The Data Discovery Revolution: Changing the Economics of Data Governance

Data In the News:
Data Consistency Problems

Software/ Solutions

Data slipups

Rick Whiting , 10-May-2006

A home valued at US$121,900 somehow wound up recorded in Porter County's computer system as being worth a whopping US$400 million. Naturally, the figure ended up on documents used to calculate tax rates. By the time the blunder was uncovered in February, the damage was done.

It's a nightmare scenario--and one like it could be yours. Bad data remains a major cause of botched marketing campaigns, failed CRM and data warehouse projects, angry customers, and lunkhead decisions. Despite all we know about the importance of data scrubbing and quality management, many companies are still using data that's redundant, incomplete, conflicting, outdated, and just plain wrong.

IT Infrastructure News

84% report problems with master data management

02 November 2006

Poor 'master data' is causing problems for organizations trying to analyse data across their IT systems according to a new report by the Data Warehousing Institute (TDWI)

EXEROS

Data in the News: Sensitive Data Leaks

eWEEK.com ENTERPRISE NEWS & REVIEWS

NEWS · REVIEWS · OPINIONS · CASE STUDIES · TOPICS ·

SEARCH

Security

H&R Block Mailing Reveals Customers' SSNs

By Paul F. Roberts

January 3, 2006

Some H&R Block customers who received free copies of the company's TaxCut software also had their Social Security numbers exposed, according to a company spokesperson.

H&R Block sent a letter to customers in late December saying that a tracking number used on packages containing TaxCut contained the customer's Social Security number as part of a unique, 47-digit tracking number.

EXEROS

Data in the News:
Data Governance and Data Sprawl

leading off November 2006

Data Governance: Struggling with the Numbers Crunch

Banks collect large volumes of data to guide their strategies. But much of that output is too fragmented and redundant to help.

By Glen Fest

Banks compile ever more diverse sets of data to help decipher business strategy and compliance issues. But they are discovering a problem with their analytical operations: the data itself.

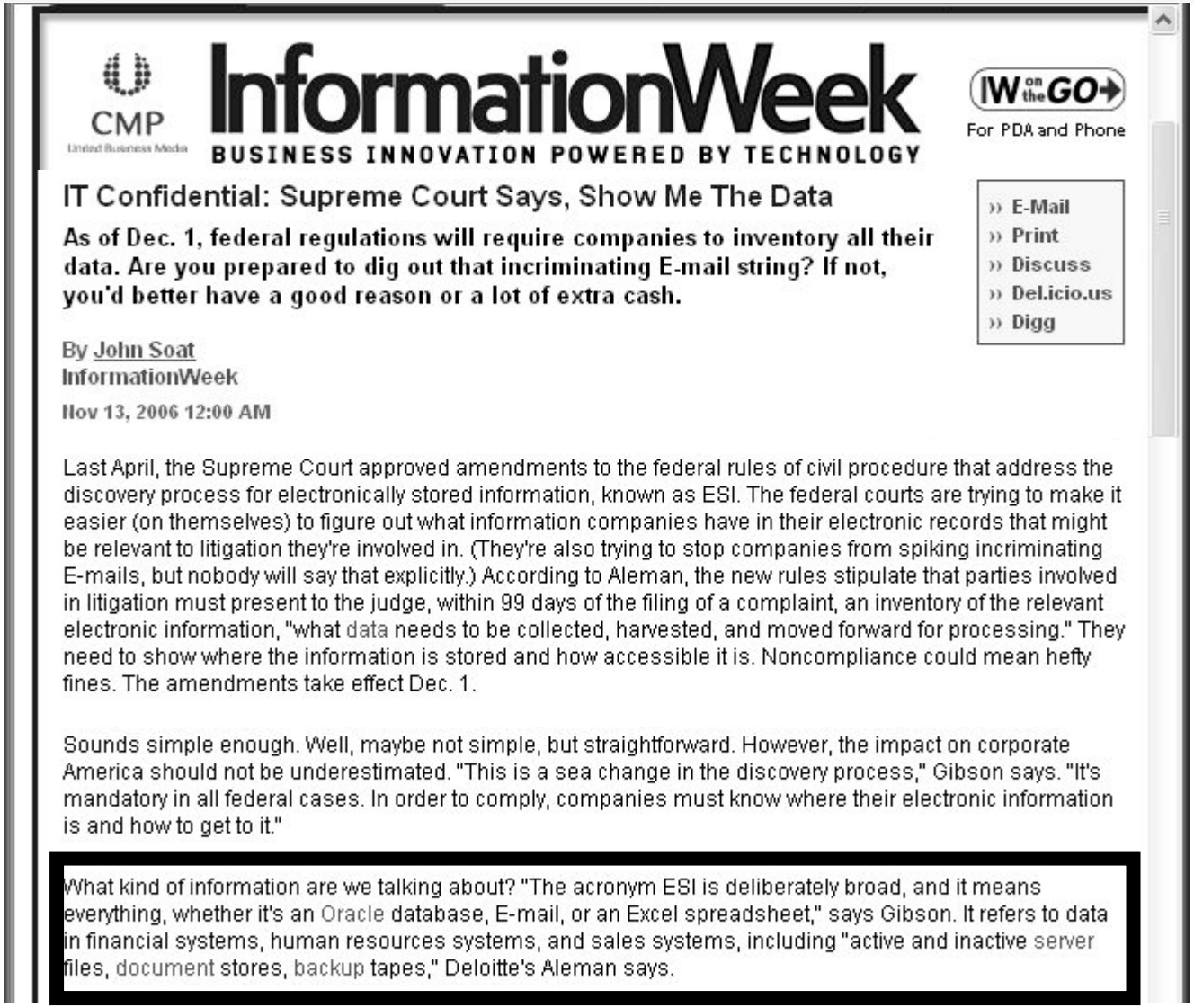
The glut of new information gleaned from siloed business units, IT systems, products and devices are for the most part leaving banks with fragmented and duplicative data. It's estimated by TowerGroup that financial services groups spend at least 50 percent of their data management resources simply trying to resolve problems in rendering data into usable metrics and resources.

BANK TECHNOLOGY NEWS btn

According to TowerGroup research by Kopp, almost no institution today has in place a formal data governance process or solution to effectively deal with redundant and sprawling data. The result has been the muck-up of several enterprise automation initiatives, and the bungling of CRM implementations.

EXEROS

Data in the News: E-Discovery



CMP
United Business Media

InformationWeek
BUSINESS INNOVATION POWERED BY TECHNOLOGY

IW on the GO
For PDA and Phone

IT Confidential: Supreme Court Says, Show Me The Data

As of Dec. 1, federal regulations will require companies to inventory all their data. Are you prepared to dig out that incriminating E-mail string? If not, you'd better have a good reason or a lot of extra cash.

›› E-Mail
›› Print
›› Discuss
›› Del.icio.us
›› Digg

By John Soat
InformationWeek
Nov 13, 2006 12:00 AM

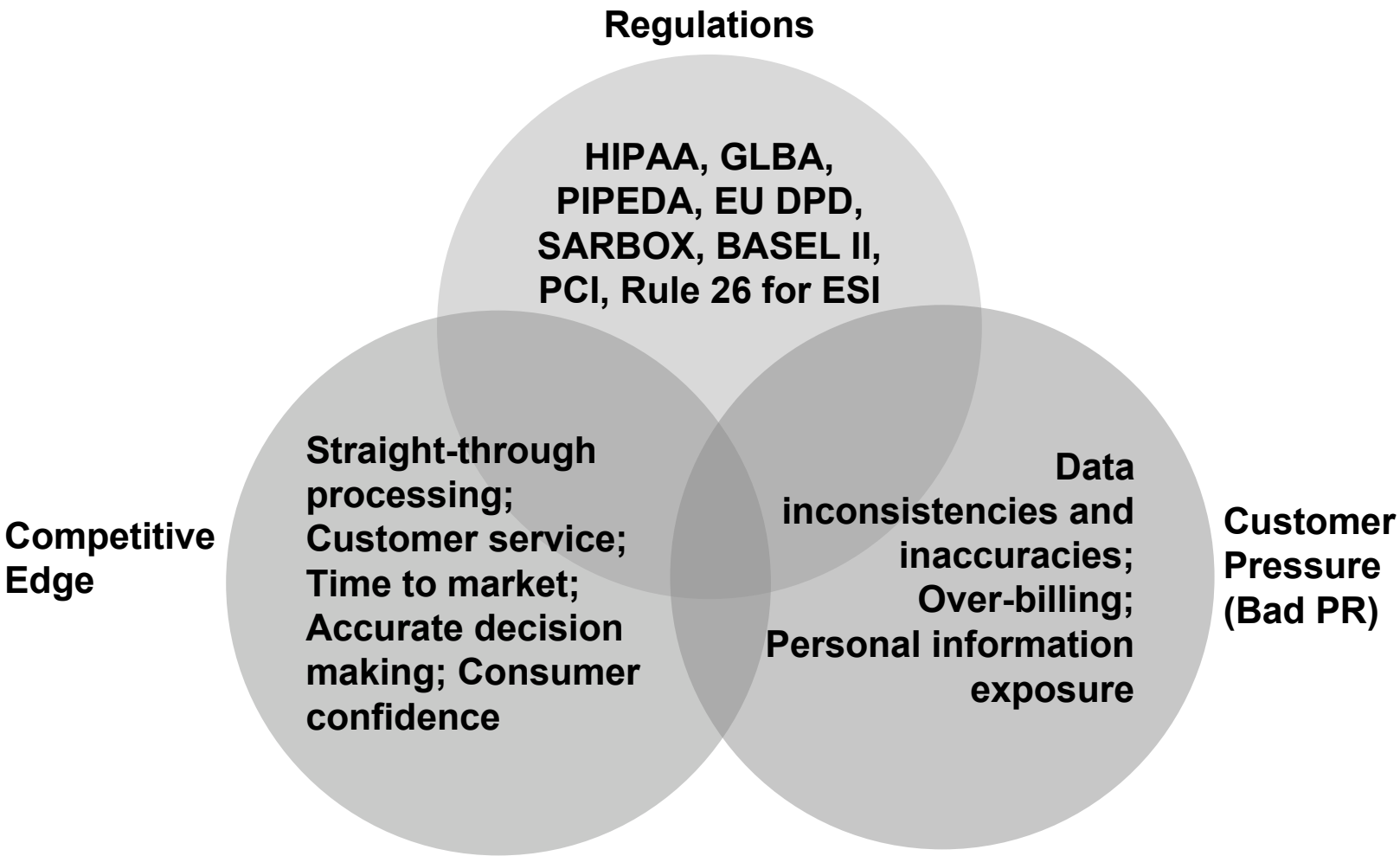
Last April, the Supreme Court approved amendments to the federal rules of civil procedure that address the discovery process for electronically stored information, known as ESI. The federal courts are trying to make it easier (on themselves) to figure out what information companies have in their electronic records that might be relevant to litigation they're involved in. (They're also trying to stop companies from spiking incriminating E-mails, but nobody will say that explicitly.) According to Aleman, the new rules stipulate that parties involved in litigation must present to the judge, within 99 days of the filing of a complaint, an inventory of the relevant electronic information, "what data needs to be collected, harvested, and moved forward for processing." They need to show where the information is stored and how accessible it is. Noncompliance could mean hefty fines. The amendments take effect Dec. 1.

Sounds simple enough. Well, maybe not simple, but straightforward. However, the impact on corporate America should not be underestimated. "This is a sea change in the discovery process," Gibson says. "It's mandatory in all federal cases. In order to comply, companies must know where their electronic information is and how to get to it."

What kind of information are we talking about? "The acronym ESI is deliberately broad, and it means everything, whether it's an Oracle database, E-mail, or an Excel spreadsheet," says Gibson. It refers to data in financial systems, human resources systems, and sales systems, including "active and inactive server files, document stores, backup tapes," Deloitte's Aleman says.

EXEROS

Various Forces are Increasing the Attention Paid to Data as a Strategic Asset



What are Organizations Doing?

Wall Street Firm:
Data Consistency will Increase Profitability

Business Problem:

- Transaction errors are expensive and the risk of regulatory fines due to inconsistent reference data is unacceptable

Proposed Solution:

- Deploy a master data management solution

Roadblock:

- 9 years to determine the business rules that relate the master data system to legacy systems

Status:

- Project cancelled

EXEROS

Health Insurance Company: Outsourcing Development vs. Data Compliance

Business Problem:

- Data must be sent to India for offshore application development.
- Sensitive data must be masked for HIPAA compliance

Proposed Solution:

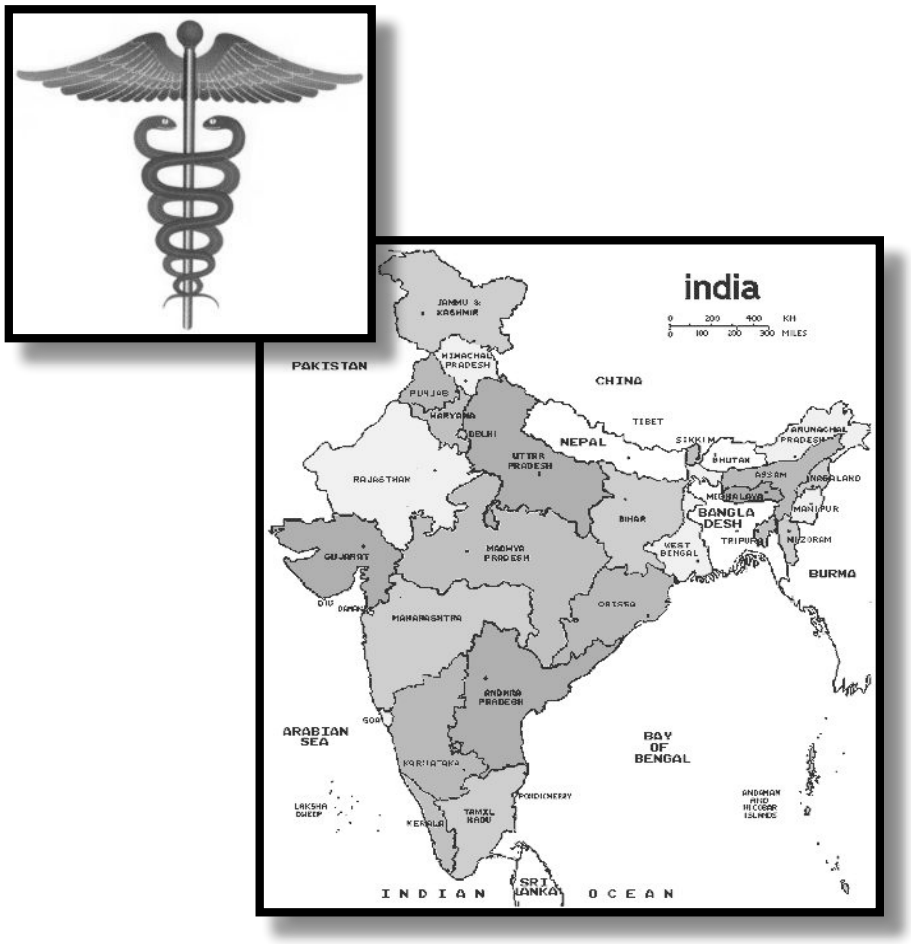
- Mask sensitive data before sending it outside the company

Roadblock:

- Sensitive data, where is it?
- Can two sets of data that individually contain no sensitive data be combined to make it sensitive?

Status:

- Manual discovery of sensitive data slows outsourcing to a crawl



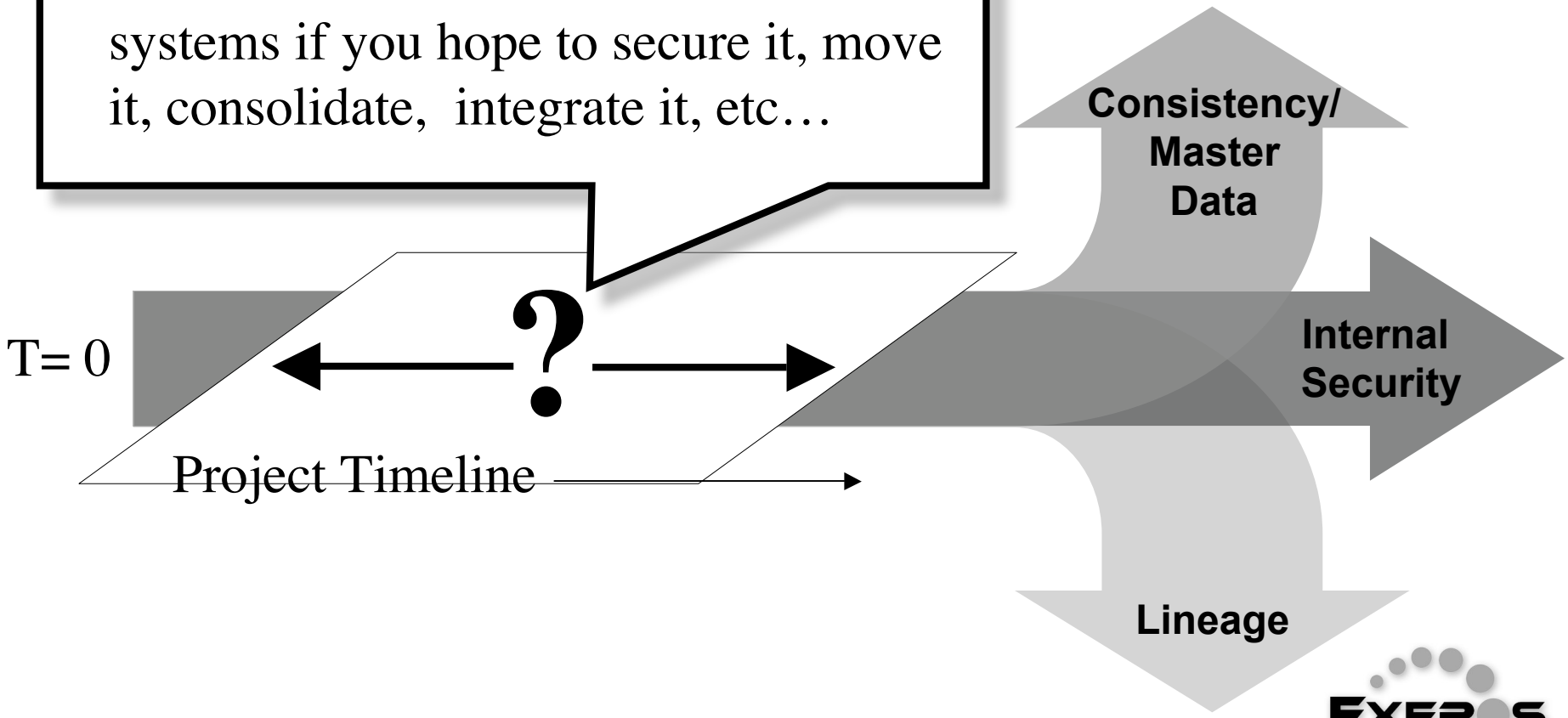
The Economics of these Projects just doesn’t look good.

Why is that?

The Common "?" in the Project Schedule

Data Relationship Discovery

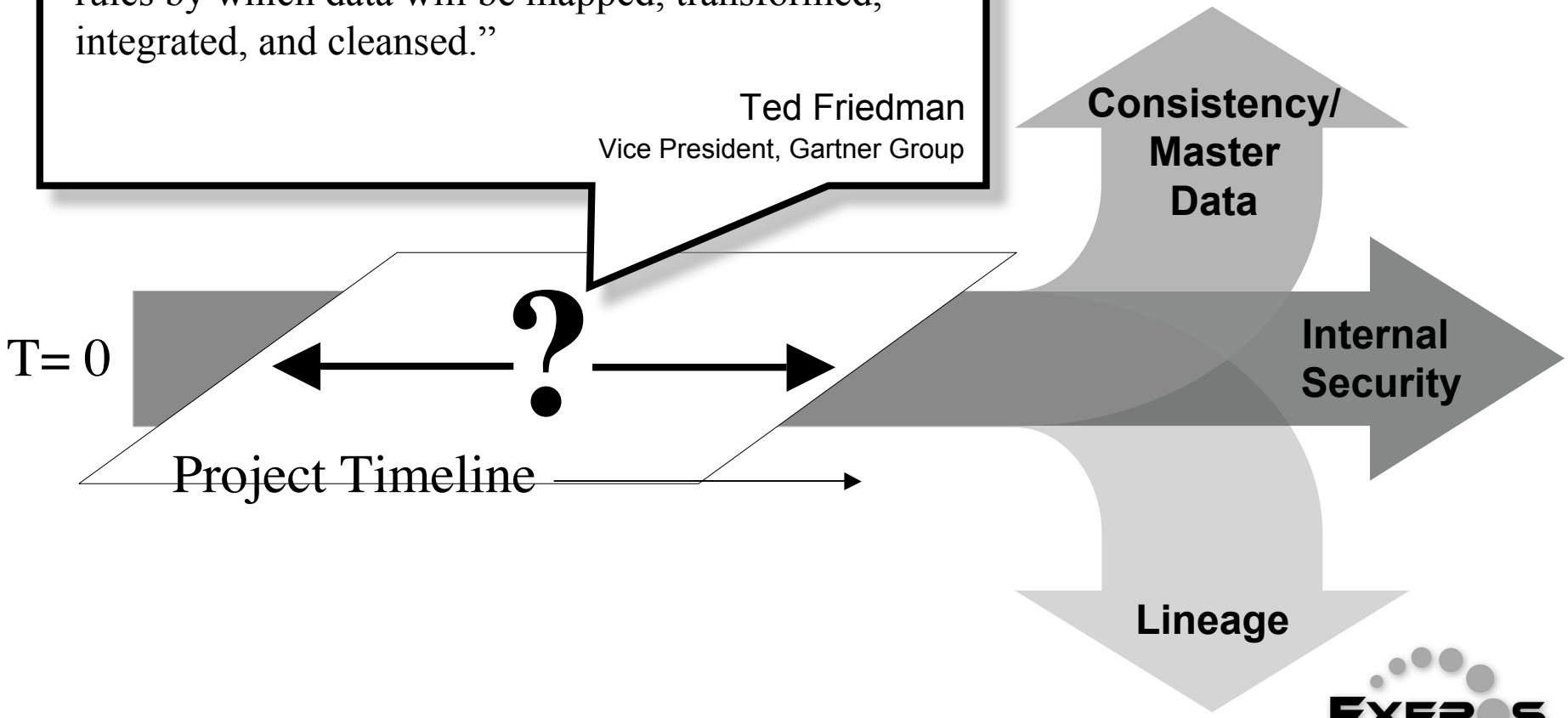
You have to know where your data is, how it flows and relates across systems if you hope to secure it, move it, consolidate, integrate it, etc…



The Common "?" in the Project Schedule

"**70%** or more of the time and effort involved in completing most data integration projects is consumed by defining and implementing the business rules by which data will be mapped, transformed, integrated, and cleansed."

Ted Friedman
Vice President, Gartner Group



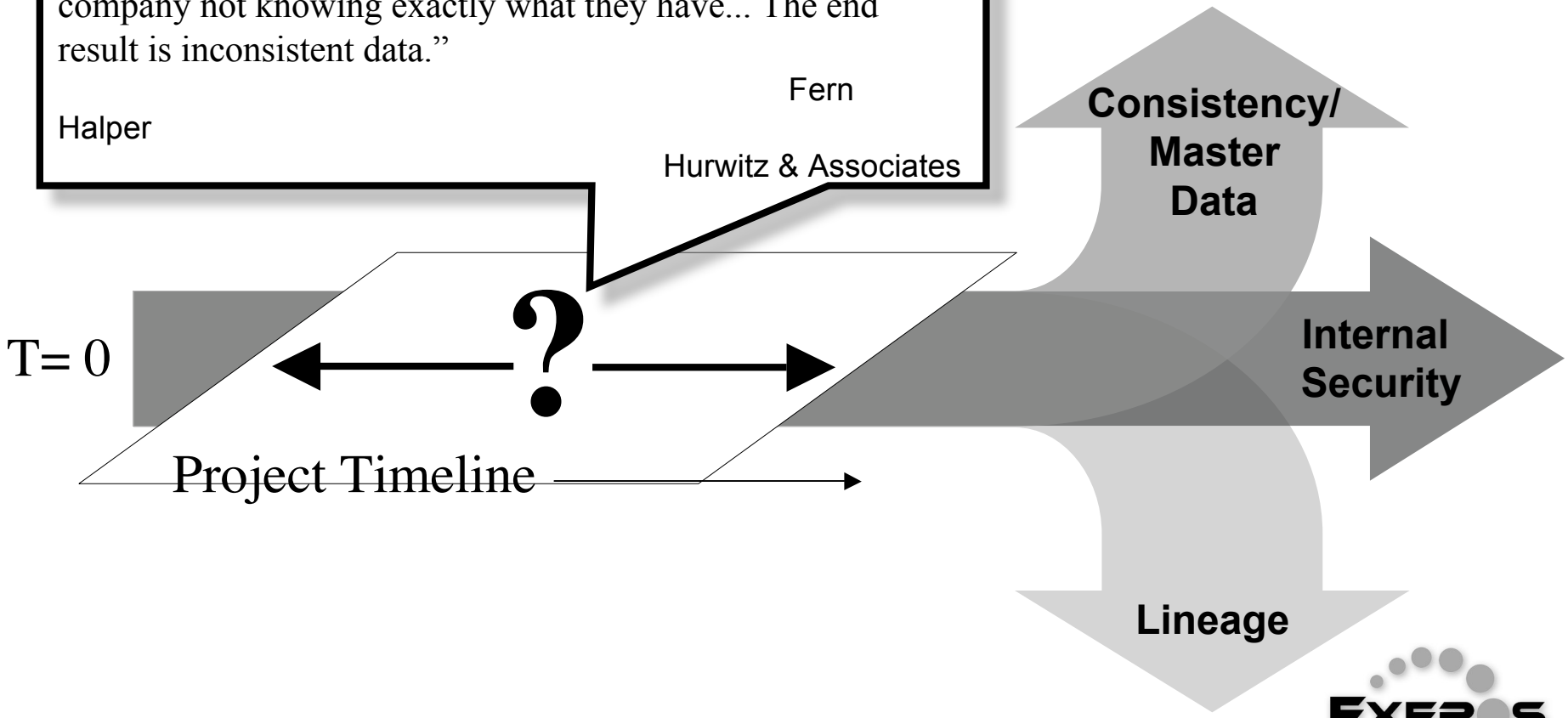
EXEROS

The Common "?" in the Project Schedule

"Design specifications get lost or outdated, subject matter experts leave companies, databases and business rules get changed without updating documentation, mergers and acquisitions wreak havoc on databases, all leading to a company not knowing exactly what they have... The end result is inconsistent data."

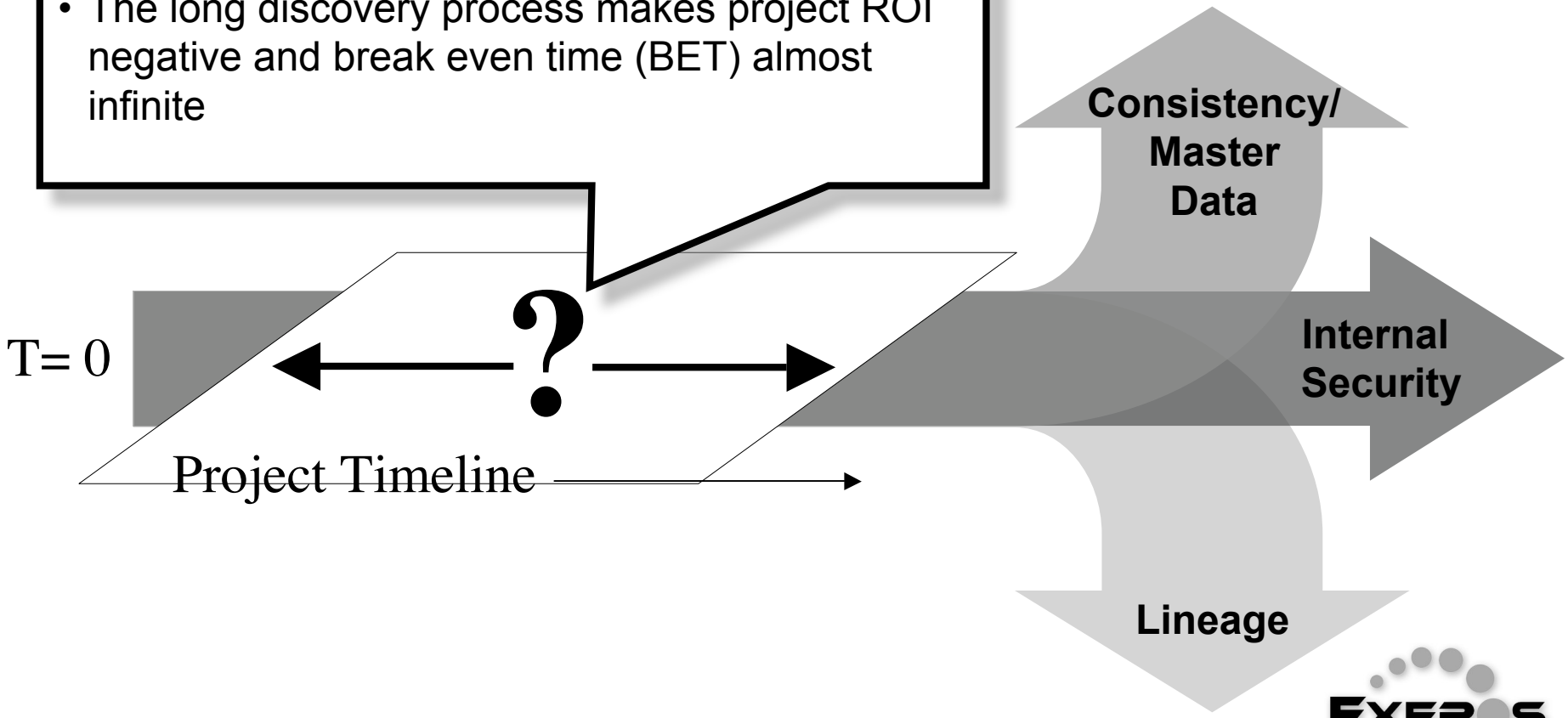
Fern Halper

Hurwitz & Associates



"?" Negatively Affects Project Economics & ROI

- Lack of consistency in discovering cross-system business rules limits your ability to leverage data maps from one project to another
- The long discovery process makes project ROI negative and break even time (BET) almost infinite



The Myth of:
“But we know our data!”

A Common Myth: “We know our data”

I’m a professional. Of course I know my data!

But, once it leaves my hands, it is someone else’s problem!

Wow, that transformation is complex. Are you sure that is in my data?

I’m going to start my own consulting firm

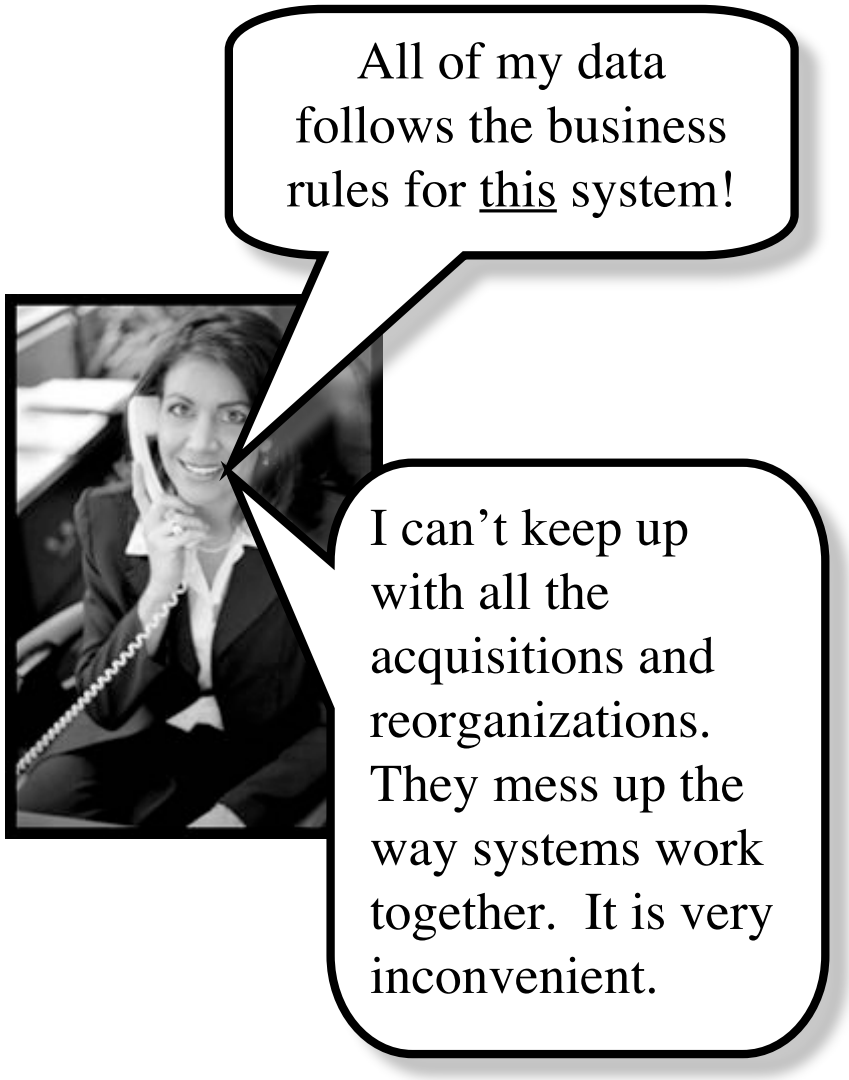
- Subject matter experts (SMEs) only know their own systems
- But they can’t tell you how it changes and is transformed as it moves from system to system
- Relationships between systems are complex:

```
CASE WHEN VTIB1_PROC_RULE_C.B1_PROC_RULE_ID in (
'370000', '370100', '400088', '455622', '474100', '4742001',
'47430011', '496604') THEN substr(VTIB1
_STIP_RESP_C.VERS_NUM, 1, 1) ELSE '0' END
```

- SMEs sometimes change jobs!

EXEROS

More of the Myth: “We know our data”



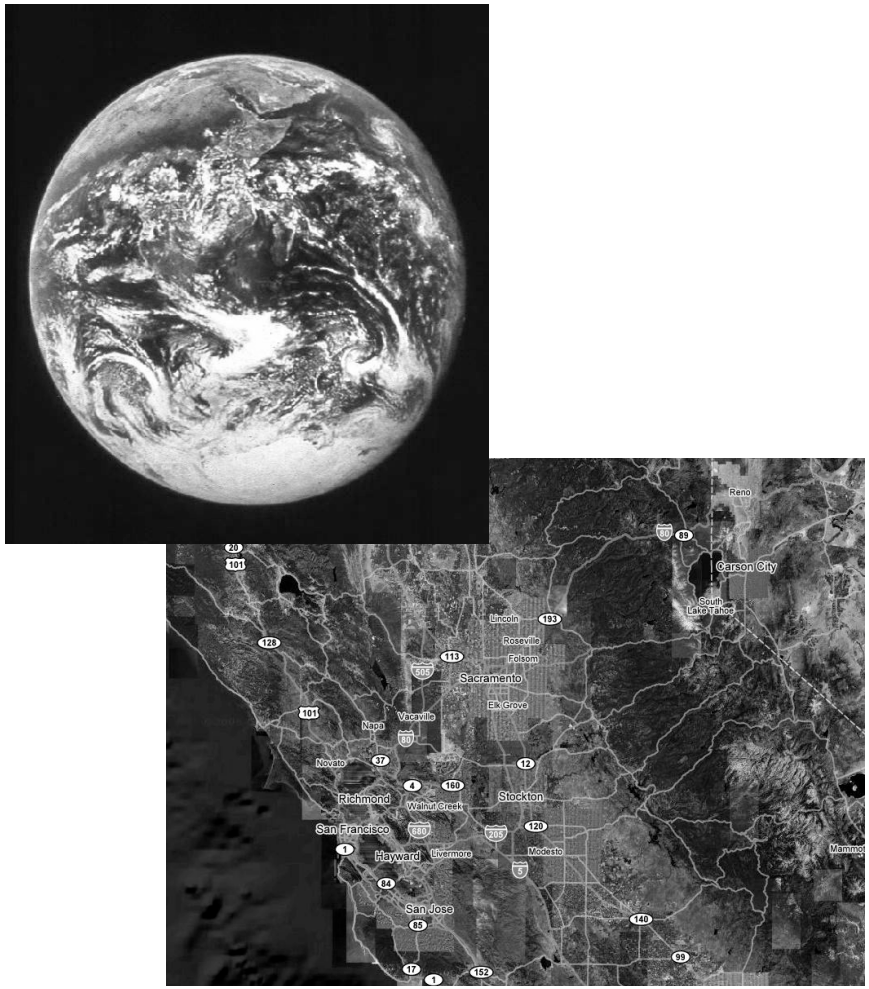
- Business rules are broken all the time as data crosses business and system boundaries:
  - 83 year old man in system A is a “youthful driver” in system B
  - Bond yield is listed as 5% in system X and 5.3% in system Y
- Exceptions result in lost revenue, customer dissatisfaction, and regulatory fines
- Business rules change as organizations change
  - Mergers and Acquisitions
  - New products or services
  - Products/services are retired
  - Reorganizations
  - New IT systems are added

EXEROS

The Reality

Organizations lack a global view of their “enterprise” data map…

Because it changes all the time



Current Trend: Data Governance

What is it?

- Data Integration **is to** Tactical **as**
  Data Governance **is to** Strategic

Definition

- Data Governance encompasses the people, processes and procedures to create a consistent, enterprise view of your data in order to:
  - Increase consistency & confidence in decision making
  - Improve data security
  - Decrease the risk of regulatory fines

EXEROS

The Challenge with Data Governance

- How do you do it?
  - Where is the sensitive data? Where does it come from?
  - What are the business rules and data relationships?
  - Where are the exceptions?
- How do you ensure a consistent, repeatable process?
- Too much time spent on the discovery process kills ROI:
  - The first step, DISCOVERY, is huge
  - DISCOVERY is a manual, slow, and error-prone process

EXEROS

Traditional Data Discovery Approaches

Traditional Answers to the Discovery Problem

- ETL, EAI, Cleansing
  - Not discovery solutions. They depend on discovery
- Metadata matching
  - Assumption based
  - Can't find transformations
- Profiling
  - Limited domain based matching
  - Can't find transformations
- Data analysts **manually** examine **data values** to figure out the data relationships in the corporate data map
- The most sophisticated tool commonly used today is:

Most Widely Used Data Relationship Discovery Tool

EXEROS

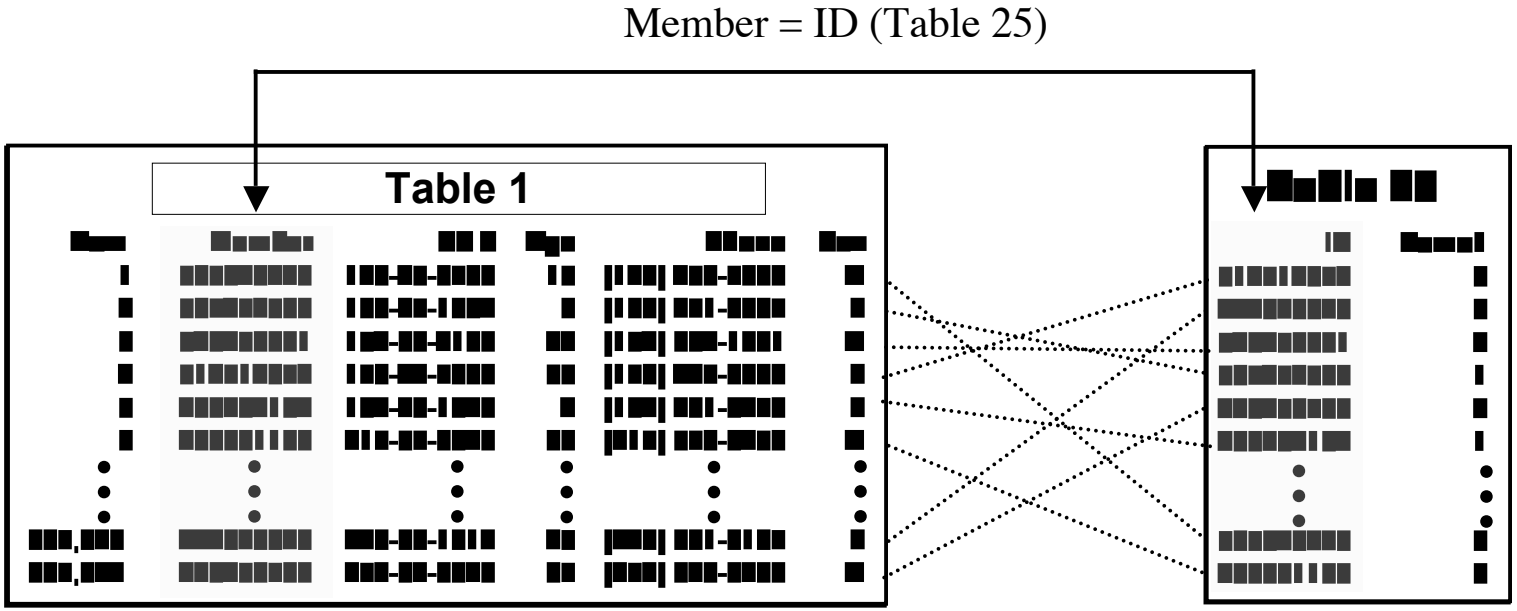
There is Hope!:
Automated Data-driven Mapping

Data-Driven Approach: Aligns Rows Across Datasets

Data-Driven Discovery Engine

Step 1: Discovery Engine analyzes the **data values** to **automatically discover** the key that aligns rows across disparate datasets:

- Works for hundreds of tables
- Works for millions of rows

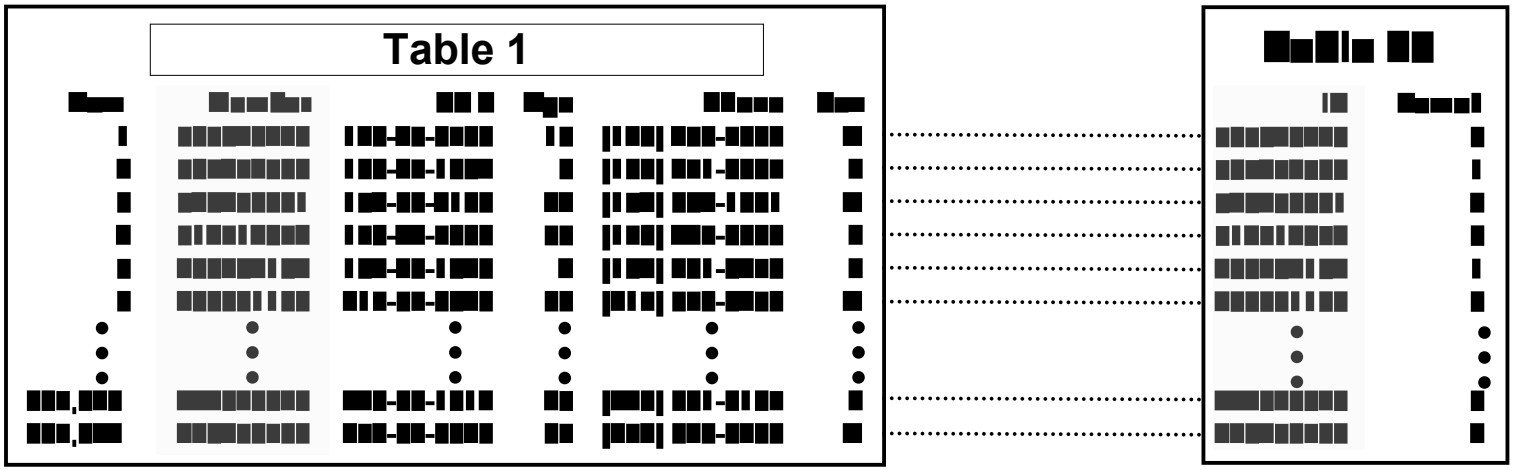


Data-Driven Approach: Aligns Rows Across Datasets

Data-Driven Discovery Engine

Step 1: Discovery Engine analyzes the **data values** to **automatically discover** the key that aligns rows across disparate datasets:

- Works for hundreds of tables
- Works for millions of rows



Data-Driven Approach: Discovers Business Rules & Sensitive Data

Data-Driven Discovery Engine

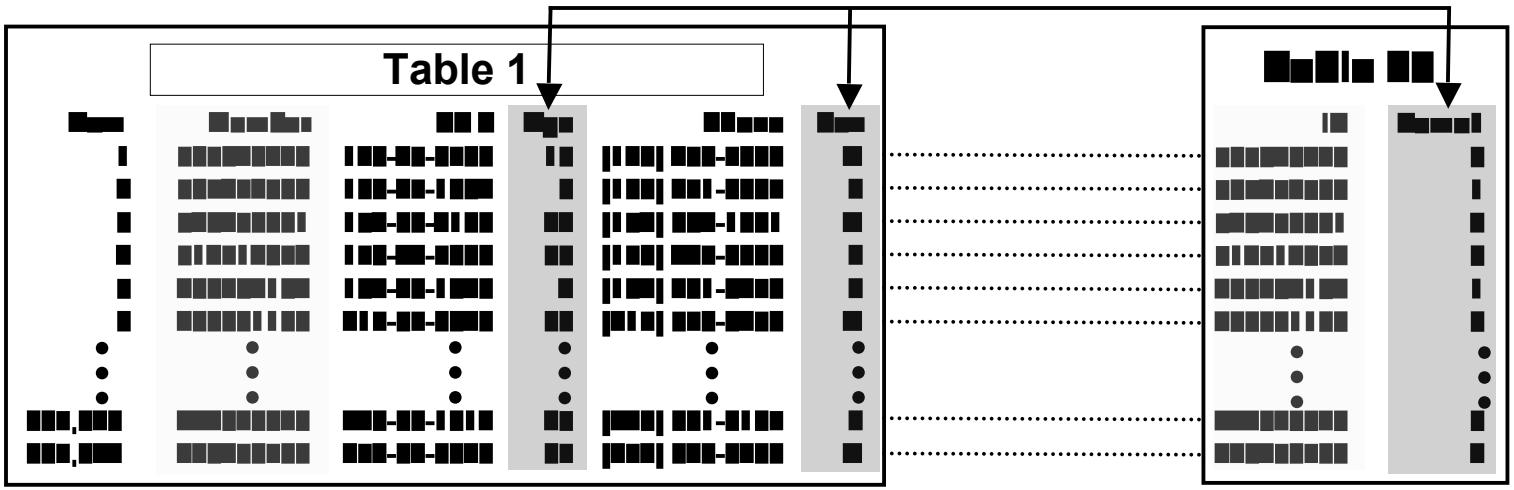
Step 2: With rows now aligned, analyzes the **data values** to **automatically discover**:

- Forgotten Business Rules
- Data Lineage
- Hidden Sensitive Data

CASE:

```
If age<18   and Sex=M then  0
If age<18   and Sex=F then  1
If age>=18 and Sex=M then  2
If age>=18 and Sex=F then  3
```

= Demo1



EXEROS

Data-Driven Approach: Discovers Business Rules & Sensitive Data

Data-Driven Discovery Engine

Step 3: With business rules now discovered, analyzes the **data values** to automatically discover:

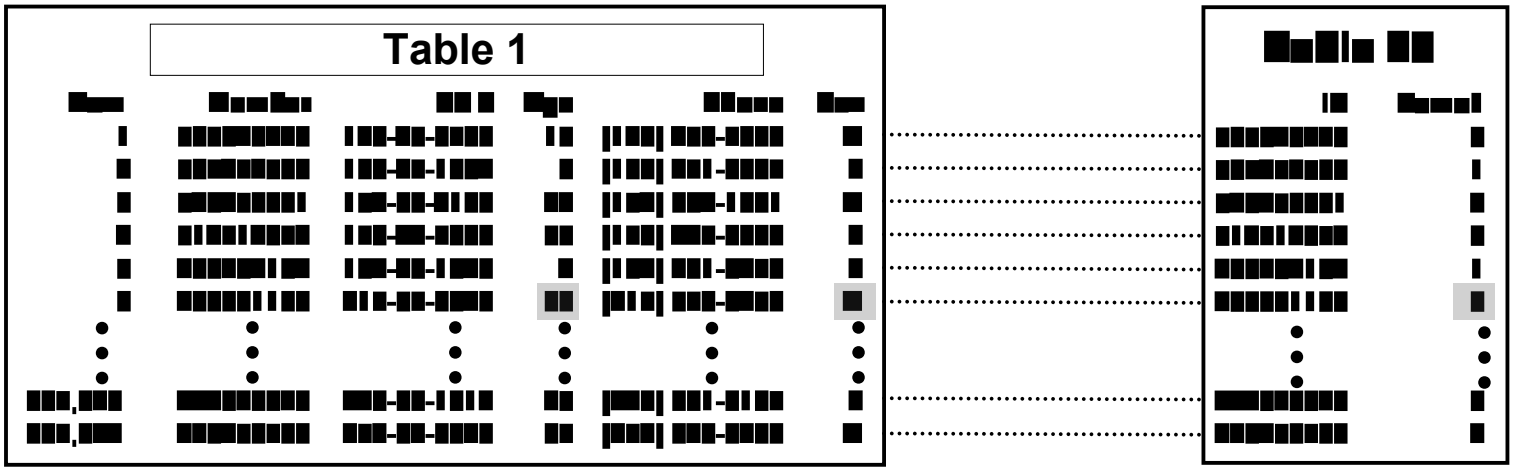
- Unknown Data Inconsistencies

Hit Rate: 98%

CASE:

```
If age<18    and Sex=M then   0
If age<18    and Sex=F then   1
If age>=18 and Sex=M then   2
If age>=18 and Sex=F then   3
```

= Demo1



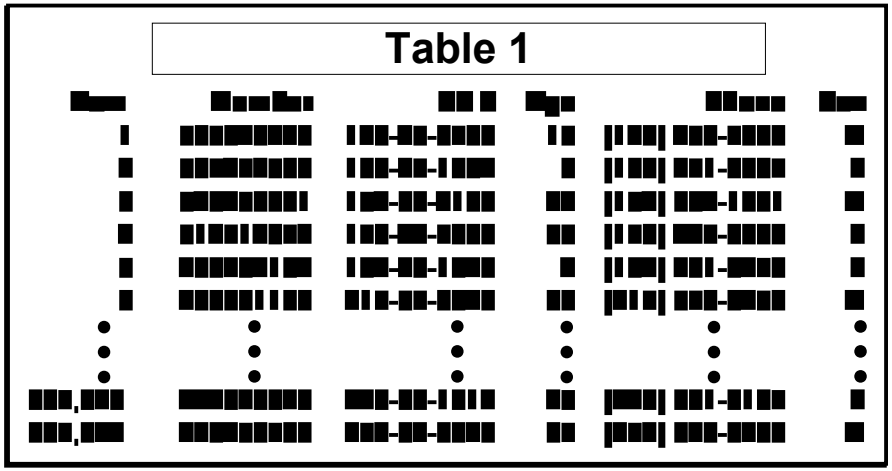
Data-Driven Approach: Re-validates Maps Against New Data

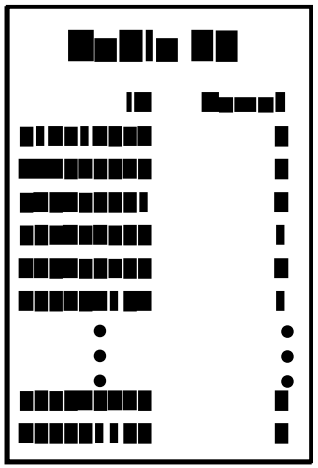
Data-Driven Discovery Engine

Step 4: Discovery Engine analyzes **NEW** data values to validate **previously discovered maps**

99% of Rows Map Correctly

- Exception: Row 32: Column Demo1: Expected Value = X, Actual Value = Y
- Exception: Row 125: Column Demo 23: Expected Value = A, Actual Value = B
- Etc…





Example:
Data Driven Discovery for Data Consistency

Credit Card Company Project:

- Integration of legacy system with product master system.

Manual results for first dataset:

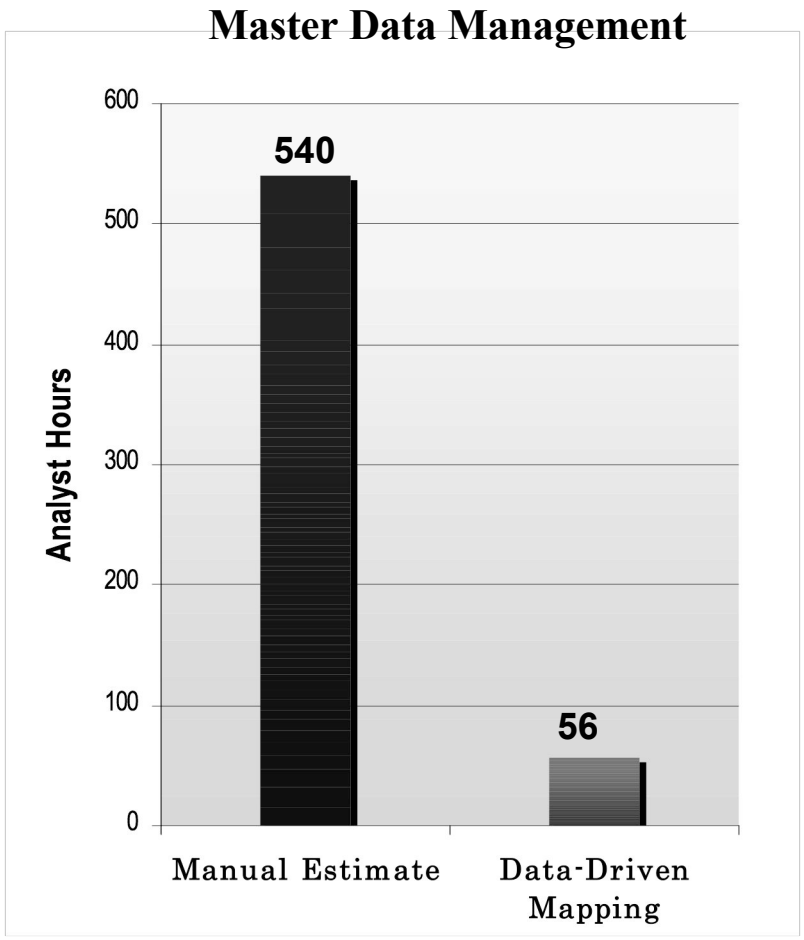
- Estimated to take 540 engineering hours
- Estimated to take 6 months elapsed

Data-Driven Mapping results:

- 56 hours total engineering hours
- 2.5 weeks of elapsed time

Benefits:

- Significant time to market savings (5 months+)
- Increased business competitiveness
- Increased accuracy of results
- Data Inconsistencies found as part of process
- Results are fully documented, auditable and can be leveraged for future projects



More Real Examples

Dataflow & Lineage Discovery

Financial Services Company

- **Project:** Discover dataflow across 1700 datasets to eliminate one-size fits all security & improve operational efficiency
- **Result:** Made a previously impossible project financial and technically feasible

Sensitive Data De-Identification

Insurance Company

- **Project:** De-Identify test and development data for off-shore team
- **Result:** Significantly reduced risk of missing hidden sensitive data and increase Chief Privacy Officer confidence in sign off
- **Result:** Cut time to discovery over 5x

EXEROS

EXEROS data relationship management

First Annual

Fastest Data Mapper Challenge!

Win $2500

Are you the fastest data mapper around?

Enter the Exeros Challenge to win $2500, a trip to DAMA International Symposium 2007, and bragging rights!!

Saturday, February 17th
12pm - 2pm Eastern time

Just map two small datasets to each other and mail back the results within the three-hour contest window. The winners will be judged on the number and accuracy of the target table columns that they mapped.

FIRST PRIZE: $2500, hotel, airfare and free registration to DAMA International Symposium and Wilshire Meta-Data Conference in Boston, March 4-8, 2007.
SECOND PRIZE: $1000
THIRD PRIZE: $500

http://www.exeros.com/html/mapoff.asp

So What does this have to do with the "Economics of Data Governance" ?

Data-Driven Discovery: Improves the Economics of Data Governance

- Improves the Top Line
  - Creates repeatable processes that provide leverage from the first project to the next
  - You will still want to focus on a first project with a critical business need or deadline
    - Regulations and compliance are OK for industries where non-compliance will shut you down… BUT
    - Improving company competitiveness and agility is a much easier sell to management
- Drives 5x faster time to initial results for your data governance program
- Improves the cost model by significantly cutting costs
  - Since 70% of the effort is spent discovering and documenting the data map… build it (quickly and cheaply), and they will come

EXEROS

Summary: Faster Time to Deployment Improves Data Governance ROI

- Companies are implementing data governance projects to:
  - Improve Security
  - Increase Consistency & Improve decision making
  - Decrease Regulatory Risk

- First technical step of data governance… Discovery

- To get positive ROI in a short time frame for data governance you must create a discovery process that is:
  - Fast
  - Cost-effective
  - Leverageable
  - Repeatable

EXEROS

Questions and Answers

Thank You for Attending!

For more information, contact :

Todd Goldman:

Web: www.exeros.com

Email: todd@exeros.com

Phone: +1.408.919.0191 ext 1115

EXEROS

The End