store history without affecting operating performance
- No standards for data definition and naming
- Queries never get same answer twice
  - Dynamic data
  - No “what ifs” possible
  - No summarized data
  - No integrated data



# Limitations to Traditional Reporting

- Processing Windows
- Programming Effort
- Inability to Integrate Disparate Systems
- Multiple End-User Communities requiring Multiple Independent Extracts
- Storage Space
- Report and Query Tools
- The Same OLTP Staff supports DW

# What is Data Warehousing?

***At its core, data warehousing is simple:***

- Insure that there is a single, consistent, integrated view of historical business data.
- Offload from On-Line Transaction Processing (OLTP) Systems and the Teams that support them the requests for non-operational extracts and reports so that the Systems and Teams can focus on their operational mission.
- Make information available in the format required as cost-effectively as possible.

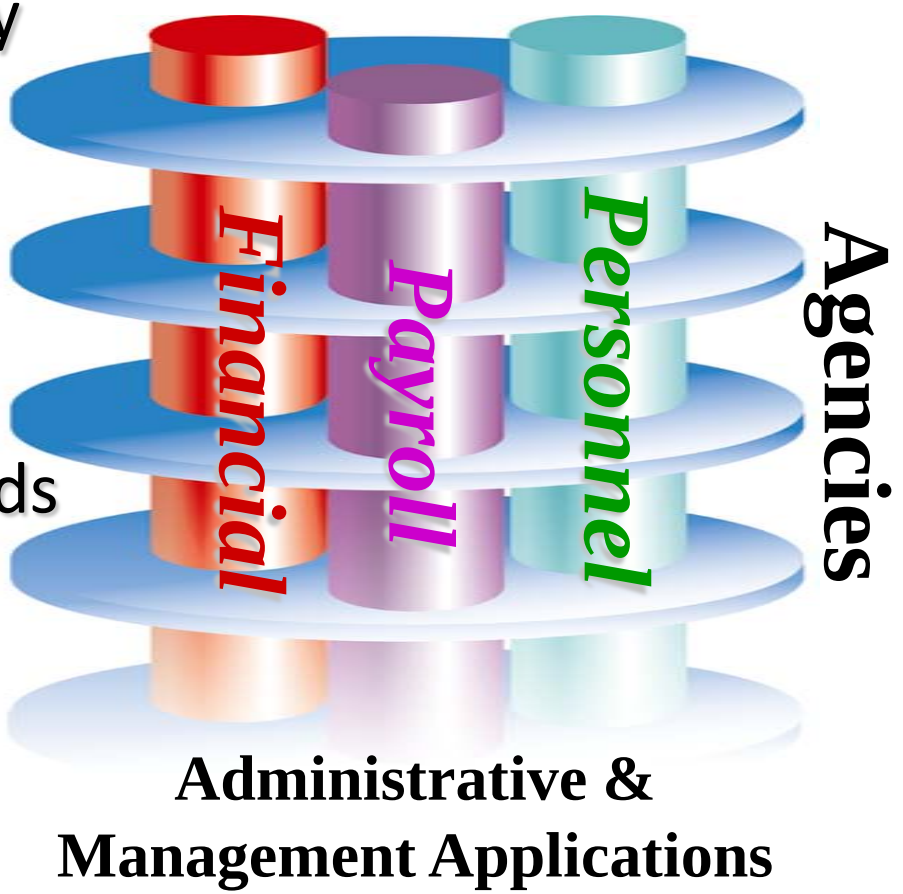
# What It Isn't

***Data Warehousing isn't…***

- A big building for your data
- Going to take all of the data and roll it up into one giant database
- Going to replace the need for transactional data systems
- Going to fix underlying data quality problems in the transactional environment

# Single Version of the Truth*

- Insure that there is a single, consistent, integrated view of historical business data.
  - Controlled Redundancy
  - Fit for Purpose
  - Agreed Definitions
  - Agreed Sources of Record
  - Identified Data Stewards




** In Context*

# Protect OLTP Systems

- Remove from the OLTP Environment:
  - Historical Data
  - Ad Hoc Query Requests
  - Non-Operational Reporting
  - Outbound Extract Processing
- And Most Importantly…
  - Those Pesky Non-Operational Users

# End-User Focus

- Make it available in the format required.
- End-Users want
  - Reports Sorted This Way
  - And That Way
  - Historical Data
  - Ad Hoc Queries
  - Integrated Views
  - Summarized Data
- And they want it Fast, Cheap and Easy

# Three Objectives of DW

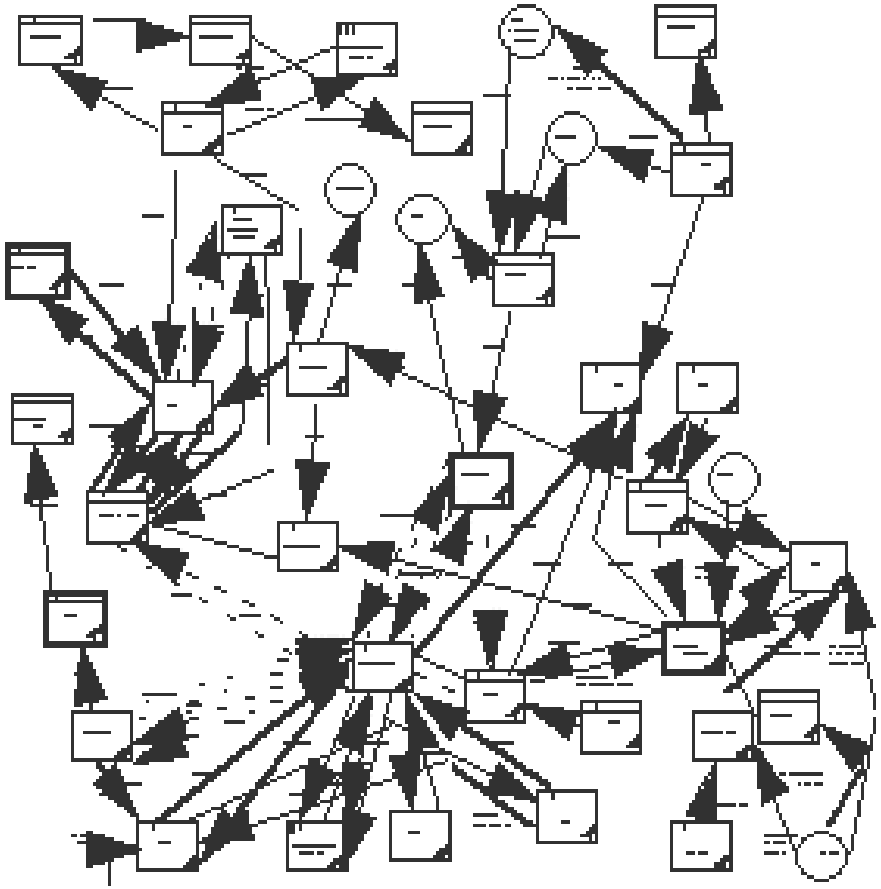
***You are only building report silos* and not doing Data Warehousing unless…***

- Data is integrated into a single, consistent version of the truth in a logical "data hub" separate from the operational systems
- Data to meet the needs of interfaces and non-operational reporting comes from that logical data hub
- Data is delivered in the form necessary to meet the needs of consumers, not the format in which it was collected

**** A Report Silo is another term for "Higher Cost, Lower Quality Data"***

# Our Goal

**A “Spaghetti Network” design for Data “Sharing”**

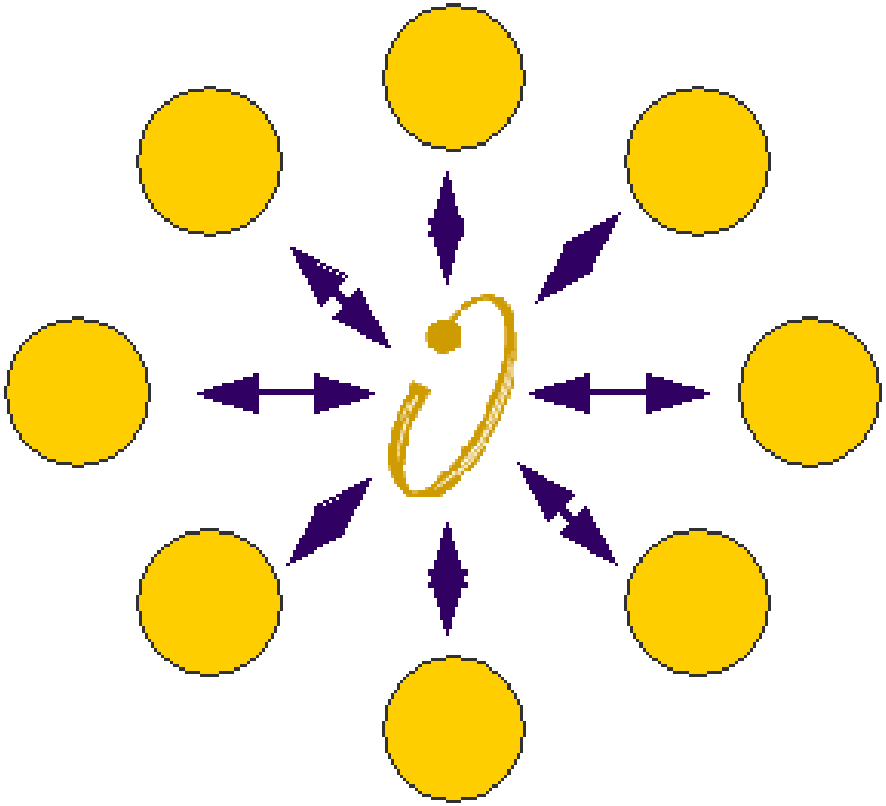


**We want to get from here……**

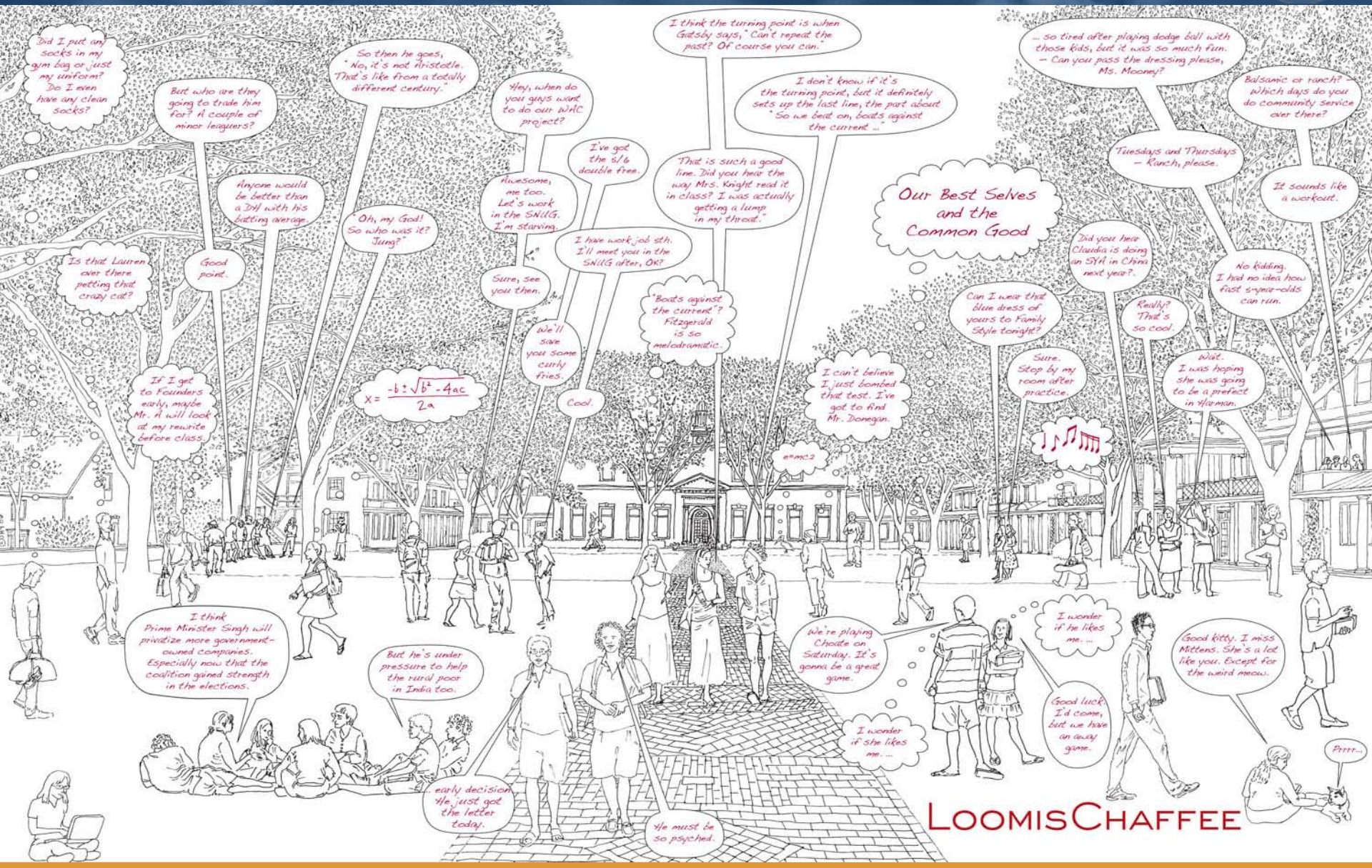
# Our Goal

**A “Hub and Spokes” design for Information Reuse**

**…….to here.**



# How Do We Know What to Ask?



# Data Mining

## The "Question Behind the Question"

# Definitions?

- Data Mining
  is the process of sifting through large amounts of data to produce data content **relationships**. It is a technique that uses software tools geared for the user who typically does not know exactly what to look for, but wants to identify patterns or trends that might point in what direction to look.

# More Definitions?

- Data Mining is a statistical analysis of data for patterns and clusters. Statisticians determine parameters for a pattern search, and then the software goes off to prove or disprove the pattern.
- Data Mining Tools can learn from earlier analyses and perform more intelligent (heuristic) analyses. They can look for patterns without guidance, to find relationships in the data that a human would never see. These techniques go to the heart of **fraud detection** and homeland security.

# So, Really... What is it?

- Data mining is the analytical process between raw data and business decisions.
- Data mining can be painfully complex or surprisingly simple.
- Data Mining does not produce answers, it produces strategies.
- It determines **semantic distance**.

# Why Is It Important?

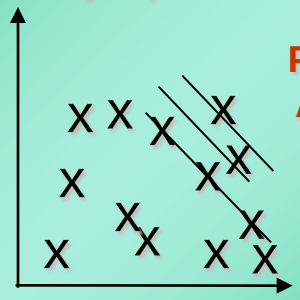
- ***Relationships:***
  What is Normal, Not Normal
  (Clustering, Regression)
- ***Outliers:***
  Best Practices, Poor Practices, Data Quality
- ***Link Analysis:***
  Association Rules, Suspicious Patterns
- ***Root Cause Analysis:***
  Forward Modeling, Reverse Engineering
- ***Balanced Scorecards***

# Results of Data Mining

## Google

Google's exploitation of the Web's link structure

"Quality" (credibility)



Page Rank Algorithm

Query relevance

## amazon.com


**Data Mining determines "How-Where-Why to Look", not what gets found**

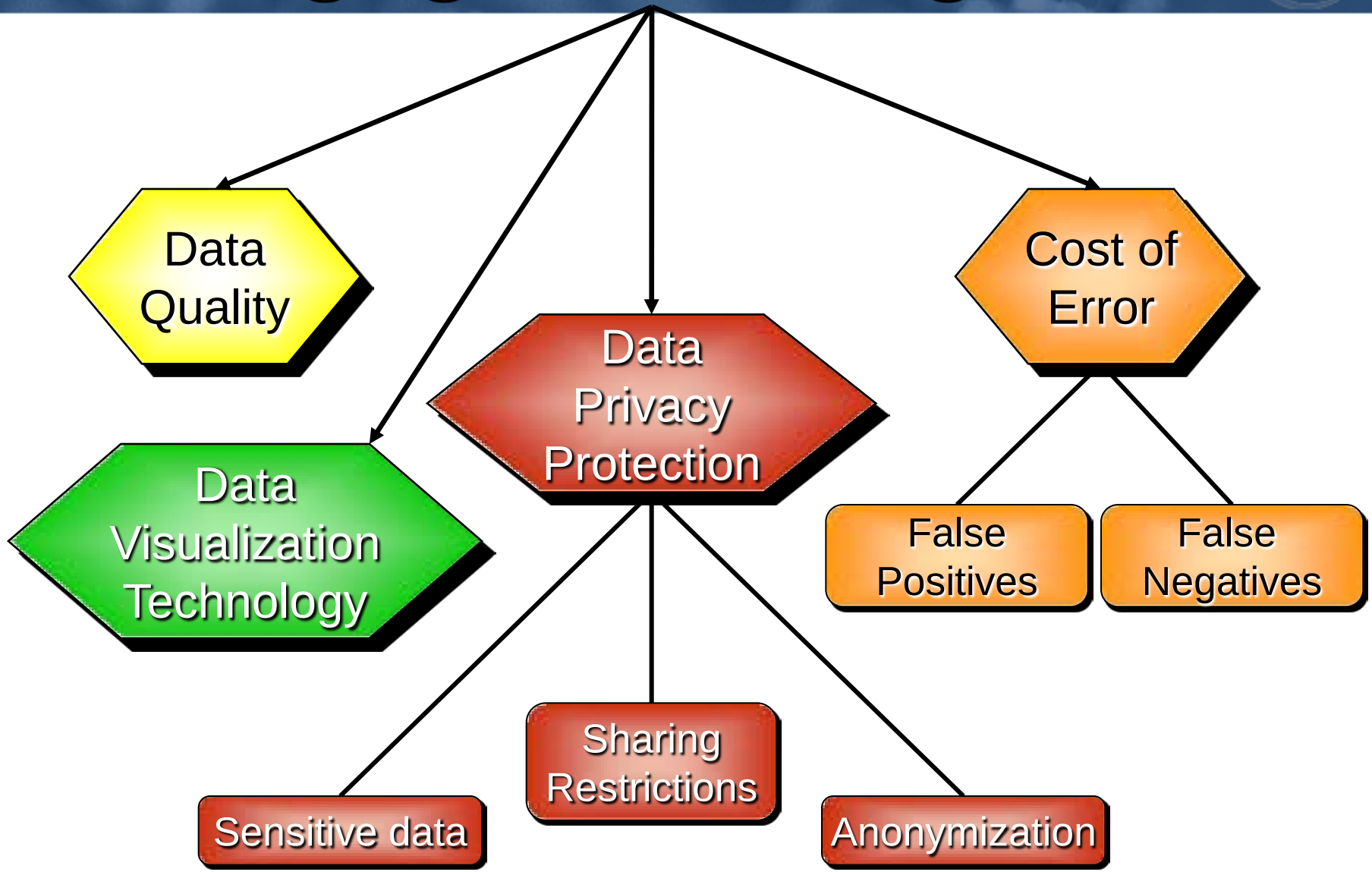
# Data Mining Approaches

- Embedded Analytics in both applications and BI tools are typically quite rudimentary and OK for starting with data mining or for moderate ambitions.
- General analytics are programming environments not geared toward special data mining needs.
- The No. 1 criterion for these tools: **the availability of staff members and how mathematically inclined they are.**

# Data Mining Approaches

- General Data Mining tools can address diverse needs with multiple algorithms and approaches.
- Algorithm-specific tools (e.g., decision trees and neural networks) implement one algorithm well.
- Application-specific tools (e.g., for **fraud detection**) focus on a single application (horizontal) or industry (vertical).
- Analytical components (from external service providers) are used for tactical business objectives when appropriate subject-matter expertise cannot be found in-house.

# Emerging Data Mining Issues



# Risk Factors and Mitigation

- Wrong Staffing Resources
  - Need Both IT and Business
  - Need Statistical Background
- Too Strategically Focused
  - Don't Wait for the Data Warehouse
  - Don't Do Everything at Once
- Too Unrealistic
  - Need "Right" Data and Assumptions
  - Need Business Orientation
  - Need Plans for Deployment of Knowledge

# Success Criteria

- Get the right team together: Analysts, Statisticians, and Business Domain Experts.
- Data Mining and Business Intelligence teams work best when closer to the business. Put the team close to the business process owner. On the other hand, the Data Warehouse team works best as a centralized core resource.
- Start simple and stop with a high ROI. Do not try to overachieve.
- Outsource if the skills are not available.
- Use an incremental ROI beyond a predetermined base line.
- Data mining does not substitute for creativity and insight! Look for new data sources, and stay “plugged in”.


# Why Should I Care?

- Data Warehousing is a more
  - Rational
  - Proactive
  - Efficient
  - Flexible
  - Leveraged
  - Cost-effective

> *"The field of* knowledge science
> — *or* knowledge engineering —
> *is about to emerge. Large enterprises are advised to consolidate their staffs into centers of excellence."*
>
> — Gartner

way to do what we have always done:

- Provide data interfaces to other systems
- Provide information to users
- Improve performance of transactional systems
- Manage security and access to data

# Why Should I Care?

- Data Mining helps your agency sift through diverse and complex data sources to determine which questions will provide the most meaningful answers (information) from reporting systems (i.e. so you can build better data marts.)

# Enterprise Data Warehousing in Support of Data Mining for Fraud and Abuse Detection

State of New Jersey
Office of Information Technology
Data Management Services
Dan Paolini, Director